

---

---



# Comparación del IPC desagregado y total con modelos ARIMA

Jarvi Arley Rodriguez Enriquez

Universidad del Valle  
Facultad de Ciencias, Departamento de Matemáticas  
Cali, Colombia

2019



# Comparación del IPC desgregado y total con modelos ARIMA

Jarvi Arley Rodriguez Enriquez

Tesis presentada como requisito parcial para optar al título de:  
**Maestría en Ciencias Matemáticas**

Director:  
Ph.D., Jaime Arango C.

Universidad Del Valle  
Departamento de Matemáticas  
Cali, Colombia  
2019



## Resumen

En este trabajo se presenta un análisis de la serie del Índice de Precios al Consumidor (IPC) a nivel nacional. El IPC es una medida de la variación en el precio de bienes y servicios representativos del consumo de los hogares del país conocido como canasta. Con el fin de desarrollar pronósticos para seis meses, entre julio y diciembre de 2018, se tomó una muestra de 89 datos que corresponden a un registro mensual para el grupo de bienes y servicios, desde Enero de 2012 hasta Diciembre de 2018. Para este análisis se utilizaron modelos autorregresivos integrados de promedio móvil (ARIMA), el orden del modelo dependerá de los criterios dados por las pruebas de hipótesis y de los criterios de información.

**Palabras clave:** IPC, ARIMA, pronóstico, serie de tiempo

# Introducción

El Índice de Precios al Consumidor (IPC) es un indicador que publica mensualmente el Departamento Administrativo Nacional de Estadística (DANE)(DANE, 2006). El IPC es una medida de los incrementos de una lista de productos, bienes y servicios que componen la canasta familiar. Este indicador no solo mide la inflación sino que a la postre incide en la fijación del salario mínimo y en las tasas de interés con que prestan los bancos, es decir, en el valor del crédito. De allí la importancia del IPC en los modelos de planeación económica y en la vida cotidiana de los ciudadanos.

La evolución del IPC da información valiosa sobre la política monetaria del banco central (Banco de la República). Por otro lado, la predicción del IPC a corto plazo conlleva ventajas para los inversionistas, pues les da la oportunidad de vender papeles de deuda pública antes de que sus precios se degraden en el escenario de una inflación grande. También permite comprar estos mismos valores cuando la inflación es poca para su posterior reventa (Banco de la República, 2008).

Investigaciones y aplicaciones de las series de tiempo a indicadores económicos han sido realizadas previamente; Santana (2006) muestra que las redes neuronales pueden ser mejores que las metodologías autorregresivas integradas de media móvil estacional (SARIMA) de Box-Jenkins y el suavizamiento exponencial para la predicción del IPC. Por otro lado, Jalil & Misas (2006) hacen uso de redes neuronales artificiales con modelos lineales y no lineales para modelar el IPC. En Gallego (2010), se establecen tres modelos para la serie de tiempo de la Tasa Representativa del Mercado colombiano (TRM), resultados obtenidos de dos de estos modelos fueron contrastados con el pronóstico por medio de un modelo autorregresivo integrado de promedio móvil (ARIMA) de parámetros 1, 1, 3; ARIMA(1,1,3). Sin embargo, en base a los resultados obtenidos se concluye que ninguno de los métodos de pronósticos implementados sobre la TRM captura satisfactoriamente la información de esta estructura ya que el error RMSE que producen en el pronóstico a un paso es mayor que la variación diaria promedio de la variable. En Espinal et al. (2016) se presenta un análisis de la serie del IPC de la ciudad de Medellín; a partir de 72 datos se establece un modelo ARIMA(0,2,1), siendo un buen modelo dados los bajos errores y el nivel de significancia del 95% de los coeficientes. Finalmente en Huwiler & Kaufmann (2013) se hace un análisis de la serie de tiempo del IPC de Suiza estableciendo que, tomar los datos desagregados que componen el IPC para desarrollar modelos ARIMA es mejor que tomar el IPC total.

Uno de los modelos de series de tiempo más utilizados y más importantes es el modelo ARIMA. La popularidad del modelo ARIMA se debe a sus propiedades estadísticas, así como a la conocida metodología de Box-Jenkins en el proceso de construcción del modelo

(Zhang, 2003), además, los modelos ARIMA pueden implementar varios modelos de suavizado exponencial. Adicionalmente, los modelos ARIMA son bastante flexibles ya que pueden representar diferentes tipos de series de tiempo como el modelo autorregresivo puro (AR), promedio móvil puro (MA) y series combinadas de AR y MA (ARMA). Nuestro objetivo mostrar en una ventana de seis meses, considerando los criterios de determinación y de información, que los modelos ARIMA para los datos desagregados del grupo de bienes y servicios del IPC predicen de forma más precisa que el modelo ARIMA considerando la serie del IPC total. Se considera el criterio de Akaike (AIC), el criterio de información bayesiano (BIC) y el indicador del porcentaje medio absoluto (MAPE) como se llevó a cabo en Espinal et al. (2016). Adicionalmente, se considera el indicador de la raíz del error cuadrático medio (RMSE) y el error medio absoluto (MAE), utilizados con mayor frecuencia (Botchkarev, 2019); así como la prueba de Dickey-Fuller y la prueba de Dickey-Fuller Aumentada (Dickey & Fuller, 1979). Se hará un contraste con el mejor modelo ARIMA, según los indicadores y criterios establecidos, usando la serie total del IPC y las series de los datos desagregados que los componen. A diferencia del trabajo de Huwiler & Kaufmann (2003), en el cual no se consideró los indicadores de información MAPE, MAE, centrándose en un análisis de los pronósticos del ajuste de estos a los datos reales haciendo uso del indicador RMSFE; este trabajo considera los criterios, indicadores y pruebas establecidas para la selección de los modelos previos. Por otro lado, Theil (1954) mostró que la desagregación hace posible explotar más la información si esta presenta heterogeneidad. El análisis de las primeras y segundas diferencias de la series desagregadas del IPC permite, en caso de tendencia, establecer series con coeficientes de correlación que satisfacen la estacionariedad débil, encontrándonos con series heterogéneas que permiten el uso de modelos ARMA. La decisión previa al análisis de la construcción de los modelos ARIMA estará determinada por el estadístico de Durbin-Watson (Tsay, 2010).

# Contenido

|                                                                              |             |
|------------------------------------------------------------------------------|-------------|
| <b>Resumen</b>                                                               | <b>VII</b>  |
| <b>Introducción</b>                                                          | <b>VIII</b> |
| <b>1. Modelos Lineales y no Lineales</b>                                     | <b>1</b>    |
| 1.1. MARCO ECONÓMICO . . . . .                                               | 1           |
| 1.1.1. Índice de precios al consumidor . . . . .                             | 1           |
| 1.1.2. Manual de índice de precios al consumidor . . . . .                   | 1           |
| 1.2. Conceptos básicos . . . . .                                             | 2           |
| 1.2.1. Series de tiempo . . . . .                                            | 2           |
| 1.2.2. Estacionariedad . . . . .                                             | 2           |
| 1.2.3. Ruido blanco . . . . .                                                | 3           |
| 1.2.4. Función de correlación y autocorrelación. . . . .                     | 3           |
| 1.2.5. Función de autocorrelación parcial . . . . .                          | 6           |
| 1.2.6. Linealidad . . . . .                                                  | 6           |
| 1.3. Modelos Autorregresivos AR . . . . .                                    | 8           |
| 1.3.1. Modelo AR(1) . . . . .                                                | 9           |
| 1.3.2. Modelo AR( $p$ ) . . . . .                                            | 11          |
| 1.4. Modelos de Medias Móviles MA . . . . .                                  | 12          |
| 1.4.1. Modelo MA(1) . . . . .                                                | 12          |
| 1.4.2. Modelo MA( $q$ ) . . . . .                                            | 13          |
| 1.4.3. Modelos ARMA( $p,q$ ) . . . . .                                       | 14          |
| 1.4.4. Modelos ARIMA . . . . .                                               | 15          |
| <b>2. Teorema de Wold, Pruebas de Hipótesis y Criterios de Determinación</b> | <b>17</b>   |
| 2.1. Teorema de representación de Wold . . . . .                             | 18          |
| 2.1.1. OLS en forma matricial . . . . .                                      | 18          |
| 2.1.2. Teorema de representación de Wold . . . . .                           | 19          |
| 2.2. Prueba Dickey-Fuller . . . . .                                          | 25          |
| 2.2.1. Test de Dickey-Fuller . . . . .                                       | 25          |
| 2.2.2. Prueba de Dickey-Fuller Aumentada . . . . .                           | 26          |
| 2.3. Indicadores RMSE, MAE y MAPE . . . . .                                  | 26          |
| 2.4. Criterios de información AIC y BIC . . . . .                            | 28          |

---

|                                                                          |           |
|--------------------------------------------------------------------------|-----------|
| <b>3. Selección de Modelos</b>                                           | <b>29</b> |
| 3.1. Prediciendo con ARIMA . . . . .                                     | 29        |
| 3.1.1. Caminata Aleatoria vs Modelo AR(1) Sobre el IPC Total Histórico . | 29        |
| 3.1.2. Normalidad para las series desagregadas . . . . .                 | 34        |
| 3.1.3. Modelos ARIMA para las series desagregadas . . . . .              | 37        |
| 3.2. Modelo desagregado del IPC vs modelo total del IPC . . . . .        | 76        |
| <b>4. Conclusiones</b>                                                   | <b>79</b> |
| 4.1. Conclusiones . . . . .                                              | 79        |
| <b>A. Anexo: Códigos utilizados para cada modelo.</b>                    | <b>82</b> |
| A.0.1. Repositorio GitHub . . . . .                                      | 82        |

# 1. Modelos Lineales y no Lineales

Se introducirán conceptos como el IPC, inflación, y otros pertinentes para poner en contexto el problema que se tratará en capítulos posteriores. Introduciremos los modelos AR, MA, ARMA, ARIMA, considerados para el desarrollo del modelo final que predice un tentativo del índice de alimentos del IPC.

## 1.1. MARCO ECONÓMICO

### 1.1.1. Índice de precios al consumidor

El IPC es un índice que mide el ritmo al que los precios de los bienes y servicios de consumo cambian de un mes a otro. Los precios se recopilan en los comercios y demás puntos de venta minoristas. El método habitual de cálculo es tomar el promedio de los cambios de los precios entre un período y otro para los distintos productos, usando como ponderaciones los montos promedio que los hogares gastan en ellos. Los IPC son estadísticas oficiales comúnmente producidas por las oficinas nacionales de estadística, los ministerios de trabajo o bancos centrales. Se publican tan pronto como sea posible, por lo general alrededor de diez días después de finalizado el último mes. (DANE, 2006).

Un IPC mide el ritmo de la inflación de los precios conforme la experimentan y perciben los hogares en su papel de consumidores. Asimismo es muy utilizada como variable representativa del índice general de la inflación para la economía en su totalidad, en parte debido a la frecuencia y puntualidad con que se produce. Ha llegado a ser una estadística clave para la determinación de políticas, en especial de índole monetaria. En las leyes y en una gran diversidad de contratos privados suele caracterizarse como la medida apropiada de inflación a los fines de reajuste de pagos (por ejemplo, salarios, alquileres, intereses y beneficios de seguridad social) para dar cuenta de los efectos de la inflación. Por lo tanto, puede tener implicaciones financieras importantes y de amplio alcance para los gobiernos y las empresas, así como para los hogares.

### 1.1.2. Manual de índice de precios al consumidor

La organización Internacional del Trabajo (OIT), el Fondo Monetario Internacional (FMI), la organización para la Cooperación y el Desarrollo Económicos (OCDE), la Oficina Estadística de las comunidades Europeas (Eurostat), la Comisión Económica de las Naciones Unidas

para Europa (UNECE) y el Banco Mundial, junto con los expertos de una serie de oficinas nacionales de estadística y universidades, han trabajado desde 1998 en forma conjunta en el desarrollo de un manual considerado por el DANE y que se puede encontrar en la página oficial de la entidad. Sin embargo, debido a las limitaciones prácticas y de recursos, algunas de las recomendaciones presentes en el manual pueden ser difíciles de cumplir inmediatamente por parte de todas las oficinas de estadística, por lo cual se siguen como pautas u objetivos a la hora de revisar los índices y mejorar los programas de IPC.

## **1.2. Conceptos básicos**

### **1.2.1. Series de tiempo**

Granger & Newbold (1986) describen una serie de tiempo como una secuencia de observaciones ordenadas por un parámetro de tiempo, por ejemplo, la temperatura, los precios de la Coca-Cola en los últimos 15 años, el precio del tomate mes a mes. El orden de los datos es ahora considerado de gran importancia, la mayoría de las técnicas estadísticas clásicas ya no son relevantes por lo que nuevas técnicas tienen que ser ideadas. Las series de tiempo son encontradas en muchos campos, como la economía, meteorología (lluvia, temperatura, velocidad del viento), medicina (electrocardiogramas y electroencefalogramas), vibración física del sistema (como el de un automóvil que viaja sobre una superficie áspera), sismología, oceanografía, y geomorfología. Los métodos diseñados para tratar estos datos también se aplican a datos no registrados a través del tiempo, si no ordenados, por ejemplo, la altura del nivel del mar sobre una latitud, o la tasa de especiación sobre la latitud. Una serie de tiempo puede ser registrada de manera continua, por ejemplo con un esfigmomanómetro digital se puede medir la presión arterial, un oscilógrafo registra oscilaciones armónicas de un amplificador de audio. Sin embargo, la mayoría de las series disponibles consisten en lecturas tomadas a un tiempo predeterminado, intervalos de tiempo iguales, de modo que uno pueda obtener información horaria, diaria, mensual o lecturas trimestrales. Estos datos forman una serie temporal discreta, designada por  $r_t$ .

### **1.2.2. Estacionariedad**

Las series temporales pueden ser estacionarias o no estacionarias. Las series estacionarias se caracterizan por presentar un tipo de equilibrio estadístico alrededor de una media constante y una dispersión constante alrededor de esa media. Existen varios tipos de estacionariedad, se dice que una serie es estacionaria en un sentido débil o de segundo orden si esta tiene media y covarianza constante; se dice que una serie es estrictamente estacionaria si tiene además de media y varianza constante, una estructura constante de autocovarianza. Cuando

una serie posee esta estacionariedad de covarianza, la estructura de covarianza es constante en el tiempo. Es decir la autocovarianza sigue siendo la misma independiente del punto de referencia temporal. Específicamente,  $\{r_t\}$ , es débilmente estacionaria si  $E(r_t) = \mu$ , siendo  $\mu$  una constante y  $\text{Cov}(r_t, r_{t-\ell}) = \gamma_\ell$ , la cual solo depende de  $\ell$ , y se dice que estrictamente estacionaria si  $F_{t_1, \dots, t_k}(r_{t_1}, \dots, r_{t_k}) = F_{t_1+\ell, \dots, t_k+\ell}(r_{t_1}, \dots, r_{t_k})$  para cualquier  $k \in \mathbb{N}$ ,  $t_1, \dots, t_k, \ell \in \mathbb{Z}$  y  $r_1, \dots, r_t \in \mathbb{R}$  donde  $F_{t_1, \dots, t_k}$  denota la distribución conjunta de  $r_{t_1}, \dots, r_{t_k}$ .

Llamaremos a la covarianza  $\text{Cov}(r_t, r_{t-\ell})$ , lag- $\ell$  autocovarianza. Tenemos que la lag- $\ell$  cumple las propiedades.

$$\gamma_0 = \text{Var}(r_t),$$

$$\gamma_{-\ell} = \gamma_\ell$$

### 1.2.3. Ruido blanco

Una serie de tiempo  $r_t$  se dice que es ruido blanco si  $\{r_t\}$  es una secuencia de variables aleatorias independientes e idénticamente distribuidas con media y varianza finita. En particular, cuando  $r_t$  está normalmente distribuido con media cero y varianza  $\sigma^2$ , la serie es llamada ruido blanco gaussiano (Tsay, 2010).

### 1.2.4. Función de correlación y autocorrelación.

La covarianza entre dos variables aleatorias determina si existe una dependencia lineal entre ellas, esta nos va permitir definir el coeficiente de correlación entre dos series de tiempo. Definimos el coeficiente de correlación de dos variables  $X$  y  $Y$  como:

$$\rho_{x,y} = \frac{\sum_{t=1}^T (x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^T (x_t - \bar{x})^2 \sum_{t=1}^T (y_t - \bar{y})^2}}$$

donde  $\bar{x} = \sum_{t=1}^T \frac{x_t}{T}$  y  $\bar{y} = \sum_{t=1}^T \frac{y_t}{T}$  son las medias de  $X$  y  $Y$ , respectivamente.

#### *Función de Autocorrelación(ACF)*

El concepto de correlación se generaliza al de autocorrelación. Considerando la serie de tiempo débilmente estacionaria  $r_t$ , el coeficiente de correlación entre  $r_t$  y  $r_{t-\ell}$ , es llamado el lag- $\ell$  autocorrelación, y lo denotamos por  $\rho_\ell$ . El lag- $\ell$  autocorrelación será de nuestro interés para poder determinar la forma en que una serie de tiempo correlaciona con su pasado.

Mas específicamente, bajo el supuesto de que la serie de tiempo es débilmente estacionaria definimos:

$$\rho_\ell = \frac{\text{Cov}(r_t, r_{t-\ell})}{\sqrt{\text{Var}(r_t)\text{Var}(r_{t-\ell})}} = \frac{\text{Cov}(r_t, r_{t-\ell})}{\text{Var}(r_t)}$$

Para esta definición tenemos que  $\rho_0 = 1$ ,  $\rho_\ell = \rho_{-\ell}$  y  $-1 \leq \rho_\ell \leq 1$

Dada una serie de tiempo  $\{r_t\}_{t=1}^T$ , y sea  $\bar{r}$  su media, el lag- $\ell$  autocorrelación de  $r_t$  es:

$$\hat{\rho}_\ell = \frac{\sum_{t=\ell+1}^T (r_t - \bar{r})(r_{t-\ell} - \bar{r})}{\sum_{t=1}^T (r_t - \bar{r})^2}, \quad 0 \leq \ell \leq T-1.$$

Bajo algunas condiciones  $\hat{\rho}_\ell$  es un estimado consistente de  $\rho_\ell$ . Si  $\{r_t\}$  es una secuencia idénticamente distribuida que satisface  $E(r_t^2) < \infty$ , entonces  $\hat{\rho}_\ell$  es asintóticamente normal con media cero y varianza  $1/T$  para cualquier entero fijo positivo  $\ell$ . En general, si  $r_t$  es una serie débilmente estacionaria satisface  $r_t = \mu + \sum_{i=0}^q \Psi_i a_{t-i}$ , donde  $\Psi_0 = 1$  y  $\{a_j\}$  es una secuencia de variables idénticamente distribuidas con media cero, entonces  $\hat{\rho}_\ell$  es asintóticamente normal con media cero y varianza  $(1 + 2 \sum_{i=0}^q \rho_i^2)/T$  para  $\ell > q$ . Esta se conoce como la fórmula de Bartlett; ver (Box, Jenkins, & Reinsel, 1994).

### ***Test Individual ACF***

Para  $\ell$  entero positivo, el anterior resultado puede ser usado para el test  $H_0 : \rho_\ell = 0$  vs  $H_a : \rho_\ell \neq 0$ . El estadístico de prueba es:

$$t = \frac{\hat{\rho}_\ell}{\sqrt{(1 + 2 \sum_{i=1}^{t-1} \hat{\rho}_i^2)/T}}$$

Si  $r_t$  es una serie normal estacionaria que satisface  $\rho_j = 0$  para  $j > \ell$ , el  $t$  ratio es una variable asintóticamente distribuida como una normal estándar. La regla de decisión del test es rechazar  $H_0$  si  $|t| > Z_{\alpha/2}$ , donde  $Z_{\alpha/2}$  es el  $100(1 - \alpha/2)$ th percentil de la distribución normal estándar.

### ***Prueba Portmanteau***

Las aplicaciones financieras a menudo requieren probar que varias autocorrelaciones de  $r_t$  son cero, Box and Pierce (1970) proponen el estadístico de Portmanteau

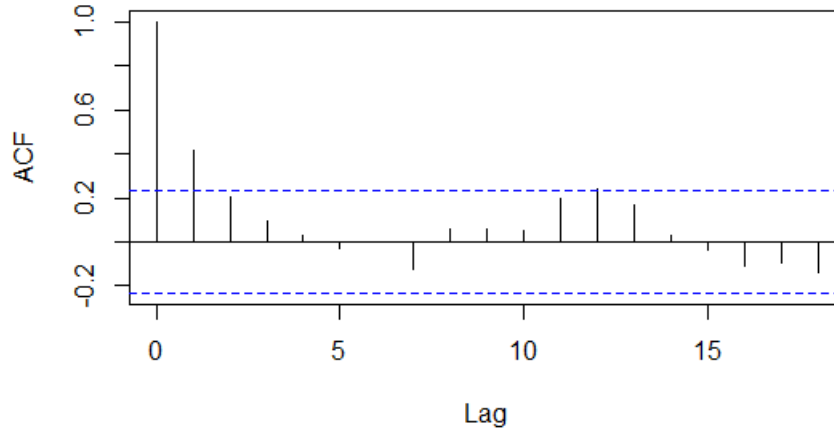
$$Q^*(m) = T \sum_{\ell=1}^m \hat{\rho}_\ell^2$$

como una prueba estadística con hipótesis nula  $H_0 : \rho_1 = \dots = \rho_m = 0$  frente a la hipótesis alternativa  $H_a : \rho_i \neq 0$  para algún  $i \in \{1, \dots, m\}$ . Bajo el supuesto de que  $\{r_t\}$  es una secuencia idénticamente distribuida,  $Q^*(m)$  es asintóticamente una distribución chi-cuadrado con  $m$  grados de libertad.

Ljung y Box modifican el  $Q^*(m)$  estadístico para aumentar la potencia de la prueba en muestras finitas

$$Q(m) = T(T+2) \sum_{\ell=1}^m \frac{\hat{\rho}_{\ell}^2}{T-\ell}$$

La regla de decisión es, rechazar  $H_0$  si  $Q(m) > \chi_{\alpha}^2$ , donde  $\chi_{\alpha}^2$  denota el  $100(1-\alpha)$ th percentíl de una distribución chi-cuadrado con  $m$  grados de libertad.



**Figura 1-1.:** Función de autocorrelación. Serie de tiempo del Dane desde Enero del 2011 hasta Febrero de 2017. Las dos líneas representan el límite del error estándar de la función ACF.

El valor  $\rho_1 = 0,5$  y el valor  $\rho_{12} = 0,268$ , el paquete de R usa  $1/T$  como la varianza asintótica de  $\hat{\rho}$  para  $\ell \neq 0$ , teniendo así como valores críticos las dos líneas horizontales que representan los límites del error estándar

$$V_{CR} = \pm \frac{1,96}{\sqrt{T}}.$$

Los valores críticos para la serie de tiempo del DANE desde Enero del 2011 hasta Febrero de 2017 son 0,23 y  $-0,23$ . Al superar el lag-1 y el lag-12 el valor crítico tenemos que son buenos predictores para el siguiente dato.

Los valores críticos de forma más general se calculan usando la fórmula

$$\text{Valores críticos} = \pm 1,96 \frac{\sigma}{\sqrt{n}}$$

Dadas las condiciones de estacionariedad débil, si los valores de la serie de tiempo están muy dispersos, el paquete de R no brindaría buenas estimaciones, pero en este caso tenemos que  $\sigma < 1$ , siendo  $V_{CR}$  un buen estimativo para los valores críticos de la función de correlación.

### 1.2.5. Función de autocorrelación parcial

La función de autocorrelación parcial (PACF) de una serie temporal estacionaria es una función de su ACF y es una herramienta útil para determinar el orden de un modelo AR. Se consideran modelos AR de órdenes consecutivos

$$\begin{aligned} r_t &= \phi_{0,1} + \phi_{1,1}r_{t-1} + e_{1t}, \\ r_t &= \phi_{0,2} + \phi_{1,2}r_{t-1} + \phi_{2,2}r_{t-2} + e_{2t}, \\ r_t &= \phi_{0,3} + \phi_{1,3}r_{t-1} + \phi_{2,3}r_{t-2} + \phi_{3,3}r_{t-3} + e_{3t}, \\ r_t &= \phi_{0,4} + \phi_{1,4}r_{t-1} + \phi_{2,4}r_{t-2} + \phi_{3,4}r_{t-3} + \phi_{4,4}r_{t-4} + e_{4t}, \\ &\vdots \end{aligned}$$

donde  $\phi_{0,j}$ ,  $\phi_{i,j}$  y  $\{e_{it}\}$  son, respectivamente, los términos constantes, el coeficiente de  $r_{t-i}$ , y el error del modelo AR( $j$ ). Estos modelos tienen la forma de una regresión lineal múltiple, y pueden ser calculados por el método de mínimos cuadrados. Están organizados de una forma secuencial para ilustrar mejor la aplicación de la prueba  $F$  parcial en el análisis de la regresión lineal múltiple. El estimado  $\hat{\phi}_{1,1}$  de la primera ecuación es llamado el lag-1 de la PACF de  $r_t$ . El estimado  $\hat{\phi}_{2,2}$  de la segunda ecuación es el lag-2 de la PACF de  $r_t$ , y así consecutivamente.

El lag-2 PACF  $\hat{\phi}_{2,2}$  muestra la contribución de  $r_{t-2}$  a  $r_t$  sobre el modelo AR(1). El lag-3 PACF muestra la contribución agregada de  $r_{t-3}$  sobre el modelo AR(2), y así sucesivamente. Por lo tanto, para un modelo AR( $p$ ), el lag- $p$  de la PACF podría no ser cero, pero  $\hat{\phi}_{j,j}$  podría ser cercano a cero para todo  $j > p$ . Haremos uso de esta propiedad para determinar el orden  $p$ .

### 1.2.6. Linealidad

Dada una serie temporal de media cero, como el valor de  $r$  en el momento  $t$  depende de su pasado, un modelo teórico capaz de describir un comportamiento sería:

$$r_t = f(r_{t-1}, r_{t-2}, \dots) + a_t \quad t = 1, 2, \dots$$

donde se exige que el comportamiento de  $r_t$  sea función de sus valores pasados, posiblemente finitos.

Dentro de los procesos estocásticos estacionarios se considerará únicamente la clase de procesos lineales que se caracterizan porque se pueden representar como una combinación lineal de variables aleatorias, en el caso de los procesos estacionarios con distribución normal y media cero, la teoría de procesos estocásticos señala que, bajo condiciones muy generales,  $r_t$  se puede expresar como una combinación lineal de los valores pasados finitos de  $r$  más una innovación ruido blanco:

$$r_t = \psi_1 r_{t-1} + \psi_2 r_{t-2} + \psi_3 r_{t-3} + \dots + a_t \quad (1-1)$$

El proceso debe ser no anticipante, es decir, que el presente no venga determinado por el futuro, luego  $r_t$  no puede depender de valores futuros de  $r$  o innovaciones  $a$ , debe ser a su vez invertible, es decir, que el presente dependa de forma convergente de su propio pasado, lo que implica que la influencia de  $r_{t-k}$  en  $r_t$  ha de ir disminuyendo conforme nos alejemos en el pasado. Esta condición se cumple si los parámetros del modelo general (1-1) cumplen la siguiente restricción:

$$\sum_{i=1}^{\infty} \psi_i^2 < \infty$$

El modelo (1-1) se puede escribir de forma más compacta en términos del operador de rezagos:

$$r_t = (\psi_1 B + \psi_2 B^2 + \dots) r_t + a_t$$

donde el operador de rezagos es tal que  $B r_t = r_{t-1}$ , de lo anterior tenemos

$$(1 - \psi_1 B - \psi_2 B^2 - \dots) r_t = a_t \longrightarrow \Psi_{\infty}(B) r_t = a_t$$

otra forma alternativa de escribir el modelo (1-1) es:

$$r_t = \frac{1}{\Psi_{\infty}(B)} a_t = \Phi_{\infty}(B) a_t = (1 + \theta_1 B + \theta_2 B^2 + \dots) a_t$$

$$r_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \dots \quad t = 1, 2, \dots$$

así, considerando media  $\mu \neq 0$  podemos expresar,

$$r_t = \mu + \sum_{i=0}^{\infty} \theta_i a_{t-i} \quad (1-2)$$

donde  $\mu$  es la media de  $r_t$ ,  $\theta_0 = 1$ , y  $\{a_t\}$  es una secuencia de variables aleatorias idénticamente distribuidas independientes con media cero (ruido blanco). El modelo (1-2) cumple la condición de estacionariedad si los parámetros satisfacen la siguiente restricción:

$$\sum_{i=1}^{\infty} \theta_i^2 < \infty$$

Esta restricción implica que el valor presente depende de forma convergente de las innovaciones pasadas, es decir, la influencia de la innovación  $a_{t-k}$  va desapareciendo conforme se aleja en el pasado.

### 1.3. Modelos Autorregresivos AR

El hecho que la tasa de retorno de una serie de tiempo tenga una autocorrelación lag-1 estadísticamente significativa indica que la tasa de retorno  $r_{t-1}$  puede ser útil para predecir el valor de  $r_t$ . Un modelo simple que hace uso de este poder predictivo es

$$r_t = \phi_0 + \phi_1 r_{t-1} + a_t \quad (1-3)$$

donde la serie  $\{a_t\}$  es ruido blanco con media cero y varianza  $\sigma_a^2$ . En la literatura de series temporales el modelo (1-3) es conocido como un modelo autorregresivo (AR) de orden 1 o simplemente modelo AR(1).

Es importante notar que el modelo AR(1) implica que, condicionado al valor pasado  $r_{t-1}$ , tenemos que:

$$E(r_t | r_{t-1}) = \phi_0 + \phi_1 r_{t-1}, \quad Var(r_t | r_{t-1}) = Var(a_t) = \sigma_a^2$$

Es decir, dado el retorno pasado  $r_{t-1}$  el retorno actual se centra en  $\phi_0 + \phi_1 r_{t-1}$  con desviación estándar  $\sigma_a^2$ . Esta es una propiedad de Markov tal que, condicionado a  $r_{t-1}$  el retorno  $r_t$  no está correlacionado con  $r_{t-i}$  para  $i > 1$ . Es evidente que existen situaciones en las cuales  $r_{t-1}$  solo, no determina la expectativa condicional de  $r_t$ , y un modelo más flexible puede dar mejores resultados. Una generalización directa del modelo AR(1) es un modelo AR(p):

$$r_t = \phi_0 + \phi_1 r_{t-1} + \dots + \phi_p r_{t-p} + a_t. \quad (1-4)$$

donde  $p$  es un entero no negativo y  $a_t$  es definido como en (1-3). Este modelo dice que las  $p$  variables pasadas  $r_{t-i}$  ( $i = 1, \dots, p$ ) determinan conjuntamente la expectativa condicional de  $r_t$  dados los datos pasados. El modelo AR(p) tiene la misma forma que un modelo de regresión múltiple con valores rezagados que sirven como variables explicativas.

Para el uso efectivo de los modelos AR estudiaremos sus propiedades básicas. Estudiaremos las propiedades de los modelos AR(1) y luego presentamos los resultados para el modelo general AR(p).

### 1.3.1. Modelo AR(1)

Asumiendo que la serie es débilmente estacionaria, tenemos que  $E(r_t) = \mu$ ,  $\text{Var}(r_t) = \gamma_0$ , y  $\text{Cov}(r_t, r_{t-j}) = \gamma_j$ , donde  $\mu$  y  $\gamma_0$  son constantes y  $\gamma_j$  es una función de  $j$ , no depende de  $t$ . Se puede obtener la media, la varianza, y autocorrelación de la serie de la siguiente manera. Dado que  $E(a_t) = 0$ , calculamos la esperanza de la ecuación (1-3) tenemos

$$E(r_t) = \phi_0 + \phi_1 E(r_{t-1}).$$

Bajo la condición de estacionariedad,  $E(r_t) = E(r_{t-1}) = \mu$ , y por tanto

$$\mu = \phi_0 + \phi_1 \mu \quad \text{or} \quad E(r_t) = \mu = \frac{\phi_0}{1 - \phi_1}.$$

Este resultado tiene dos implicaciones para  $r_t$ . Primero, la media de  $r_t$  existe si  $\phi_1 \neq 1$ . Segundo, la media de  $r_t$  es cero si y solo si  $\phi_0 = 0$ . Así, para un proceso estacionario AR(1), el término constante  $\phi_0$  está relacionado con la media de  $r_t$  via  $\phi_0 = (1 - \phi_1)\mu$  y  $\phi_0 = 0$  implica que  $E(r_t) = 0$ .

Usando  $\phi_0 = (1 - \phi_1)\mu$ , el modelo AR(1) puede ser escrito como

$$r_t - \mu = \phi_1(r_{t-1} - \mu) + a_t. \quad (1-5)$$

si sustituimos de forma repetitiva, la anterior ecuación, tenemos

$$\begin{aligned} r_t - \mu &= a_t + \phi_1 a_{t-1} + \phi_1^2 a_{t-2} + \dots \\ &= \sum_{i=0}^{\infty} \phi_1^i a_{t-i}. \end{aligned} \quad (1-6)$$

Observe que esta ecuación tiene la misma forma que la ecuación (1-2). Note que  $r_t - \mu$  es una función lineal de  $a_{t-i}$  para  $i \geq 0$ . Usando esta propiedad y la independencia de la serie  $\{a_t\}$ , obtenemos  $E[(r_{t-1} - \mu)a_t] = 0$ . Por la estacionariedad, tenemos  $\text{Cov}(r_{t-1}, a_t) = E[(r_{t-1} - \mu)a_t] = 0$ . De la ecuación (1-5) tenemos

$$\text{Var}(r_t) = \phi_1^2 \text{Var}(r_{t-1}) + \sigma_a^2.$$

donde  $\sigma_a^2$  es la varianza de  $a_t$ , se consideró que la covarianza entre  $r_{t-1}$  y  $a_t$  es cero. Bajo el supuesto de estacionariedad,  $\text{Var}(r_t) = \text{Var}(r_{t-1})$ , así que

$$\text{Var}(r_t) = \frac{\sigma_a^2}{1 - \phi_1^2}$$

siempre que  $\sigma_1^2 < 1$ . El requisito de  $\phi_1^2 < 1$  resulta de que la varianza de la variable aleatoria es acotada y no negativa. Consecuentemente, la estacionariedad débil de un modelo AR(1) implica que  $-1 < \phi_1 < 1$ , es decir,  $|\phi_1| < 1$ , y para que el proceso sea estacionario  $|\phi_1| < 1$ . Usando  $\phi_0 = (1 - \phi_1)\mu$ , podemos reescribir el modelo estacionario AR(1) como

$$r_t = (1 - \phi_1)\mu + \phi_1 r_{t-1} + a_t.$$

### **Función de autocorrelación de un Modelo AR(1)**

Multiplicando la ecuación (1-5) por  $a_t$ , usando la independencia entre  $a_t$  y  $r_{t-1}$ , y calculando la esperanza, tenemos

$$E[a_t(r_t - \mu)] = \phi_1 E[a_t(r_{t-1} - \mu)] + E(a_t^2) = E(a_t^2) = \sigma_a^2,$$

donde  $\sigma_a^2$  es la varianza de  $a_t$ . Multiplicando la ecuación (1-5) por  $r_{t-\ell} - \mu$ , y calculando la esperanza, y usando el resultado anterior, tenemos

$$\gamma_\ell = \begin{cases} \phi_1 \gamma_1 + \sigma_a^2 & \text{si } \ell = 0 \\ \phi_1 \gamma_{\ell-1} & \text{si } \ell > 0 \end{cases}$$

Donde  $\gamma_\ell = \gamma_{-\ell}$ . En consecuencia, para un modelo AR(1) débilmente estacionario en la ecuación (1-3) tenemos

$$\text{Var}(r_t) = \gamma_0 = \frac{\sigma_a^2}{1 - \phi_1^2} \quad y \quad \gamma_\ell = \phi_1 \gamma_{\ell-1}, \quad \text{para } \ell > 0.$$

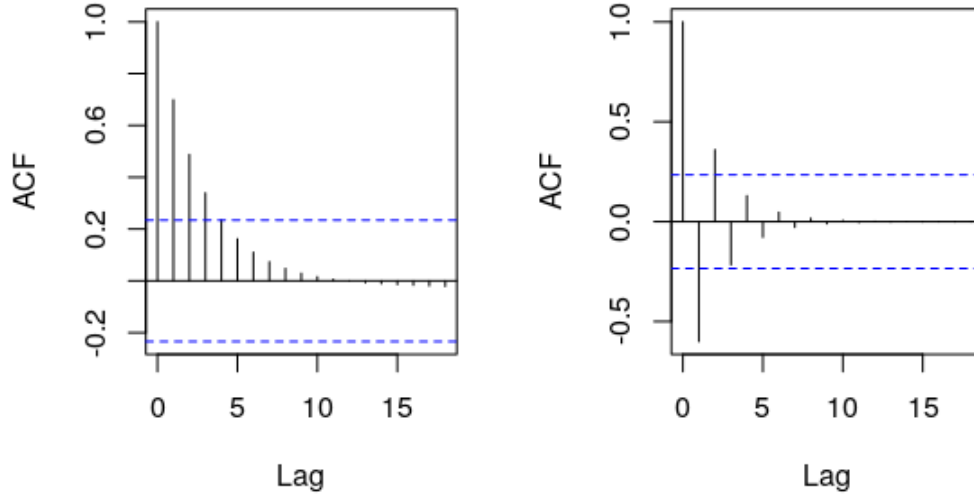
de la última ecuación, el ACF de  $r_t$  satisface

$$\rho_\ell = \phi_1 \rho_{\ell-1}, \quad \text{para } \ell > 0$$

dado que  $\rho_0 = 1$ , tenemos que  $\rho_\ell = \phi_1^\ell$ . Este resultado dice que el ACF de una serie débilmente estacionario AR(1) decae exponencialmente con tasa  $\phi_1$  y comenzando en el valor  $\rho_0 = 1$ .

Para un valor positivo de  $\phi_1$  la gráfica del ACF de un AR(1) decae exponencialmente. Para un valor negativo de  $\phi_1$ , la gráfica consiste de dos decaimientos alternantes, con tasa  $\phi_1^2$ . La figura (1-2) muestra el ACF de dos modelos AR(1),  $\phi_1 = -0,6$ .

Los datos que se usaron fueron generados usando  $r_t = 0,7r_{t-1} + \mu$ , con dato inicial  $r_0 = 1$ , bajo estas condiciones el correlograma siempre tendrá la regularidad correspondiente mostrada en la figura (1-2), ¿que el correlograma de los datos del DANE tenga un lag-1 por encima del límite del error estándar (Ver figura (1-1)) implica que un AR(1) es un buen ajuste a los datos? Se deben considerar pruebas de estacionariedad entre otros criterios para decidir si los datos se ajustan a un modelo AR(1).



**Figura 1-2.:** Función de autocorrelación de un modelo AR(1): izquierda  $\phi_1 = 0,7$ , derecha  $\phi_1 = -0,6$

### 1.3.2. Modelo AR( $p$ )

El proceso autorregresivo de orden  $p$ , expresa  $r_t$  en función de su pasado hasta el retardo  $t - p$  y una innovación contemporánea:

$$r_t = \phi_1 r_{t-1} + \phi_2 r_{t-2} + \dots + \phi_p r_{t-p} + a_t \quad t = 1, 2, \dots$$

donde  $a_t$  es ruido blanco. En términos del operador de retardos,

$$(1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p) r_t = a_t \longrightarrow \Phi_p(B) r_t = a_t$$

donde  $\Phi_p(B)$  recibe el nombre de polinomio autorregresivo y  $(\phi_1, \phi_2, \dots, \phi_p)$  es el vector de parámetros autoregresivos.

En primer lugar es preciso comprobar si el proceso AR( $p$ ) cumple las condiciones de estacionariedad para cualquier valor de los parámetros. Esta comprobación que es relativamente sencilla para el modelo AR(1), se complica para modelos autorregresivos de orden mayor cuyos parámetros han de satisfacer restricciones complejas para ser estacionarios. [2].

El siguiente teorema proporciona condiciones necesarias y suficientes para que el modelo AR( $p$ ) sea estacionario.

**Teorema 1.3.1** *Un proceso autorregresivo finito  $AR(p)$  es estacionario si y solo si el módulo de las raíces del polinomio autorregresivo  $\Phi_p(B)$  está fuera del círculo unidad.*

MirE mos el teorema para el modelo  $AR(2)$ .

### **Estacionariedad modelo $AR(2)$**

$$r_t = \phi_1 r_{t-1} + \phi_2 r_{t-2} + a_t \longrightarrow (1 - \phi_1 B - \phi_2 B^2) r_t = a_t$$

Polinomio autorregresivo  $\Phi_2(B) = 1 - \phi_1 B - \phi_2 B^2$ , así, por la condición de estacionariedad, tenemos

$$|B_1| = \left| \frac{\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2} \right| \quad \text{y} \quad |B_2| = \left| \frac{\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2} \right|$$

Si el radicando es positivo,  $\phi_1^2 + 4\phi_2 > 0$ , las raíces son reales y el módulo es el valor absoluto.

Si el radicando es negativo,  $\phi_1^2 + 4\phi_2 < 0$ , las raíces son complejas, y la gráfica de ACF de  $r_t$  mostraría una imagen de ondas de seno y coseno de amortiguamiento.

## **1.4. Modelos de Medias Móviles MA**

El modelo de medias móviles de orden finito  $q$ ,  $MA(q)$ , es una aproximación natural al modelo lineal general (1-4), se considera que la variable actual puede ser predicha por los residuos. El modelo sería:

$$r_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q}$$

donde  $a_t$  tiene media cero y varianza  $\sigma^2$  (ruido blanco). Empezaremos estudiando el modelo de medias móviles de primer orden,  $MA(1)$  y posteriormente el modelo más general,  $MA(q)$ .

### **1.4.1. Modelo $MA(1)$**

El modelo  $MA(1)$  determina el valor de  $r$  en el momento  $t$  en función de la innovación contemporánea y su primer retardo:

$$r_t = a_t - \theta_1 a_{t-1} \tag{1-7}$$

Observe que el modelo  $MA(1)$  es estacionario en media para cualquier valor del parámetro  $\theta_1$ ,  $E(r_t) = E(a_t - \theta_1 a_{t-1}) = 0$ . Además,

$$\begin{aligned}
\gamma_0 &= E(r_t - E(r_t))^2 = E(r_t)^2 = E(a_t - \theta_1 a_{t-1})^2 \\
&= E(a_t)^2 + \theta_1^2 E(a_{t-1})^2 - 2\theta_1 E(a_t a_{t-1}) = \sigma^2 + \theta_1^2 \sigma^2 - 0 \\
\gamma_0 &= (1 + \theta_1^2) \sigma^2 < \infty
\end{aligned}$$

Es decir, la varianza del modelo MA(1) es constante y finita para cualquier valor de  $\theta_1$ .

### ***Función de autocorrelación de un Modelo MA(1)***

Las autocovarianzas,  $\gamma_k$ ,  $k > 0$ :

$$\begin{aligned}
\gamma_1 &= E(r_t - E(r_t))(r_{t-1} - E(r_{t-1})) = E(a_t - \theta_1 a_{t-1})(a_{t-1} - \theta_1 a_{t-2}) \\
&= E(a_t a_{t-1}) - \theta_1 E(a_{t-1})^2 - \theta_1 E(a_t a_{t-2}) + \theta_1^2 E(a_{t-1} a_{t-2}) = -\theta_1 \sigma^2 < \infty \\
\gamma_2 &= E(r_t - E(r_t))(r_{t-2} - E(r_{t-2})) = E(a_t - \theta_1 a_{t-1})(a_{t-2} - \theta_1 a_{t-3}) = \\
&= E(a_t a_{t-2}) - \theta_1 E(a_{t-1} a_{t-2}) - \theta_1 E(a_t a_{t-3}) + \theta_1^2 E(a_{t-1} a_{t-3}) = 0
\end{aligned}$$

La función de autocovarianzas de un MA(1) es:

$$\gamma_\ell = \begin{cases} (1 + \theta_1^2) \sigma^2 & \text{si } \ell = 0 \\ -\theta_1 \sigma^2 & \text{si } \ell = 1 \\ 0 & \text{si } \ell > 0 \end{cases}$$

La función de autocovarianzas es finita y depende solo de  $\ell$  y no del tiempo, para cualquier valor del parámetro  $\theta_1$ , por lo tanto, no es necesario poner restricciones sobre el valor del parámetro  $\theta_1$  para que el modelo MA(1) sea estacionario. Tenemos así que la función de autocorrelación de un proceso MA(1) es:

$$\rho_\ell = \begin{cases} 1 & \text{si } \ell = 0 \\ \frac{-\theta_1}{1 + \theta_1^2} & \text{si } \ell = 1 \\ 0 & \text{si } \ell > 0 \end{cases}$$

La función de autocorrelación de un proceso MA(1) se caracteriza por ser una función truncada que sufre un corte brusco tomando el valor cero a partir del retardo 1.

### **1.4.2. Modelo MA( $q$ )**

El modelo medias móviles finito de orden  $q$ , es la aproximación natural al modelo lineal general (1-2), ya que supone su truncamiento, es decir que  $\phi_i = 0$  para todo  $i > q$ , este

modelo expresa el valor de  $r_t$  en un función de la innovación contemporánea y de su pasado hasta el retardo  $q$ :

$$r_t = a_t - \theta_1 a_{t-1} + \theta_2 a_{t-2} - \dots - \theta_q a_{t-q}$$

con  $a_t$  ruido blanco gaussiano. En términos del operador de retardos,

$$r_t = (1 - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q}) \longrightarrow r_t = \Theta_q(B) a_t$$

donde  $\theta_q(L)$  se denomina polinomio de medias móviles y  $(\theta_1, \theta_2, \theta_3, \dots, \theta_q)$  es el vector de parámetros de medias móviles.

El modelo  $MA(q)$  es una generalización del modelo  $MA(1)$ , por lo tanto sus características van a ser muy similares. Ahora bien, al introducir más retardos de la perturbación en el modelo, la memoria aumenta y la estructura dinámica representada por el modelo es más robusta. Si una perturbación entra en el modelo  $t$ , influye en  $r_t$ , en  $r_{t+1}$ , y en valores posteriores hasta  $r_{t+q}$ .

**Teorema 1.4.1** *Un proceso de medias móviles finito  $MA(q)$  es invertible si y solo si el módulo de las raíces del polinomio de medias móviles  $\Theta_q(B)$  está fuera del círculo unidad.*

Considerando el teorema para un  $MA(1)$

**Condición de invertibilidad de un Modelo  $MA(1)$**

$$r_t = a_t - \theta_1 a_{t-1} \longrightarrow r_t = (1 - \Theta(B)) a_t$$

Polinomio de medias móviles,  $\Theta_B = 1 - \theta_1 B$ , por la condición de invertibilidad,

$$|B| = \left| \frac{1}{\theta_1} \right| > 1 \longrightarrow |\theta_1| < 1$$

### 1.4.3. Modelos ARMA(p,q)

Los procesos autorregresivos de medias móviles (ARMA) determinan  $r_t$  en función de su pasado hasta el retardo  $p$ , la innovación contemporánea y el pasado de la innovación hasta el retardo  $q$

$$r_t = a_t + \sum_{i=1}^p \phi_i r_{t-i} + \sum_{i=1}^q \theta_i a_{t-i} \quad (1-8)$$

donde  $a_t$  es ruido blanco gaussiano. Podemos escribir el anterior modelo en términos del operador de retardos,

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

$$\Phi_p(B) r_t = \Theta_q(B) a_t$$

donde  $\Phi_B$  es el polinomio autorregresivo y  $\Theta_q(B)$  es el polinomio de medias móviles.

Las condiciones de estacionariedad del modelo ARMA( $p, q$ ) vienen impuestas por la parte autorregresiva dado que la parte de medias móviles finita siempre es estacionaria. Adicional a esto las condiciones de invertibilidad del modelo vienen impuestas por la parte de la medias móviles, dado que la parte autorregresiva finita siempre es invertible porque está directamente escrita en forma autorregresiva.

#### 1.4.4. Modelos ARIMA

Supongamos el modelo ARMA( $p, q$ ):

$$\Phi_p(B) r_t = \Theta_q(B) a_t \quad (1-9)$$

donde el polinomio AR se puede factorizar en función de sus  $p$  raíces  $B_1, B_2, \dots, B_p$ ,

$$\Phi_p(B) = (1 - B_1^{-1} B)(1 - B_2^{-1} B) \dots (1 - B_p^{-1} B)$$

supongamos que  $p - 1$  raíces son estacionarias (módulo fuera del círculo unidad) y una de ellas es unitaria,  $B_i = 1$ . Entonces, el polinomio AR, se puede escribir como sigue,

$$\Phi_p B = \Phi_{p-1}(B)(1 - B)$$

donde el polinomio  $\Phi_{p-1}(B)$  resulta del producto de los  $p - 1$  polinomios de orden 1 asociados a las raíces  $B_i$  con módulo fuera del círculo unidad. Sustituyendo en el modelo ARMA ( $p, q$ ) (1-9) se tiene que,

$$\Phi_{p-1}(B)(1 - B) r_t = \Theta_q(B) a_t \quad \longrightarrow \quad \Phi_{p-1}(B) \Delta r_t = \Theta_q(B) a_t \quad (1-10)$$

El polinomio  $\Phi_{p-1}(B)$  es estacionario porque todas sus raíces tienen modulo fuera del círculo unidad y el polinomio  $\Delta = 1 - B$  es el que recoge la raíz unitaria.

El modelo (1-10) representa el comportamiento de un proceso  $r_t$  que no es estacionario porque tiene una raíz unitaria. A un proceso  $r_t$  con estas características se le denomina *proceso integrado de orden 1*.

En general, el polinomio AR del modelo (1-9) puede contener más de una raíz unitaria, por ejemplo,  $d$ , entonces se puede descomponer como,

$$\Phi_p(B)\Delta^d r_t = \Theta_q(B)a_t$$

donde el polinomio  $\Phi_{p-d}(B)$  es estacionario porque sus  $p-d$  raíces tienen modulo fuera del círculo unidad, y el polinomio, y el polinomio  $\Delta^d = (1-B)^d$ , de orden  $d$ , contiene las  $d$  raíces unitarias no estacionarias. A un proceso  $r_t$  con estas características se le denomina *proceso integrado de orden  $d$*  y se denota por  $r_t \sim I(d)$

**Definición:** Un proceso  $r_t$  es integrado de orden  $d$ ,  $r_t \sim I(d)$ , si  $r_t$  no es estacionario, pero su diferencia de orden  $d$ ,  $\Delta^d r_t$ , sigue un proceso ARMA( $p-d, q$ ) estacionario invertible.

En la práctica, los procesos que surgen más habitualmente en el análisis de las series temporales económicas son los  $I(0)$  e  $I(1)$ , encontrándose los  $I(2)$  con mucha menos frecuencia (Gonzales, 2009).

En general, si una serie  $r_t$  es integrada de orden  $d$ , se puede representar por el siguiente modelo,

$$\Phi_p(B)\Delta^d r_t = \delta + \Theta_q(B)a_t \tag{1-11}$$

El modelo (1-11) se denomina modelo Autorregresivo Integrado de Medias Móviles de orden  $(p, d, q)$  o ARIMA( $p, d, q$ ), donde  $p$  es el orden del polinomio autorregresivo *estacionario*,  $d$  es el *orden de integración* de la serie, es decir, el número de diferencias que hay que tomar en la serie para que sea estacionaria, y  $q$  es el orden del polinomio de medias móviles *invertible*.

## 2. Teorema de Wold, Pruebas de Hipótesis y Criterios de Determinación

Considerar indicadores con criterio de información es lo previo a la selección del modelo. Se consideran las pruebas de hipótesis para establecer si una serie es estacionaria en media y en covarianza, pero es evidente en muchos casos que al observar la gráfica de la serie de tiempo se concluya que no se cumple dichas condiciones. Los indicadores con criterio de información abordan el compromiso entre la complejidad y la capacidad predictiva de un modelo. Cuanto más complejo sea un modelo peor será su capacidad para predecir en un abanico amplio de situaciones. Es decir, cuantas más variables y cuantas más interrelaciones entre los componentes incorpore un modelo más concretas serán sus predicciones pero también menos generalizables. Si un modelo es sencillo e incluye solo los componentes con mayor importancia en el sistema, podrá predecir en un mayor número de escenarios aceptando un cierto error de precisión.

El fin considerando las pruebas de hipótesis y los criterios de información es encontrar el modelo más parsimonioso, este será aquel que explique lo máximo con el menor número de parámetros. El estadístico de Durbin Watson y la prueba ADF permiten la escogencia del orden del modelo, una vez establecido el modelo, los criterios de información AIC (Akaike Information Criterion) y BIC (Bayesian Information Criterion), criterios que utilizan estimadores de máxima verosimilitud como bondad de ajuste y el número de parámetros como medida de complejidad, y los criterios de determinación MAE, RMASE y MAPE serán los que permitirán la selección del mejor modelo.

Los modelos se depuran hasta el punto de que haya ruido blanco, se parte del hecho de que la serie es débilmente estacionaria, comúnmente usado en la práctica sin dar cuenta del sustento teórico que garantiza que bajo estas condiciones siempre es posible construir un modelo ARIMA, dicho sustento lo establece el teorema de representación de Wold que establece que si una serie de tiempo es estacionaria en covarianza se puede descomponer en una parte determinista y una parte estocástica, en este capítulo demostramos el teorema en caso de que la serie es débilmente estacionaria, que pone una condición más, estacionariedad en media.

## 2.1. Teorema de representación de Wold

El proceso de modelar una serie de tiempo acaba cuando se ha expresado como una suma de un componente determinístico y uno estocástico que se comporta como ruido blanco. Las técnicas de regulación van encaminadas a que los residuales se comporten como ruido blanco, y se trabaja bajo este supuesto. El teorema de representación de Wold garantiza que si la serie temporal es estacionaria en covarianza se podrá expresar como la suma de un componente determinístico y uno estocástico, es el sustento teórico de peso que garantiza el modelamiento de series temporales cuando se cumplen las hipótesis.

### 2.1.1. OLS en forma matricial

Sea  $\Phi$  una matriz donde tenemos  $k$  variables independientes para  $n$  observaciones

1. Sea  $y$  el vector de las variables dependientes, de tamaño  $n \times 1$ .
2. Sea  $a$  el vector de errores  $n \times 1$ .
3. Sea  $\beta$  el vector de parámetros a estimar, de tamaño  $n \times 1$

tenemos así nuestro modelo

$$y = \Phi\beta + a$$

Bajo el supuesto de que tenemos un modelo preciso de los datos reales. El modelo tiene una componente determinista  $\Phi\beta$  y una estocástica  $a$ . Nuestro objetivo será determinar los parámetros del vector  $\beta$

#### ***Criterio para la estimación.***

Obtener nuestra estimación, es obtener una estimación  $\hat{\beta}$  que minimice la suma de los cuadrados residuales ( $\sum_i e_i^2$  en notación escalar)<sup>1</sup>.

El vector de los residuales es dado por:

$$e = y - \Phi\hat{\beta}$$

La suma de cuadrados residuales es  $e'e$ , escribiéndola como:

$$\begin{aligned} e'e &= (y - \Phi\hat{\beta})'(y - \Phi\hat{\beta}) \\ &= y'y - 2\hat{\beta}'\Phi'y + \hat{\beta}'\Phi'\Phi\hat{\beta} \end{aligned} \tag{2-1}$$

---

<sup>1</sup>Hay una diferencia entre perturbaciones que no pueden ser observadas ( $a$ ) y residuales ( $e$ ) que pueden ser observados

se tuvo en cuenta que la transpuesta de un escalar es un escalar, es decir  $y'\Phi\hat{\beta} = (y'\Phi\hat{\beta})' = \hat{\beta}'\Phi'y$

Para encontrar el  $\hat{\beta}$  que minimiza la suma de los cuadrados residuales, hacemos la derivada de Eq. 2-1 con respecto a  $\hat{\beta}$ . Teniendo así la siguiente ecuación.

$$\frac{\partial e'e}{\partial \hat{\beta}} = -2\Phi'y + 2\Phi'\Phi\hat{\beta} = 0 \quad (2-2)$$

El valor  $\hat{\beta}$  que buscamos es un mínimo, observe que al derivar la expresión Eq. 2-2 con respecto a  $\hat{\beta}$  obtenemos  $2\Phi'\Phi$ , matriz simétrica. En caso de que  $\Phi'\Phi$  sea de rango total, es positiva definida.

De la ecuación Eq. 2-2 tenemos la ecuación normal

$$(\Phi'\Phi)\hat{\beta} = \Phi'y \quad (2-3)$$

### 2.1.2. Teorema de representación de Wold

**Teorema 2.1.1 (Teorema de representación de Wold)** *Cualquier serie de tiempo  $\{r_t\}$  débilmente estacionaria de media cero puede ser expresada como*

$$r_t = v_t + s_t$$

donde

1.  $\{v_t\}$  es un proceso lineal determinístico, es decir una combinación lineal de los valores pasados de  $v_t$ .
2.  $s_t = \sum_{i=0}^{\infty} \phi_i e_{t-i}$  es un proceso infinito de medias móviles de los errores, donde  $\phi_0 = 1$ ,  $\sum_{i=0}^{\infty} \phi_i^2 < \infty$ , y  $\{e_t\}$  es ruido blanco, es decir  $E(e_t) = 0$ ,  $E(e_t^2) = \sigma^2$ ,  $E(e_t e_s) = 0$  para todo  $t$  y  $s$  con  $t$  distinto de  $s$ , y  $E(e_t v_s) = 0$ , para todo  $t$  y  $s$ .

Antes de proceder con la demostración del teorema haremos un ejemplo general de la aplicación del teorema.

#### ***Aplicación del teorema de representación de Wold***

Supongamos que queremos especificar una serie débilmente estacionaria  $\{r_t\}$  para modelar datos reales de una serie temporal real

$$\{r_t, t = 0, 1, \dots, T\}$$

se considera la siguiente estrategia:

- Se fija un parámetro  $p$ , que será el número de observaciones pasadas en la parte lineal determinista de la descomposición de  $\{r_t\}$
- Estimar la proyección lineal de  $r_t$  sobre  $(r_{t-1}, r_{t-2}, \dots, r_{t-p})$ 
  - Considerar una estimación de la muestra de tamaño  $n$  con punto final  $t_0 \leq T$ .
  - Dejemos que  $\{j = -(p-1), \dots, 0, 1, 2, \dots, n\}$  indexe la subserie de  $\{t = 0, 1, \dots, T\}$  correspondiente a la estimación de la muestra, y definimos  $\{y_j : y_j = r_{t_0-n+j}\}$ , (con  $t_0 \geq n + p$ )
  - Definimos el vector  $y = y_{(n \times 1)}$  y la matriz  $\Phi_{n \times [p+1]}$  como:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \Phi = \begin{pmatrix} 1 & y_0 & y_{-1} & \cdots & y_{-(p-1)} \\ 1 & y_1 & y_0 & \cdots & y_{-(p-2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & y_{n-1} & y_{n-2} & \cdots & y_{n-p} \end{pmatrix} \quad (2-4)$$

- Aplicando OLS tenemos

$$\begin{aligned} \hat{y} &= \Phi(\Phi^T \Phi)^{-1} \Phi y \\ &= \hat{y}^{(p)} \end{aligned} \quad (2-5)$$

- Calculando la proyección residual

$$\hat{e}^{(p)} = y - \hat{y}^{(p)}$$

- Se aplica un modelo MA a la serie de tiempo de los valores residuales  $\{\hat{e}_j^{(p)}\}$ , especificando un modelo:

$$e_t^{(p)} = \sum_{i=1}^{\infty} \phi_i e_{t-i}$$

produciendo  $\{\hat{\phi}_i\}$  y  $\{\hat{e}_t\}$ , estimaciones de los parámetros e innovaciones.

- Se realiza un análisis de consistencia al modelo establecido
  - Evaluar la ortogonalidad de  $\hat{e}^{(p)}$  sobre  $y_{t-s}$ ,  $s > p$ , si evidencia correlación, incrementar  $p$  y comenzar de nuevo.
  - Evaluar la consistencia de  $\{\hat{e}_t\}$  con la hipótesis de ruido blanco en el teorema. Si evidenciamos lo contrario, consideramos replantear el modelo.
    - Cambiar la especificación del modelo de medias móviles MA

- Adicionar variables deterministas al modelo.

*Prueba:* **(Teorema de representación de Wold)**

En Bierns (1996) se demuestra que con las hipótesis que planteamos se satisface que la serie temporal se puede expresar como un ruido blanco y uno determinístico y los valores donde  $r_{t+s}$  puede ser expresado como una combinación de los valores pasados de  $v_t$  y  $s_t$ , para efectos del trabajo haremos una prueba donde  $r_t$  puede ser expresada como la suma de los valores pasados de  $v_t$  y  $s_t$ . En el teorema de representación de Wold tenemos que la serie se expresa como la suma de dos términos, el primero completamente determinista, el segundo como algo parecido a un modelo de medias móviles de orden infinito.

Primero encontraremos los  $\phi_i$  y  $e_t$  que establezcan las propiedades requeridas. Después encontraremos los  $v_t$ , que resultan de el error de una proyección.

Sea  $r_t$  una serie de tiempo débilmente estacionaria. Sea

$$\hat{r}_t^{(n)} = P[r_t | r_{t-1}, \dots, r_{t-n}]$$

$P$  es la proyección de  $r_t$  sobre  $r_{t-1}, \dots, r_{t-n}$ . Escribimos

$$r_t = \hat{r}_t^{(n)} + e_t^{(n)}$$

Así, tenemos que

$$e_t^{(n)} \perp (r_{t-1}, \dots, r_{t-n})$$

De la estacionariedad débil de  $r_t$  se sigue que:

$$E(e_t^{2(n)}) = \sigma^{2(n)}$$

Tenemos entonces lo siguiente:

$$\hat{r}_t^{(n)} \rightarrow \hat{r} = P[r_t | r_{t-1}, r_{t-2}, \dots] \quad (2-6)$$

$$r_t = \hat{r}_t + e_t, \quad E(e_t^2) = \sigma^2 \quad (2-7)$$

$$e_t \perp (r_{t-1}, r_{t-2}, \dots) \quad (2-8)$$

de donde:

$$e_t = r_t - P[r_t | r_{t-1}, r_{t-2}, \dots]$$

es decir,  $e_t$  es una combinación lineal de  $r_t$  y su pasado, de donde se tiene que  $e_t$  es ortogonal al pasado de  $r_t$  y al pasado de  $e_t$ .

Observe que  $e_t$  es un proceso en serie no correlacionado.

### Proyección de $r_t$ sobre el presente y el pasado de $e_t$

Consideramos la proyección de  $r_t$  sobre el presente y el pasado de los  $e_t$

$$\tilde{r}_t^m = \sum_{j=0}^m \phi_j e_{t-j}$$

la notación,  $\tilde{r}_r^m$ , indica que la proyección usada aquí es diferente a la que se usó para definir los  $e_t$ . Tendremos en cuenta la ausencia de correlación entre los  $e_t$  para el análisis de los coeficientes de la proyección. De la condición de ortogonalidad asociada a la proyección tenemos:

$$E \left( r_t - \sum_{j=0}^m \phi_j e_{t-j} \right) e_{t-k} = 0, \quad k = 0, \dots, m,$$

de donde tenemos que:

$$E(r_t e_{t-k}) - \phi_k E(e_{t-k}^2) = 0,$$

considerando (2-7):

$$\phi_k = \begin{cases} \frac{E(r_t e_{t-k})}{\sigma^2}, & k = 1, 2, \dots, m \\ 1, & k = 0 \end{cases} \quad (2-9)$$

Así  $E(r_t e_t) = \sigma^2$ , esto porque

$$E(r_t e_t) = E(\hat{r}_t + e_t) e_t = \sigma^2,$$

Una propiedad clave de la proyección es que  $\phi_k$  no depende de  $m$ , esto es un reflejo de la falta de correlación serial de los  $e_t$ .

Pasamos a establecer la suma de los cuadrados de los  $\phi_j$ . Teniendo en cuenta que:

$$E \left( r_t - \sum_{j=0}^m \phi_j e_{t-j} \right)^2 \geq 0,$$

o,

$$\begin{aligned} E(r_t^2) - 2 \sum_{j=0}^m \phi_j E(r_t e_{t-j}) + \sum_{j=0}^m \phi_j^2 \sigma^2 \\ = E(r_t^2) - \sigma^2 \sum_{j=0}^m \phi_j^2 \geq 0 \end{aligned}$$

Esto es verdad para cualquier valor de  $m$ . Dado que  $E(r_t^2)$  es un número fijo por la estacionariedad débil, concluimos que

$$\lim_{m \rightarrow \infty} \sum_{j=0}^m \phi_j^2 < \infty$$

De lo anterior concluimos que esta suma converge a algún número finito:

$$\lim_{m \rightarrow \infty} \sum_{j=0}^m \phi_j^2 \rightarrow \sum_{j=0}^{\infty} \phi_j^2$$

dado que los terminos  $\phi_t$  son cuadrados sumables, podemos mostrar que  $\tilde{r}_t^m$  es una sucesión de Cauchy, es decir

$$\tilde{r}_t^m = \sum_{j=0}^m \phi_j e_{t-j} \rightarrow \tilde{r}_t = \sum_{j=0}^{\infty} \phi_j e_{t-j}.$$

Observe que para cada  $\epsilon > 0$ , dado el resultado anterior, existe  $N$  talque si  $m > n > N$ ,

$$E(\tilde{r}_t^m - \tilde{r}_t^n)^2 = \left( \sum_{j=n+1}^m \phi_j^2 \right) \sigma^2 < \epsilon$$

### Construcción de los $v_t$

Definimos  $v_t$  como la diferencia entre  $r_t$  y su proyección sobre el presente y el pasado de los  $e_t$ .

$$v_t = r_t - \tilde{r}_t \quad (2-10)$$

$E(v_t e_s) = 0$  para  $s > t$  porque  $v_t$  es una combinación lineal de  $r_t$  y el pasado de los  $e_t$ , y  $e_s$  es ortogonal a estos para  $s > t$ . Que  $E(v_t e_s) = 0$  para  $s \leq t$ , se concluye de que  $v_t$  es el error en la proyección de  $r_t$  sobre el presente y pasado de  $e_t$ . Concluimos que:

$$E(v_t e_s) = 0 \text{ para todo } t, s$$

Procedemos a establecer que  $v_t$  es perfectamente predecible desde el pasado de los  $r_t$ . Usando la propiedad de la linealidad de la proyección,  $P[A + B|\Omega] = P[A|\Omega] + P[B|\Omega]$ :

$$P[v_t | r_{t-1}, r_{t-2}, \dots] = P[r_t | r_{t-1}, r_{t-2}, \dots] - \sum_{j=0}^{\infty} \phi_j P[e_{t-j} | r_{t-1}, r_{t-2}, \dots], \quad (2-11)$$

Considerando el último término que involucra la proyección de los  $e_t$ . En el caso  $j = 0$ :

$$P[e_t|r_{t-1}, r_{t-2}, \dots] = \sum_{j=1}^{\infty} \lambda_j r_{t-j}.$$

De la ortogonalidad asociada a la proyección:

$$E \left( e_t - \sum_{j=1}^{\infty} \lambda_j r_{t-j} \right) r_{t-k} = 0 \quad k = 1, 2, 3, \dots$$

recordemos que  $e_t$  es ortogonal a  $(r_{t-1}, r_{t-2}, \dots)$ , de modo que  $\lambda_j = 0$  satisface la condición de ortogonalidad. Esta condición garantiza que

$$P[e_t|r_{t-1}, r_{t-2}, \dots] = 0$$

Consideremos ahora

$$P[e_t|r_t, r_{t-1}, \dots]$$

recordemos que  $e_t$  es una combinación lineal del presente y pasado de los  $r_t$ :

$$e_t = r_t - P[r_t|r_{t-1}, r_{t-2}, \dots],$$

de donde:

$$P[e_t|r_t, r_{t-1}, \dots] = e_t$$

Si consideramos el problema de optimización que define una proyección:

$$\min_{\{\lambda_j\}_{j=0}^{\infty}} E \left( e_t - \sum_{j=0}^{\infty} \lambda_j r_{t-j} \right)^2$$

al escoger los  $\lambda_j$  para que coincidan con la función lineal,  $r_t - P[r_t|r_{t-1}, r_{t-2}, \dots]$ , este valor mínimo puede ser cero, el cual es un valor que no se puede mejorar. Observe que si se cumple para  $j$  desde cero, de forma análoga se cumple

$$P[e_t|r_{t+j}, r_{t+j-1}, \dots] = e_t, \quad j \geq 0$$

con el previo resultado, escribimos (2-10) de la siguiente forma:

$$P[v_t|r_{t-1}, r_{t-2}, \dots] = P[r_t|r_{t-1}, r_{t-2}, \dots] - \sum_{j=1}^{\infty} \phi_j e_{t-j}$$

con lo anterior y (2-10) tenemos:

$$v_t - P[v_t|r_{t-1}, r_{t-2}, \dots]$$

$$\begin{aligned}
&= r_t - P[r_t | r_{t-1}, r_{t-2}, \dots] - \left( \sum_{j=0}^{\infty} \phi_j e_{t-j} - \sum_{j=1}^{\infty} \phi_j e_{t-j} \right) \\
&= e_t - \phi_0 e_t = 0
\end{aligned}$$

,  
estableciendo que

$$v_t = P[v_t | r_{t-1}, r_{t-2}, \dots]$$

*Q.E.D..*

## 2.2. Prueba Dickey-Fuller

La idea básica de una caminata aleatoria es que el valor de mañana de una serie es el valor de hoy más un cambio impredecible, es decir, el proceso estocástico estaría modelado por  $r_t = r_{t-1} + a_t$  siendo  $a_t$  ruido blanco, en el capítulo 3 hablaremos de la posibilidad de que la serie de tiempo del IPC del Dane sea una caminata aleatoria. La caminata aleatoria no capta tendencia ni estacionariedad, es decir, únicamente encontramos la serie temporal con el error  $a_t$ . La caminata aleatoria de la forma más simple, tiene una raíz unitaria. En general, para modelos  $AR(p)$  tener una raíz unitaria representa un problema, no se tendría estacionariedad, la prueba de Dickey-Fuller Aumentada nos permite analizar el caso de la posibilidad de una raíz unitaria en los modelos a tratar.

### 2.2.1. Test de Dickey-Fuller

Para probar si una serie de tiempo  $\{r_t\}$  sigue una caminata aleatoria o una caminata aleatoria con deriva, empleamos los modelos

$$r_t = \phi_1 r_{t-1} + a_t \quad (2-12)$$

$$r_t = \phi_0 + \phi_1 r_{t-1} + a_t \quad (2-13)$$

donde  $\phi_1$  es un número real y  $a_t$  es una secuencia de variables aleatorias distribuidas normalmente con media cero y varianza  $\sigma_a$ . La serie  $r_t$  converge a una serie estacionaria si  $\phi_1 < 1$ . Si  $|\phi_1| = 1$  la serie  $r_t$  no es estacionaria y la varianza de  $r_t$  es  $\sigma_a t$ . Cuando  $\phi_1 = 1$  como ya hemos mencionado, se dice que  $r_t$  es una caminata aleatoria. Consideremos como hipótesis nula  $H_0 : \phi_1 = 1$  versus la hipótesis alternativa  $H_a : \phi_1 < 1$ . (Dickey & Fuller, 1979). Dadas  $T$  observaciones  $r_1, r_2, \dots, r_T$  el estimador de máxima verosimilitud de  $\phi_1$  es el estimador por mínimos cuadrados

$$\hat{\phi}_1 = \frac{\sum_{t=1}^T r_{t-1} r_t}{\sum_{t=1}^T r_{t-1}^2}, \quad \hat{\sigma}_a^2 = \frac{\sum_{t=1}^T (r_t - \hat{\phi}_1 r_{t-1})^2}{T - 1}$$

,  
donde  $r_0 = 0$ . El  $t$  ratio de la prueba de *Dickey-Fuller* (DF) es

$$\text{DF} \equiv t \text{ ratio} = \frac{\hat{\phi}_1 - 1}{\text{std}(\hat{\phi}_1)} = \frac{\sum_{t=1}^T r_{t-1} a_t}{\hat{\sigma}_a \sqrt{\sum_{t=1}^T r_{t-1}^2}}$$

### 2.2.2. Prueba de Dickey-Fuller Aumentada

Se estableció la prueba basado en Tsay (2010). Esta prueba se restringe a los modelos planteados (2-12) y (2-13), pero en general los modelos  $\text{ARIMA}(p, d, q)$  son más apropiados para modelar series de tiempo, como vimos un modelo  $\text{ARIMA}$ , es un modelo  $\text{ARMA}$  con  $d$  diferencias y en la literatura, los modelos  $\text{AR}(p)$  son usados a menudo, y es la parte del modelo  $\text{ARMA}$  de tratar la estacionariedad. Si  $r_t$  es la serie de tiempo, para verificar la existencia de una raíz unitaria en un proceso  $\text{AR}(p)$ , realizamos la prueba,  $H_0 : \beta = 1$  vs.  $H_a : \beta < 1$  usando la regresión

$$r_t = c_t + \beta r_{t-1} + \sum_{i=1}^{p-1} \phi_i \Delta r_{t-i} + e_t, \quad (2-14)$$

donde la función  $c_t$  es determinística y  $\Delta r_j = r_j - r_{j-1}$ , es la serie de diferencias de  $r_t$ , en la practica  $c_t$  puede ser cero o constante, dependiendo de si tiene o no tendencia. El  $t$  ratio de  $\hat{\beta} - 1$  de la prueba de *Dickey-Fuller Aumentada* (ADF) es,

$$\text{ADF} - \text{test} = \frac{\hat{\beta} - 1}{\text{std}(\hat{\beta})}$$

donde  $\hat{\beta}$  es el estimador por mínimos cuadrados de  $\beta$ . Note que la primera diferencia en la ecuación 2-14 es equivalente a un modelo  $\text{AR}(p)$  con función determinística  $c_t$ . Haciendo  $\beta_c = \beta - 1$ , podemos escribir la ecuación 2-14 de la forma,

$$\Delta r_t = c_t + \beta_c r_{t-1} + \sum_{i=1}^{p-1} \phi_i \Delta r_{t-i} + e_t$$

considerando las hipótesis  $H_0 : \beta_c = 0$  vs.  $H_a : \beta_c < 0$  se tienen que la prueba es equivalente a la prueba ADF.

## 2.3. Indicadores RMSE, MAE y MAPE

Sea  $r_t$  una serie de tiempo, se definen los residuales de las series de tiempo como,  $r_t - \hat{r}_t$ . De forma general la estimación del error promedio que se genera al estimar un modelo se ve como,

$$\bar{e}_\gamma = \left( \frac{\sum_{i=1}^T w_i |e_i|^\gamma}{\sum_{i=1}^T w_i} \right)^{\frac{1}{\gamma}}$$

donde  $\gamma \leq 0$  y  $w_i$  es la escala asignada a cada  $|e_i|^\gamma$  según su influencia sobre el error total.. Este indicador es comúnmente usado cuando  $w_i = 1$  para todo  $i$ , y cuando  $\gamma = 2$ ; que es más conocido como la raíz del error cuadrático medio (RMSE),

$$\text{RMSE} = \left( T^{-1} \sum_{i=1}^T |e_i|^2 \right)^{\frac{1}{2}} \quad (2-15)$$

De uso menos frecuente (Willmott & Matsuura, 2005), se establece el error absoluto medio (MAE), cuando  $\gamma = 1$  como

$$\text{MAE} = \left( T^{-1} \sum_{i=1}^T |e_i| \right) \quad (2-16)$$

Tanto MAE como RMSE expresan el error promedio de predicción del modelo. Ambas métricas pueden variar de 0 a  $\infty$  y son indiferentes a la dirección de los errores. Son puntuaciones con orientación negativa, lo que significa que los valores más bajos son mejores. Sacar la raíz cuadrada de los errores cuadrados promedio tiene algunas implicaciones interesantes para RMSE. Como los errores se potencian antes de promediar, el RMSE otorga un peso relativamente alto a los errores grandes. Esto significa que el RMSE debería ser más útil cuando los errores grandes son particularmente indeseables.

Las métricas RMSE, MAE y en el error porcentual medio absoluto MAPE son las más usadas en estudios de investigación. (Botchkarev, 2019)

Los indicadores MAE, RMSE, y error porcentual absoluto medio MAPE son comúnmente usados para comparar medidas de rendimiento en el pronóstico, establecidos los modelos, pero MAE y RMSE pueden ser inapropiadas. Esto se debe a que a menudo puede haber grandes variaciones en la escala de las observaciones entre series, de modo que unas pocas series con valores grandes pueden dominar las comparaciones. En estos casos es necesario utilizar medidas libres de estos errores, empleando un ajuste que considera el valor real de la serie, y el error de porcentaje absoluto medio (MAPE) es probablemente la medida más utilizada de este tipo (Goodwin & Lawton, 1999). Definimos el error porcentaje absoluto medio (MAPE) como,

$$\text{MAPE} = \frac{\sum_{t=1}^T APE_t}{T}$$

donde,

$$APE_t = \left| \frac{r_t - \hat{r}_t}{r_t} \right| \times 100$$

$r_t$  es el valor real de la serie en el tiempo  $t$  y  $\hat{r}_t$  su estimado según el modelo.

## 2.4. Criterios de información AIC y BIC

Ver Tsay (2010). Con los criterios de información de Akaike (AIC) y de información Bayesiano (BIC) daremos una medida de bondad de ajuste de los modelos, estas nos darán un estimado entre la bondad de ajuste y complejidad del modelo.

Teniendo en cuenta el modelo  $AR(p)$  expuesto en el capítulo 1, en general tenemos que el modelo AIC se define como,

$$AIC(\ell) = \ln \bar{\sigma}_\ell^2 + \frac{2\ell}{T}$$

donde  $\bar{\sigma}_\ell^2$  es el estimador de máxima verosimilitud de  $\sigma_a^2$ , la cual es la varianza de  $a_t$  y  $T$  es el tamaño de la muestra. Tenemos que entre mayor sea la cantidad de parámetros hay un sobreajuste de los datos. El estimador de la varianza nos mide la dispersión del error de los datos, así considerando la formula que define el AIC tenemos que entre más pequeño sea este valor, mejor la calidad relativa del modelo.

El criterio de información Bayesiana (BIC), ha sido uno de los métodos más populares usado para la selección de modelos. Este es un criterio de evaluación de modelos en términos de sus probabilidades posteriores.

Para el modelo  $AR(p)$  el criterio de información de información Bayesiano (BIC) es

$$BIC(\ell) = \ln \bar{\sigma}_\ell^2 + \frac{\ell \ln T}{T}$$

La diferencia esta en la penalidad, en el AIC la penalidad es 2, mientras que la penalidad es  $\ln T$ , el BIC tiende a seleccionar un modelo AIC inferior cuando el tamaño de la muestra es grande.

## 3. Selección de Modelos

Las caminatas aleatorias nacen del movimiento Browniano, el movimiento Browniano fue nombrado después de que el botánico inglés, Robert Brown, que en 1827 observó que las partículas pequeñas inmersas en un líquido, exhibió un movimiento aleatorio. Sin embargo algunas personas habían observado este fenómeno antes que él, varios de ellos, incluido Einstein trabajaron con la serie, hasta que Nobert Wiener le dió un tratamiento matemático riguroso en 1918.

Podrá pensarse que la serie histórica del Dane sigue un movimiento browniano en una dimensión, para este caso el modelo sería  $Ar(1)$ , donde la raíz unitaria es uno, procedemos a analizar el IPC histórico haciendo notar que se evidencian regularidades que deben ser descritas con técnicas de modelamiento.

Se procede después al análisis de las series desagregadas considerando el estadístico de Durbin Watson, la prueba ADF, los criterios de determinación AIC y BIC, y los indicadores de determinación MAE, RMSE y MAPE estas series se pueden encontrar en DANE (2019b). Se tomará una ventana de seis meses para hacer el análisis debido a que, aunque los modelos cambian, el orden de los modelos permanece en ese intervalo de tiempo.

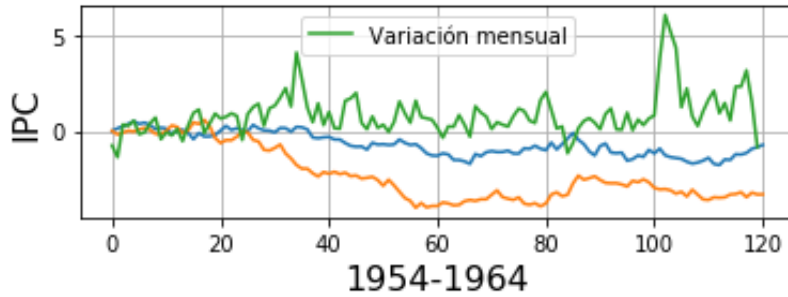
### 3.1. Prediciendo con ARIMA

#### 3.1.1. Caminata Aleatoria vs Modelo AR(1) Sobre el IPC Total Histórico

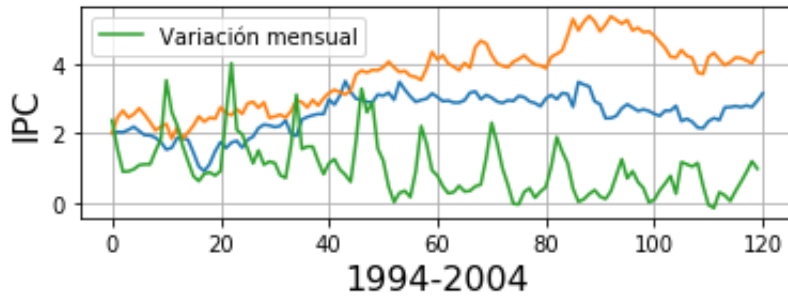
Decir que la serie de tiempo del DANE sigue una caminata aleatoria es predecirla con un  $Ar(1)$  donde el error sigue una distribución normal aleatoria y no depende del tiempo, es decir,

$$D_t = D_{t-1} + \epsilon_t$$

En la siguiente gráfica observamos los datos del IPC del DANE comparados con caminatas aleatorias.



**Figura 3-1.:** Se tomaron 120 datos. La línea azul y naranja caminatas aleatorias, línea verde IPC dado por el DANE entre el año de 1954 al año 1964



**Figura 3-2.:** Se tomaron 120 datos. La línea azul y naranja caminatas aleatorias, línea verde IPC de alimentos dado por el DANE entre el año de 1994 al año 2004

Los valores de la caminata aleatoria están generados con una distribución normal. La escala  $\delta$  específica la desviación estándar de la distribución normal que usamos para el movimiento browniano simulado.

$$\delta = \sum_{t=1}^{120} \frac{\delta_t}{120}$$

donde cada  $\delta_t$  es la diferencia de dos índices consecutivos del IPC, para los IPC en el período correspondiente.

La caminata aleatoria se hizo considerando 120 pasos que corresponden a los meses. Es claro que si se ha de predecir el IPC con una caminata aleatoria se debe hacer en períodos cortos, esto debido a que la varianza de una caminata aleatoria aumenta con el tiempo.

$$D_t = D_{t-1} + \epsilon_t$$

$$var(D_t) = var(D_{t-1}) + var(\epsilon_t)$$

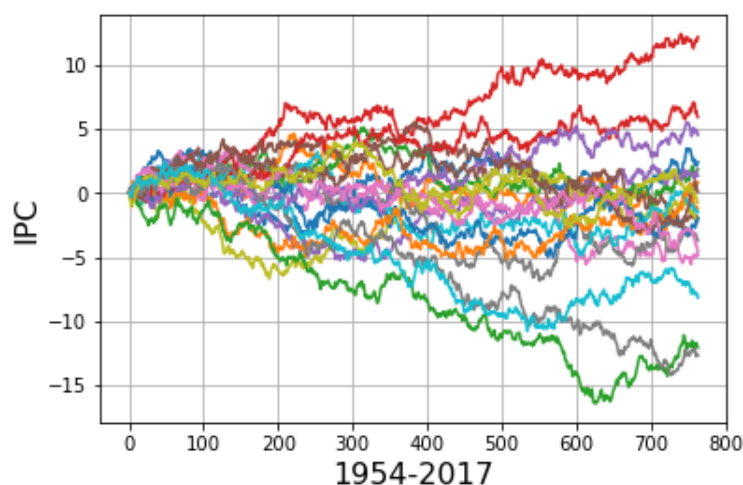
O podemos verlo de la forma

$$D_t = D_0 + \epsilon_1 + \epsilon_2 + \dots + \epsilon_t$$

De donde

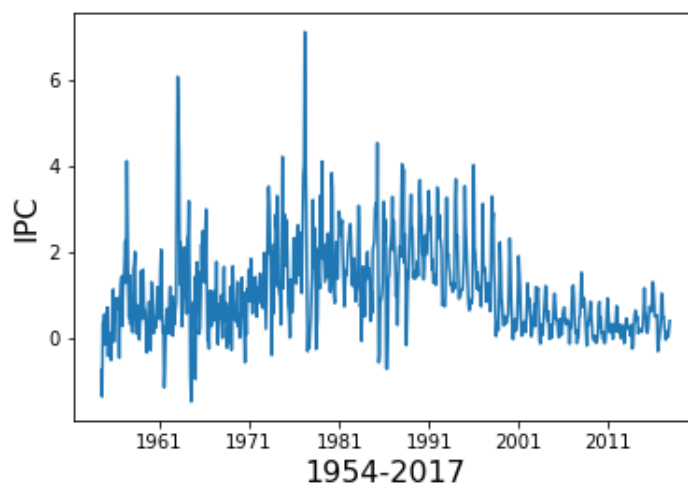
$$\text{var}(D_t) = t\delta^2$$

Este dato se ve reflejado al simular la caminata aleatoria.



**Figura 3-3.:** Simulación de 20 caminatas aleatorias para 764 pasos que corresponde a la cantidad de meses del histórico del IPC dado por el DANE.

Si observamos el histórico de la serie de tiempo del DANE



**Figura 3-4.:** Histórico del IPC del DANE

Notamos que una caminata aleatoria no es un buen modelo. La desviación estándar no aumenta en función del tiempo, determinar tendencias y correlaciones podrían apuntar a un mejor modelo.

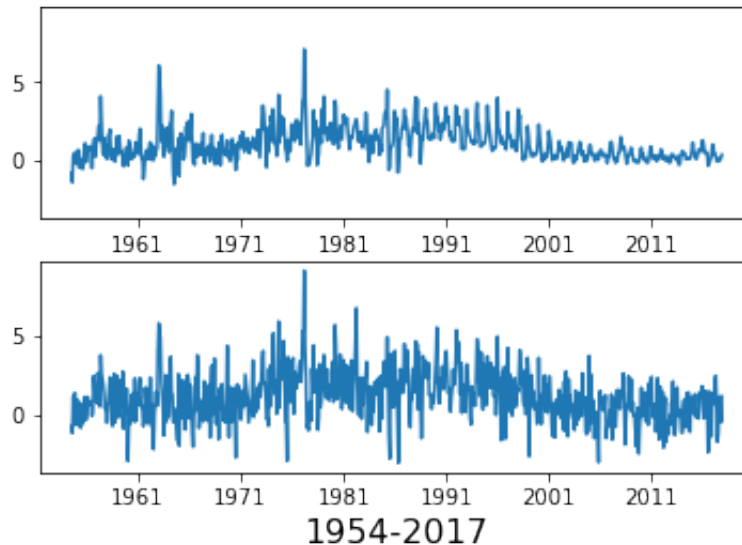
Pero, podríamos preguntarnos, ¿por qué considerar que un buen predictor es una caminata aleatoria?, es evidente que los valores de la serie de tiempo del DANE se dispersan sin explotar alrededor de una media constante, caso contrario a lo que sucede con una caminata aleatoria. La variable del IPC es de interés en el sector económico del País, y para efectos prácticos se desea conocer este valor una o dos semanas antes del valor oficial dado por el Dane, por ello, dado el contexto y las simulaciones hechas, si el IPC es una variable que se comporta de forma aleatoria, es más sensato pensar que si hay un modelo, este depende del último dato dado por el DANE.

En la caminata aleatoria se considera un punto de partida que evolucina a travez del tiempo sin depender del último dato real dado por el Dane, si consideramos este dato tendríamos un  $Ar(1)$  con raíz unitaria que tiene como variable explicativa el último dato dado por el DANE, es decir la única información que nos da es que el mejor pronóstico para mañana es el valor de hoy mas un error modelado por la distribución normal.

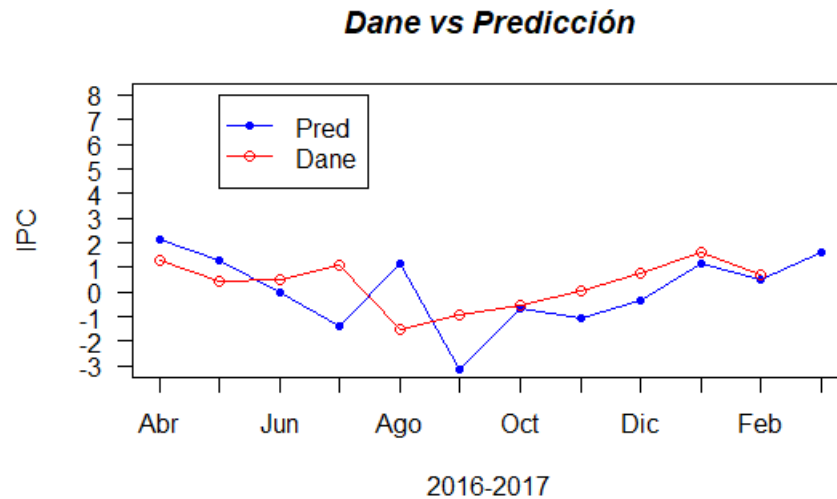
Considerando que se conoce el valor del IPC de alimentos el 5 de cada mes, podría pensarse que el siguiente valor es un valor aleatorio, es decir, si  $(D_t)_t$  es la serie de tiempo histórica del DANE, conocido el valor de un mes dado por el DANE,  $D_t$ , la predicción del siguiente mes sería,

$$D_{t+1}^* = D_t + N(0, \delta_t^2)$$

Donde  $\delta_t$  es la desviación estándar del conjunto de datos  $\{D_T\}_{T \in I}$ , donde  $I$  son las fechas mes a mes desde Enero del año 1954 hasta el mes del año que corresponde al tiempo  $t$ .



**Figura 3-5.:** Arriba, histórico del IPC del Dane, abajo, predicción.



**Figura 3-6.:** Histórico del IPC del Dane vs Predicción del modelo con raíz unitaria entre abril de 2016 y Marzo de 2017.

A pesar de que las predicciones se mantienen en un rango entre -3 y 5, la dispersión de los datos en la predicción es muy alta, consideramos a continuación 12 datos para analizar el comportamiento.

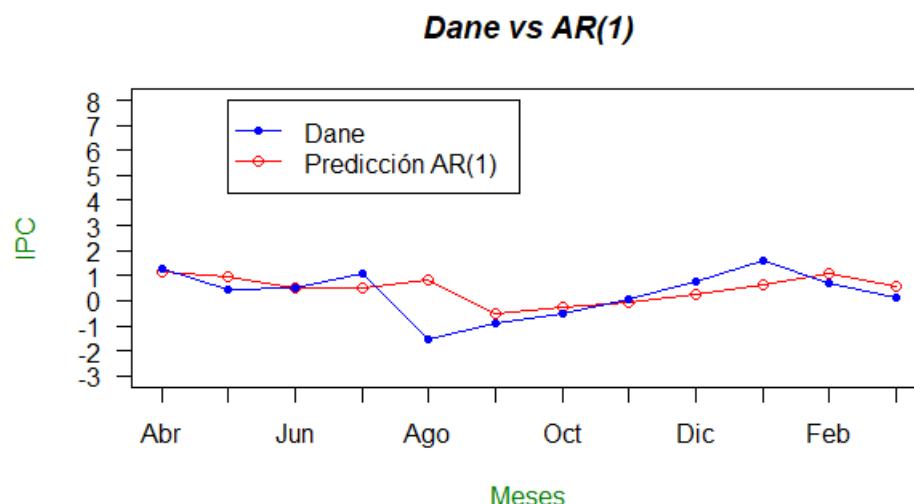
Tomamos 12 datos comprendidos entre los años 2016 y 2017

|            | Abr  | May  | Jun  | Jul   | Ago   | Sep   | Oct   | Nov   | Dic   | Ene  | Feb  | Mar  |
|------------|------|------|------|-------|-------|-------|-------|-------|-------|------|------|------|
| Dane       | 1.26 | 0.46 | 0.50 | 1.11  | -1.54 | -0.91 | -0.53 | 0.05  | 0.78  | 1.62 | 0.71 |      |
| Predicción | 2.13 | 1.31 | 0.01 | -1.37 | 1.18  | -3.15 | -0.63 | -1.05 | -0.32 | 1.17 | 0.53 | 1.59 |

**Tabla 3-1.:** Valores del IPC dados por el DANE y predicción desde Abril del 2016 hasta Marzo de 2017

Tenemos para estos 12 valores un ajuste por mínimos cuadrados de DANE vs la predicción un R cuadrado de 0.12, y un R cuadrado ajustado de 0.02, muy bajos para indicar que los valores de la predicción son un buen ajuste a los valores del DANE.

El modelo expuesto no considera el pasado, es un  $Ar(1)$  y posee una raíz unitaria. Si se considera el ajuste de los datos pasados en un modelo  $Ar(1)$  que considere el mejor ajuste de los datos, no suponiendo que depende de manera aleatoria del último dato dado por el Dane, este podría ser un mejor modelo como veremos a continuación.



**Figura 3-7.:** Histórico del IPC del DANE vs Predicción dada por un modelo AR(1) entre Abril de 2016 Y Marzo de 2017.

|            | Abr  | May  | Jun  | Jul  | Ago   | Sep   | Oct   | Nov   | Dic  | Ene  | Feb  | Mar  |
|------------|------|------|------|------|-------|-------|-------|-------|------|------|------|------|
| DANE       | 1.26 | 0.46 | 0.50 | 1.11 | -1.54 | -0.91 | -0.53 | 0.05  | 0.78 | 1.62 | 0.71 | 0.11 |
| Predicción | 1.12 | 0.93 | 0.48 | 0.5  | 0.84  | -0.52 | -0.25 | -0.07 | 0.23 | 0.61 | 1.07 | 0.59 |

**Tabla 3-2.:** Valores del IPC dados por el DANE y predicción con un modelo AR(1) desde Abril del 2016 hasta Marzo de 2017

El ajuste por mínimos cuadrados de los datos del DANE vs la predicción por un modelo  $AR(1)$  nos da un  $R$  cuadrado de 0.18, y un  $R$  cuadrado ajustado de 0.1, 10 % mejor que el modelo  $AR(1)$  de raíz unitaria. Aún así, son valores bajos, la predicción no es un buen ajuste de los datos.

Este análisis le podría llevar a cabo con los datos desagregados, en cualquier caso, se descarta que su comportamiento sea como el de una caminata aleatoria, es evidente, que si se considera solo la serie de tiempo del IPC, el mejor modelo  $AR(1)$  se ajusta mejor que una caminata aleatoria.

### 3.1.2. Normalidad para las series desagregadas

#### *Test de Anderson-Darling*

Usamos el estadístico de Anderson-Darling para determinar el ajuste de los datos con una

distribución normal. Ver Paul & Brani, (2006). El estadístico AD también se usa para comprobar si una muestra de datos proviene de una población con una distribución especificada. Por ejemplo, en nuestro caso, para comprobar si sus datos cumplen con el supuesto de normalidad para una prueba t.

Las hipótesis para la prueba de Anderson-Darling son:

- $H_0$ : Los datos siguen una distribución normal.
- $H_a$ : Los datos no siguen una distribución normal.

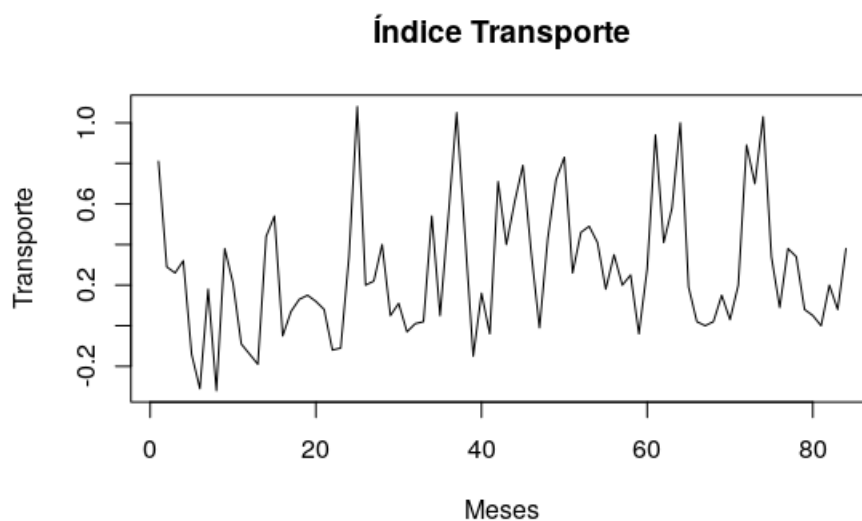
Si el valor p para la prueba de Anderson-Darling es menor que el nivel de significancia seleccionado (por lo general 0.05 o 0.10), concluimos que los datos no siguen la distribución normal.

Se procede a calcular el test de Anderson-Darling para las series de transporte y diversión.

Transporte:  $p$ -valor=0,004665

Diversión:  $p$ -valor=0,000135

Rechazando la hipótesis nula de normalidad, para las otras series el comportamiento del  $p$ -valor es similar, pero entonces ¿Se descarta que los modelos ARIMA permitan establecer un modelo como la suma de una regresión lineal y un proceso aleatorio?. Se hace necesario en esta instancia dar un enfoque al problema desde su contexto.



**Figura 3-8.:** Serie del índice de transporte

Apesar de que la serie de transporte presenta varios picos de forma regular, se puede notar la relevancia de los picos cada 12 datos. En algunos índices como los de educación y salud se puede apreciar de forma clara este comportamiento (ver **3-33** y **3-38**) y es un comportamiento que se encuentra en cada una de las series. Esto es evidencia de un componente estacional, que no se tratara en este trabajo, pero que es relevante resaltar, a pesar de que no se trata, podremos evidenciar que en general se presenta un buen comportamiento de los modelos escogidos que al no considerar el componente estacional, es decir el rezago 12, tendrá una incidencia sobre el análisis de los residuos.

### *Supuesto de normalidad*

En Espinal et al. (2006) y Huwiler & Kaufmann (2003) se calculó el valor del test de la ACF sin establecer la normalidad de los datos, requisito previo para el cálculo de la ACF. Nos preguntamos entonces, ¿qué permite hacer este supuesto?

Vale la pena resaltar que el supuesto de normalidad que se plantean en los dos trabajos no es atípico, en Pedrosa et al. (2015) se menciona; "la realidad es que cuando se lleva a cabo la investigación aplicada, la mayor parte de investigadores emplean pruebas paramétricas, suponiendo habitualmente la normalidad de los datos y sin ningún tipo de comprobación sobre este supuesto".

En matemáticas aplicadas se obtienen algunos modelos teniendo en cuenta algunas conexiones e inevitablemente ignorando otras. En Silviu & Guiasu (1985) se muestran algunos argumentos a favor del problema variacional de la máxima entropía. Si suponemos por ejemplo el experimento de lanzar una moneda, sin tener datos experimentales, la distribución que maximiza la entropía es la uniforme, es decir la distribución de probabilidad de menos supuestos que se le puede atribuir a este sistema es la uniforme.

Sabemos con certeza el valor de la media y la varianza de las series de los índices desagregados, entonces bajo el principio de máxima entropía tenemos que la distribución más acertada con la que se distribuyen los datos es la normal. La distribución con menos supuestos es aquella para la cual la desinformación es máxima. En cierto modo, toda asignación de probabilidades supone la afirmación de algún conocimiento o desconocimiento. Si consideramos la entropía como una medida de la información, entonces deberíamos elegir la distribución de probabilidad que incluye solo la información proporcionada a priori en nuestro problema. En otras palabras, podríamos elegir otra distribución, pero la que maximiza la entropía es la menos informativa o, como dicen algunos, la más honesta con respecto a nuestra ignorancia, y en nuestro caso es la normal.

### 3.1.3. Modelos ARIMA para las series desagregadas

Tradicionalmente el proceso de construcción de un modelo temporal se basa en la utilización de un conjunto de test de hipótesis, haciendo uso de estos se determina la correlación que la serie guarda con su pasado, se escogen los lag- $\ell$  más significativos y se usan para modelar la serie de tiempo con el mejor modelo  $AR(p)$ .

Una forma sencilla que permite la identificación de que el modelo a plantear es un  $AR(p)$  puro se siguen los siguientes pasos, el PACF presenta los  $p$  primeros valores distintos de cero o muy próximos a cero con un comportamiento sinusoidal, la ACF presenta un decrecimiento exponencial y un comportamiento sinusoidal. De igual forma, el patrón que debe seguir ACF y PACF para la identificación del orden de un modelo puro  $MA(q)$  es el siguiente, la ACF presenta los primeros  $q$  valores distintos de cero y el resto de valores son cero o muy próximos a cero con un comportamiento sinusoidal, y la PACF presenta decrecimiento exponencial y un comportamiento sinusoidal. (Guerrero, 2003).

Todos los análisis y gráficas siguientes se realizaron con programa R versión 3.5.2.

## COMPONENTE ESTACIONAL

Como se menciona en la sección de normalidad, el contexto tiene un impacto que de antemano prevemos y se evidenciará en los correlogramas, que es la incidencia del rezago 12. Esta se da debido al ajuste de la economía comenzando cada año, el cambio del salario mínimo, del IVA, que incide de forma significativa en el cambio del IPC. Además para el índice de alimentos tenemos una gran incidencia de las estaciones, como lluvias o sequías, hacemos énfasis en esta serie dado que es la que presenta la mayor volatilidad. A pesar de ello considerando los criterios establecidos e indicadores se escogerán los modelos. Este proceso obedece a que esperamos poder expresar la serie como la suma de un proceso que se puede predecir a partir del pasado de la serie y un proceso aleatorio que será ruido blanco. Y se espera que los correlogramas de los residuales permanezcan dentro de los límites de aceptación de la hipótesis de no correlación.

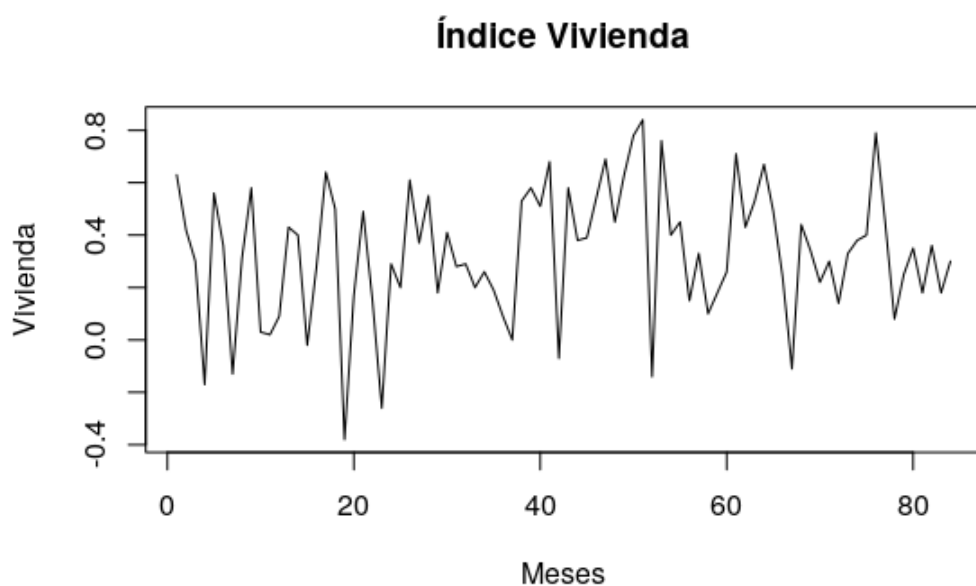
## SIGNIFICANCIA DE LOS PARÁMETROS

Para determinar si la asociación entre la respuesta y cada término del modelo es estadísticamente significativa, compararemos el valor  $p$  con su nivel de significancia para evaluar la hipótesis nula. La hipótesis nula establece que el término no es significativamente diferente de cero, indicando que no existe una buena asociación entre el término y la respuesta. Se establece como un buen nivel de significancia el valor  $p$  de 0.05. Establecemos el test de Wald para determinar la significancia de los parámetros con,

- $H_0$ : Si el valor  $p$  es mayor que o igual al nivel de significancia, concluimos que el coeficiente no es estadísticamente significativo.
- $H_a$  : Si el valor  $p$  es menor que o igual al nivel de significancia, concluimos que el coeficiente es estadísticamente significativo.

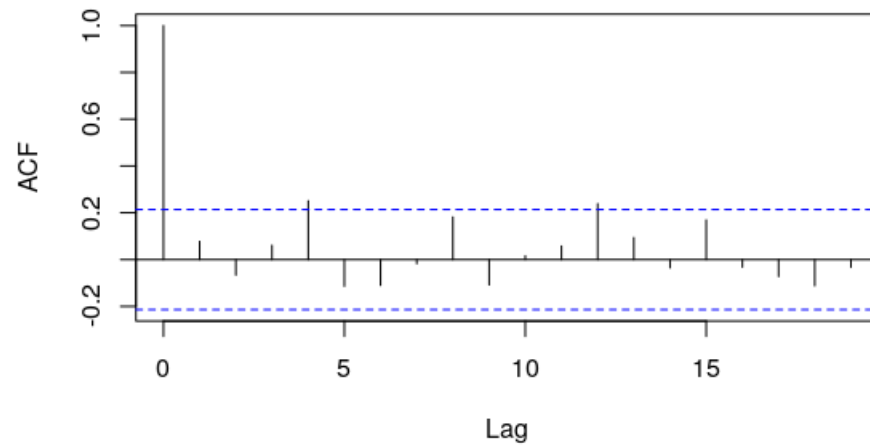
### *Serie de tiempo Vivienda*

Consideramos la serie de tiempo de Vivienda para el primer análisis.

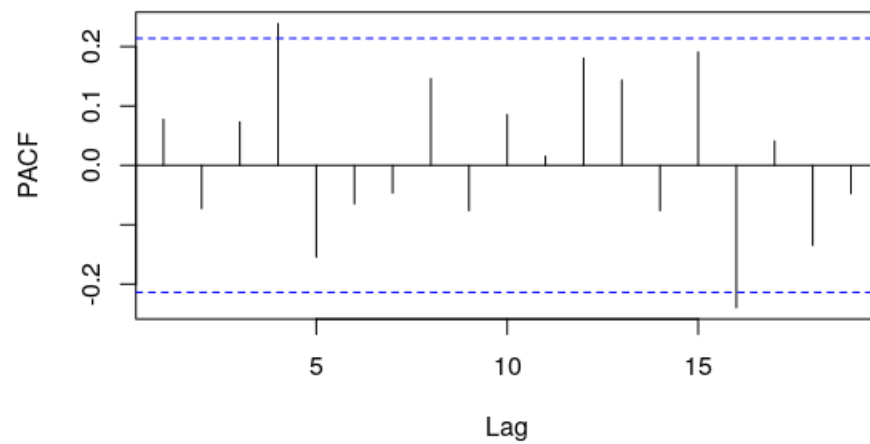


**Figura 3-9.:** Serie de tiempo del índice de vivienda, entre el 1 de Enero de 2012 hasta el 12 de Diciembre de 2018.

Considerando la figura (3-9), la serie no presenta una tendencia marcada, ni estacionalidad, se podría pensar que la serie es estacionaria, para ello tendremos en cuenta la ACF y la PACF.



**Figura 3-10.:** Función de autocorrelación del índice de vivienda.



**Figura 3-11.:** Función de autocorrelación parcial del índice de vivienda.

En Colombia, como en otros países, existe una escasa disponibilidad de información en los precios de la vivienda. En gran medida, esto se debe a las características particulares del bien en cuestión. De hecho, no existen estándares internacionales que establezcan la mejor metodología de construcción de precios en este mercado. (Salazar et al., 2013). En el primer trimestre del 2019 el índice de precios de vivienda registró una variación del 2,28 % en comparación con el cuarto trimestre del 2018. Esta tasa es superior en 0,18 puntos porcentuales al registrado en el mismo trimestre del año anterior. (DANE, 2019).

La poca variabilidad en el informe presentado por el DANE (2019), hace que sea natural la heterogeneidad y es consecuente con la ACF y la PACF mostradas en las figura (3-10) y (3-11).

Considerando el lag-4 en la ACF proponemos el modelo AR(4), la PACF nos indica por el lag-3, que supera el umbral del error estándar, que se deben considerar  $r_{t-1}$ ,  $r_{t-2}$ ,  $r_{t-3}$  como variables explicativas. Incluimos la cuarta variable explicativa dada la significancia del lag-4 en la ACF. Por otro lado los lags que superan el umbral lo hacen con mucho menos que una décima, es decir, no son significativos, indicando que es posible que el mejor modelo sea un ARIMA(0,0,0). Se comparan los indicadores y criterios de determinación de un ARIMA(4,0,0) y un ARIMA(0,0,0) para seleccionar el mejor modelo.

Comparación de los modelos en la ventana de seis meses.

|              | AIC   | BIC   | RMSE   | MAE   | MAPE |
|--------------|-------|-------|--------|-------|------|
| ARIMA(0,0,0) | 13.62 | 18.33 | 0.257  | 0.205 | Inf  |
| ARIMA(4,0,0) | 15.09 | 29.23 | 0.2462 | 0.198 | Inf  |

**Tabla 3-3.:** Modelos del mes de Junio de 2018.

El Inf en el MAPE se debe a que el valor real  $r_t$  es menor que el épsilon de la maquina en comparación a la diferencia  $r_t - \hat{r}_t$

|              | AIC   | BIC    | RMSE  | MAE   | MAPE |
|--------------|-------|--------|-------|-------|------|
| ARIMA(0,0,0) | 12.85 | 17.59  | 0.255 | 0.204 | Inf  |
| ARIMA(4,0,0) | 14.29 | 28.508 | 0.245 | 0.197 | Inf  |

**Tabla 3-4.:** Modelos del mes de Julio de 2018.

Los indicadores RMASE y MAE son similares para cada mes, mientras que el AIC y el BIC tienen una deferencia notable siendo los del ARIMA(0,0,0) menores en todos los meses, indicando que este es el mejor modelo para serie de Vivienda. Ver Tabla 3-3 a la Tabla 3-8.

|              | AIC   | BIC    | RMSE  | MAE   | MAPE |
|--------------|-------|--------|-------|-------|------|
| ARIMA(0,0,0) | 11.96 | 16.96  | 0.254 | 0.201 | Inf  |
| ARIMA(4,0,0) | 13.54 | 28.508 | 0.243 | 0.196 | Inf  |

**Tabla 3-5.:** Modelos del mes de Agosto de 2018.

|              | AIC   | BIC    | RMSE  | MAE   | MAPE |
|--------------|-------|--------|-------|-------|------|
| ARIMA(0,0,0) | 11.42 | 16.21  | 0.253 | 0.201 | Inf  |
| ARIMA(4,0,0) | 13.08 | 28.508 | 0.243 | 0.195 | Inf  |

**Tabla 3-6.:** Modelos del mes de Septiembre de 2018.

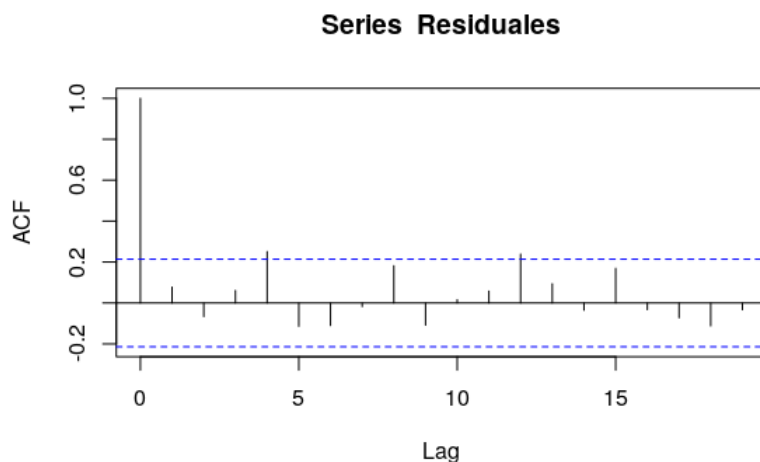
|              | AIC   | BIC    | RMSE  | MAE   | MAPE |
|--------------|-------|--------|-------|-------|------|
| ARIMA(0,0,0) | 10.67 | 15.33  | 0.251 | 0.199 | Inf  |
| ARIMA(4,0,0) | 12.27 | 26.711 | 0.241 | 0.194 | Inf  |

**Tabla 3-7.:** Modelos del mes de Octubre de 2018.

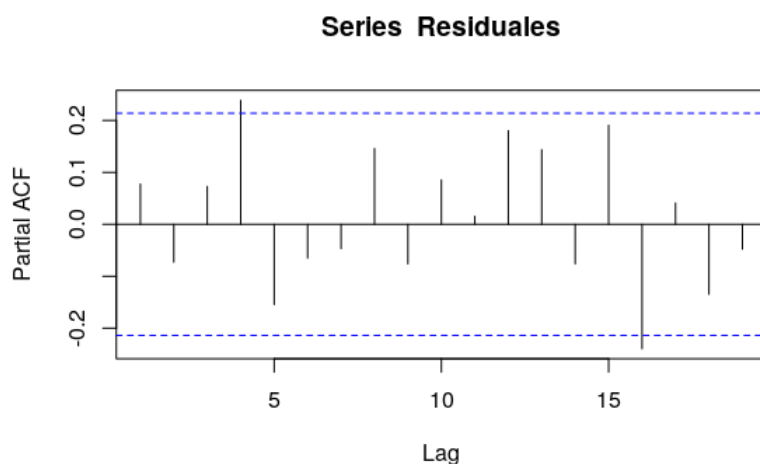
|              | AIC   | BIC   | RMSE  | MAE   | MAPE |
|--------------|-------|-------|-------|-------|------|
| ARIMA(0,0,0) | 9.96  | 14.79 | 0.251 | 0.198 | Inf  |
| ARIMA(4,0,0) | 11.62 | 26.13 | 0.241 | 0.194 | Inf  |

**Tabla 3-8.:** Modelos del mes de Noviembre de 2018.

Tomando como modelo para la serie de vivienda un  $ARIMA(0,0,0)$ . Observe que al no tener coeficiente autorregresivo, no aplicamos el DF-test. Procedemos a ver el trazado del ACF y PACF de los residuos del modelo.



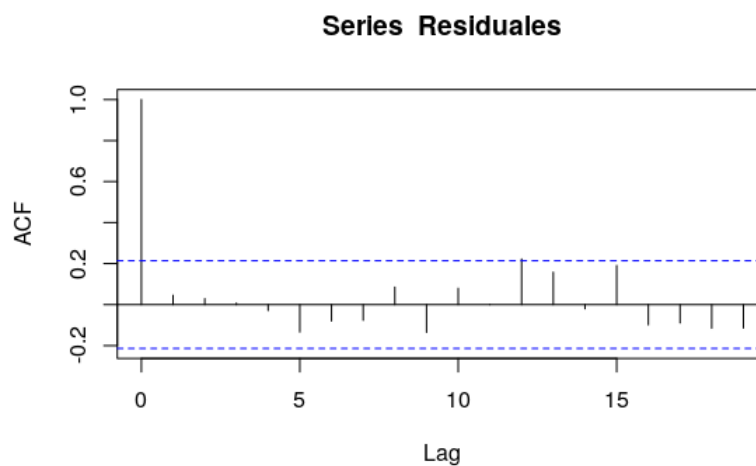
**Figura 3-12.:** ACF de los residuales del modelo  $ARIMA(0,0,0)$  de vivienda



**Figura 3-13.:** PACF de los residuales del modelo  $ARIMA(0,0,0)$  de vivienda

En el gráfico de la ACF observamos que el lag-4 y el lag-12 pasan por un orden de menos del 0.1 el umbral del error estándar. El lag-4 evidencia el impacto del valor autorregresivo de orden 4 que no se consideró, el lag-12 por la naturaleza estacional de la serie que se da forma anual.

A pesar de que el ACF de los residuales del modelo  $ARIMA(4,0,0)$  muestra el escenario ideal para un modelo por la no correlación de estos, si hacemos el promedio de los residua-



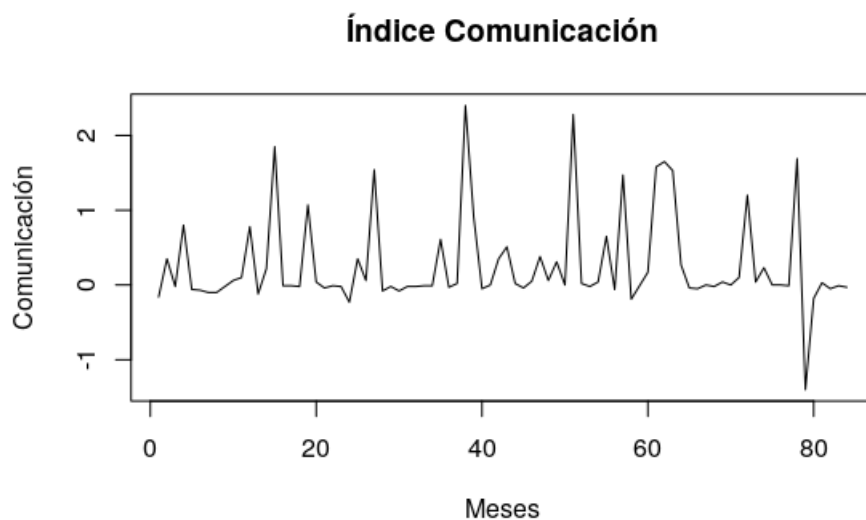
**Figura 3-14.:** ACF de los residuales del modelo ARIMA(4,0,0) de vivienda.

les del modelo ARIMA(0,0,0) que da un valor de 0,196 y comparamos con el promedio de los residuales del modelo ARIMA(4,0,0) que es 0,192 tenemos que al diferencia es 0,004, es decir que aumentar la cantidad de variables no mejora la capacidad predictiva del modelo.

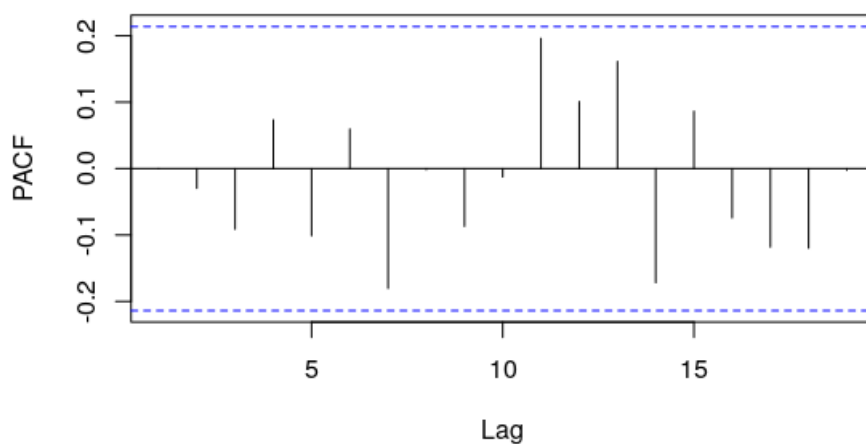
Significancia de los parámetros.

Esta serie se ajusta a la media como lo establecimos, un valor  $p$  de  $2,2 \times 10^{-16}$ .

*Serie de tiempo Comunicación.*

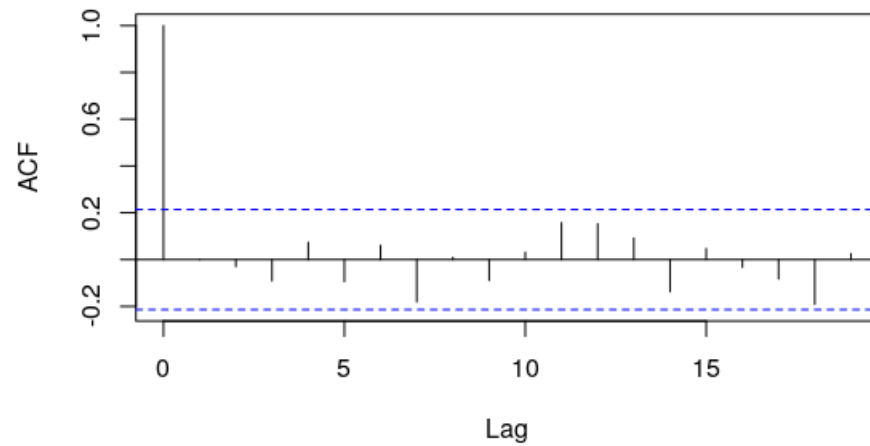


**Figura 3-15.:** Serie de tiempo del índice de comunicación, entre el 1 de enero de 2012 hasta el 12 de diciembre de 2018



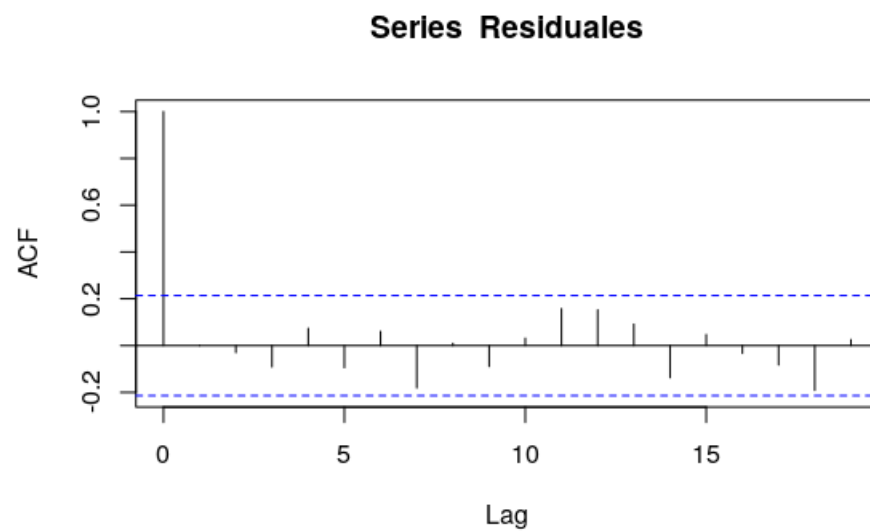
**Figura 3-17.:** Función de autocorrelación parcial del índice de comunicación.

La serie de comunicación no presenta tendencia, ni lags en la AFC y PACF que pasen el umbral del error estándar, con un AIC=161.59, y BIC=166.45. El modelo que mejor describe este comportamiento es un ARIMA(0,0,0):  $r_t = 0,2668$ .

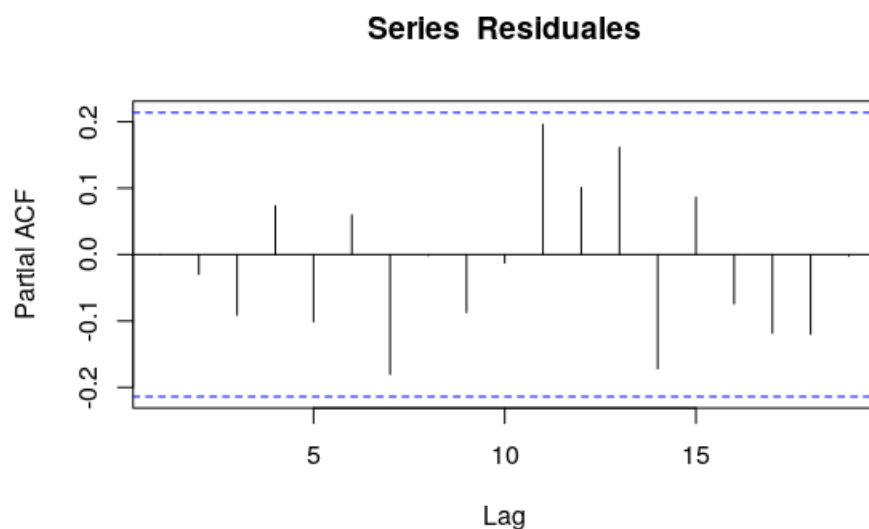


**Figura 3-16.:** Función de autocorrelación del índice de comunicación

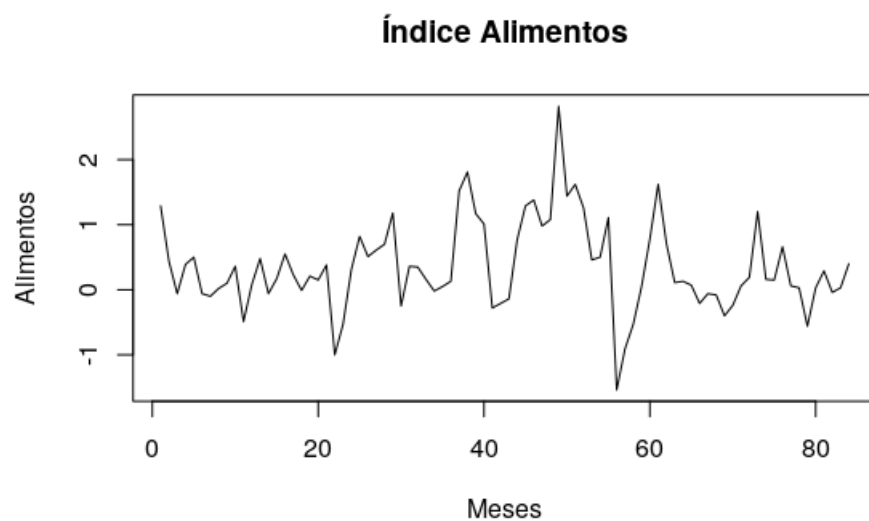
Pasamos a ver la independencia de los residuos, para ello vemos el comportamiento de la ACF y la PACF.



**Figura 3-18.:** Función de autocorrelación de los residuos del modelo ARIMA(0,0,0) del índice de comunicación.



**Figura 3-19.:** Función de autocorrelación parcial de los residuos del modelo ARIMA(0,0,0) del índice de comunicación.

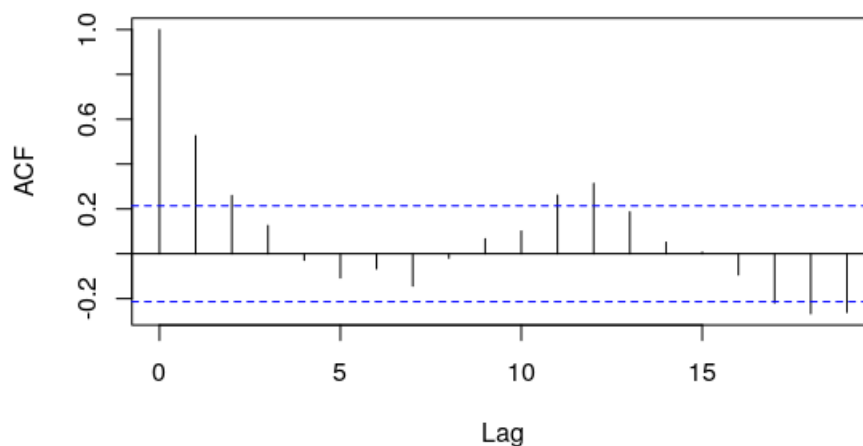


**Figura 3-20.:** Serie de tiempo del índice de alimentos, entre el 1 de Enero de 2012 hasta el 12 de Diciembre de 2018

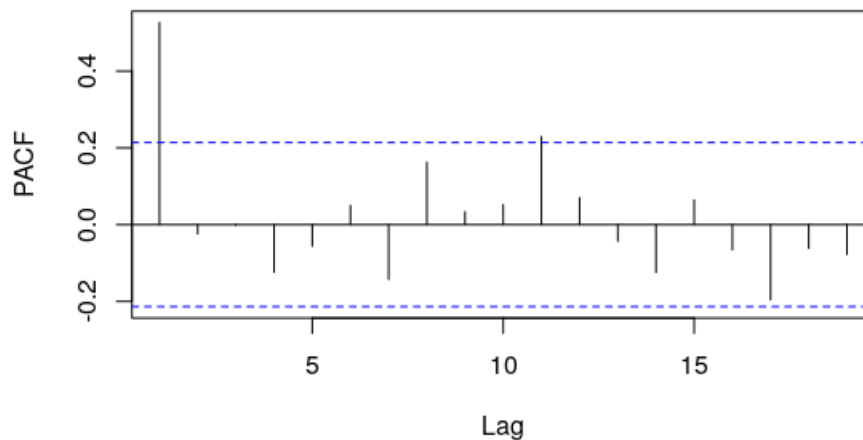
Como podemos ver la serie de los residuales son ruido blanco, es decir que el modelo ARIMA(0,0,0) establecido es un buen modelo para los datos.

***Serie de tiempo Alimentos.***

Procedemos a hacer un análisis de la serie de tiempo de Alimentos, ver fig (3-20).



**Figura 3-21.:** Función de autocorrelación índice de Alimentos



**Figura 3-22.:** Función de autocorrelación parcial del índice de Alimentos

Se tiene un valor del  $\text{lag-1} = 0,527$ , y  $(\text{lag-1})^2 = 0,277 \approx \text{lag-2} = 0,26$ ,  $(\text{lag-1})^3 = 0,146 \approx \text{lag-3} = 0,126$ , es decir, los primeros valores decrecen exponencialmente, en conjunto con una

PACF que muestra evidentemente un lag-1 pasando el umbral del error estándar seguido de valores que son cero, indicando que el AR(1) es un buen modelo. El anterior análisis no considera el efecto que tiene el lag-11 y lag-12 en la ACF sobre la serie y el lag-11 de la función PACF pasando el umbral del error estándar indicando la incidencia directa que tiene  $r_t \rightarrow r_{t+11}$ , postulamos un ARIMA(12,0,0) como un modelo. Procedemos a hacer la comparación entre los dos modelos considerando los indicadores y criterios de determinación.

|               | AIC    | BIC    | RMSE  | MAE   | MAPE   |
|---------------|--------|--------|-------|-------|--------|
| ARIMA(1,0,0)  | 143.9  | 150.97 | 0.584 | 0.41  | 206.44 |
| ARIMA(12,0,0) | 154.56 | 196.48 | 0.539 | 0.369 | 178.55 |

**Tabla 3-9.:** Modelos mes de junio 2018

|               | AIC    | BIC    | RMSE  | MAE   | MAPE   |
|---------------|--------|--------|-------|-------|--------|
| ARIMA(1,0,0)  | 146.32 | 153.43 | 0.587 | 0.414 | 200.53 |
| ARIMA(12,0,0) | 151    | 196.48 | 0.582 | 0.403 | 206.32 |

**Tabla 3-10.:** Modelo mes de julio 2018

|               | AIC    | BIC    | RMSE  | MAE   | MAPE   |
|---------------|--------|--------|-------|-------|--------|
| ARIMA(1,0,0)  | 147.16 | 154.3  | 0.583 | 0.411 | 205.63 |
| ARIMA(12,0,0) | 159.38 | 192.73 | 0.545 | 0.377 | 187.27 |

**Tabla 3-11.:** Modelo mes de agosto 2018

|               | AIC    | BIC    | RMSE  | MAE   | MAPE    |
|---------------|--------|--------|-------|-------|---------|
| ARIMA(1,0,0)  | 143.94 | 155.12 | 0.58  | 0.407 | 204.062 |
| ARIMA(12,0,0) | 152.63 | 192.73 | 0.575 | 0.398 | 212.86  |

**Tabla 3-12.:** Modelo mes de septiembre 2018

|               | AIC    | BIC     | RMSE  | MAE   | MAPE    |
|---------------|--------|---------|-------|-------|---------|
| ARIMA(1,0,0)  | 149.09 | 165.31  | 0.578 | 0.406 | 210.784 |
| ARIMA(12,0,0) | 161.21 | 194.903 | 0.541 | 0.373 | 193.211 |

**Tabla 3-13.:** Modelo mes de octubre 2018

La raíz del error cuadrático medio, el error medio absoluto (MAE) y el error medio porcentual absoluto (MAPE) son menores en los seis meses en el ARIMA(12,0,0), no se compensa dada la poca diferencia, la cantidad de variables explicativas, y la diferencia notable entre el AIC y el BIC, siendo menor en el ARIMA(1,0,0) así el mejor modelo es AR(1,0,0).

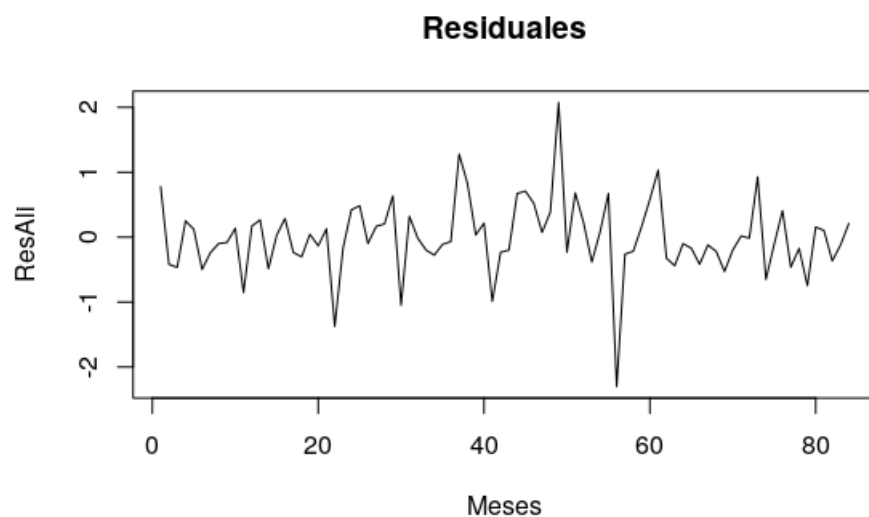
Modelo que predice el mes de diciembre,

$$r_t = 0,5325r_{t-1} + 0,3666$$

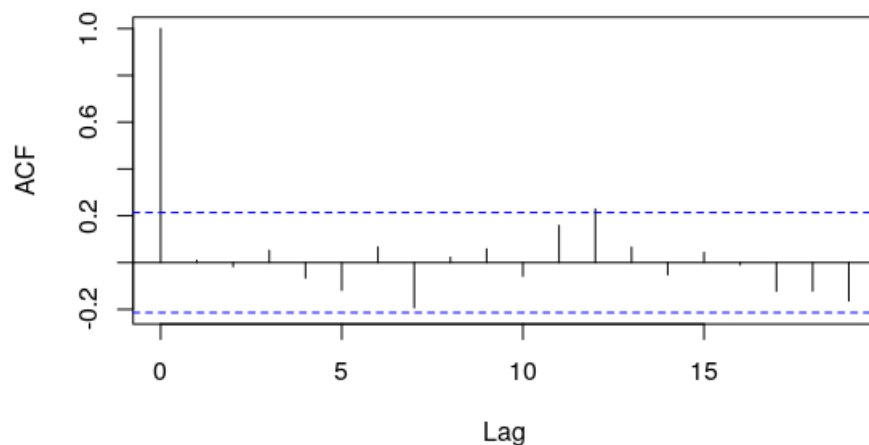
Consideramos el modelo AR(1) planteado, procedemos a hacer el análisis de los residuales.

|               | AIC    | BIC     | RMSE  | MAE   | MAPE    |
|---------------|--------|---------|-------|-------|---------|
| ARIMA(1,0,0)  | 149.86 | 157.12  | 0.574 | 0.403 | 212.292 |
| ARIMA(12,0,0) | 161.84 | 194.903 | 0.537 | 0.369 | 193.615 |

**Tabla 3-14.:** Modelo mes de noviembre 2018

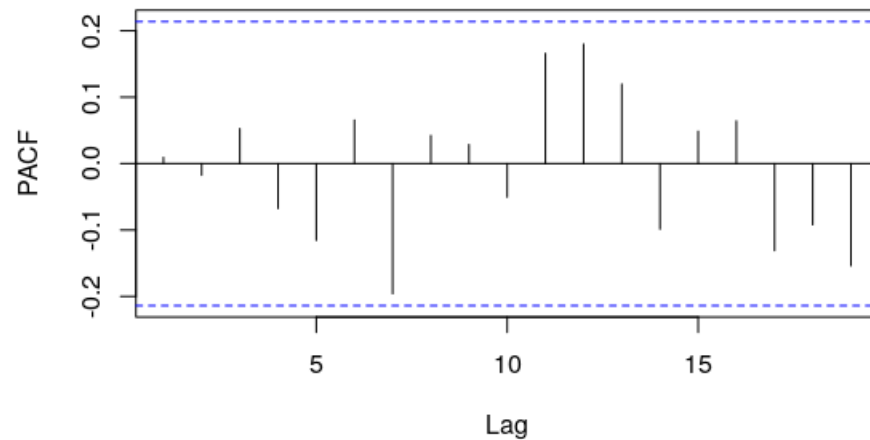


**Figura 3-23.:** Residuales del modelo AR(1)



**Figura 3-24.:** Función de autocorrelación de los residuales del modelo AR(1) planteado para Alimentos.

El gráfico del PACF de los residuales no muestra correlación, mientras que en los primeros 11 lags la AFC de los residuales no presentan un patrón discernible, pero podemos observar en la ACF de la serie de alimentos (**3-21**), junto al de los residuales (**3-23**), que hay una persistencia en el lag-12, pero esta esta por menos un valor menor a 0,01 del error estándar, y la PACF nos nos indica que no hay incidencia de  $r_{t+12}$  sobre  $r_t$ , teniendo así un compor-



**Figura 3-25.:** Función de autocorrelación parcial de los residuales del modelo AR(1) planteado para Alimentos.

tamiento aceptable de los residuales.

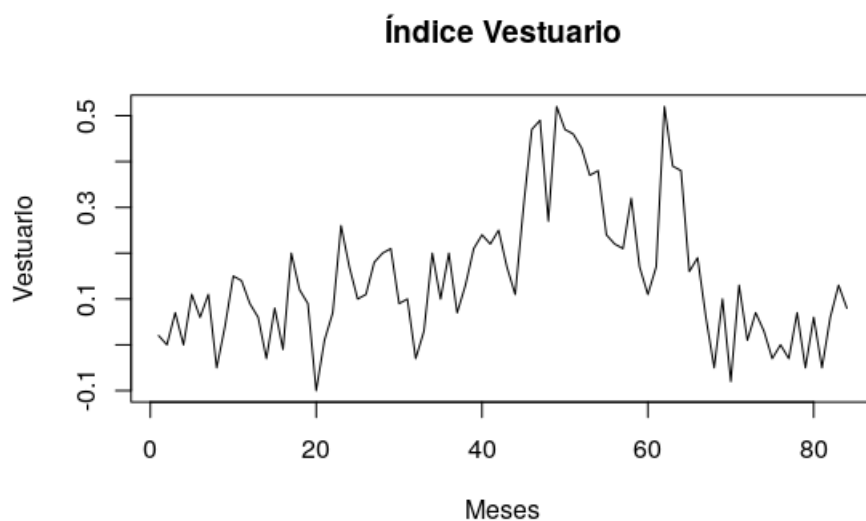
Significancia de los parametros:

Intercepto:  $p\text{-valor}=0,005384$

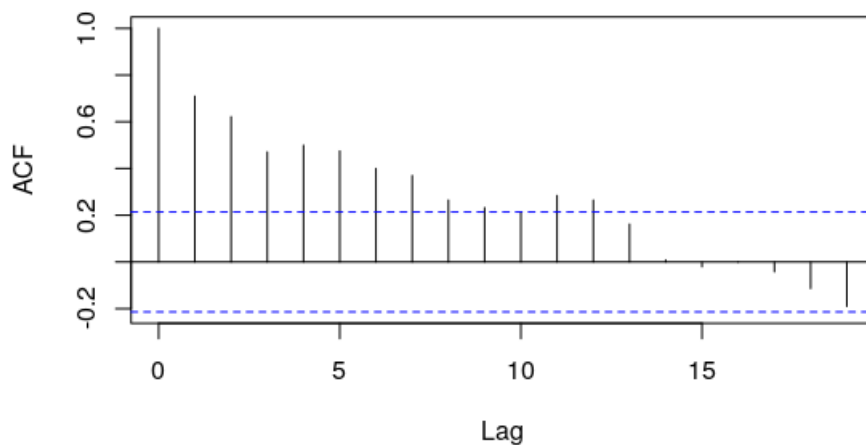
Parámetro AR1:  $p\text{-valor}=8,19 \times 10^{-9}$ .

### *Serie de tiempo Vestuario.*

La media no es constante, se observa una tendencia marcada que crece hasta el dato 50, y después decrece hasta el dato 70, ver fig (3-26). Se procede a analizar la FAC.

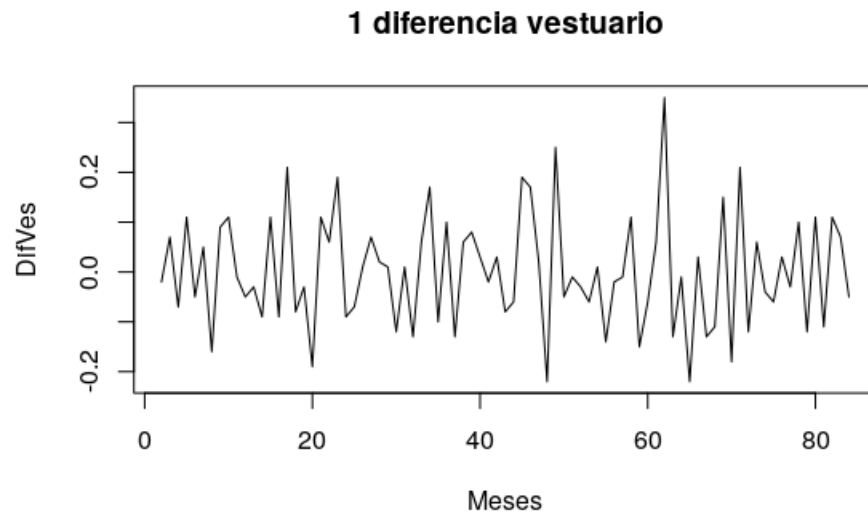


**Figura 3-26.:** Serie de tiempo del índice de vestuario, entre el 1 de Enero de 2012 hasta el 12 de Diciembre de 2018.

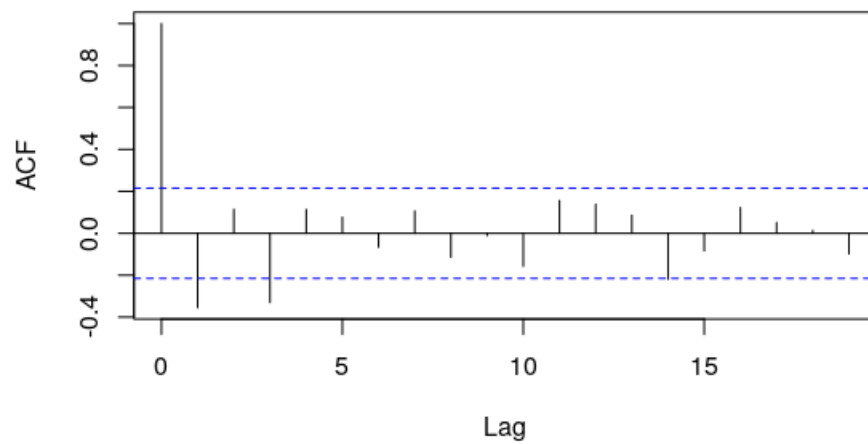


**Figura 3-27.:** ACF del índice de vestuario

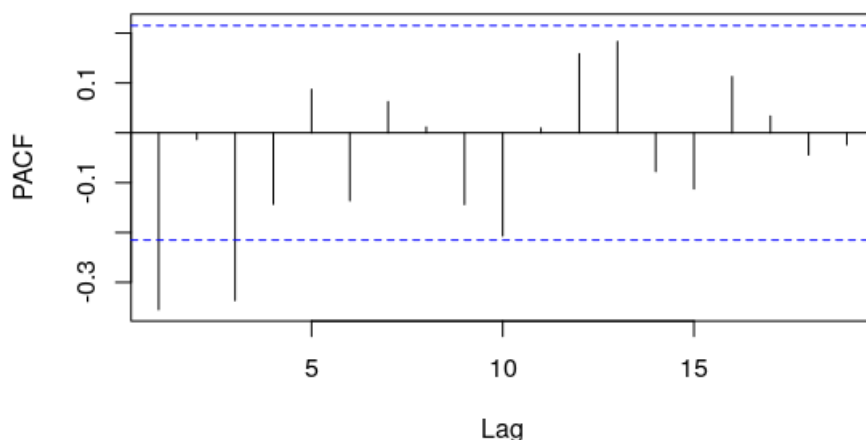
Los primeros lags son elevados con decrecimiento muy lento, característica de las series no estacionarias. Se aplica una diferenciación de orden 1 a la serie, la primera diferencia homogeniza notablemente la serie, se procede a calcular la ACF y PACF para identificar los parámetros del modelo ARIMA.



**Figura 3-28.:** Serie de la primera diferencia del índice de vestuario



**Figura 3-29.:** ACF de la primera diferencia del índice de vestuario



**Figura 3-30.:** PACF de la primera diferencia del índice de vestuario

Considerando esta información podemos plantear modelos

Modelo ARIMA(0,1,3) debido a que ACF presenta el lag-1 y el lag-3 significativos, y los valores cercanos a estos casi cero, el lag-3 en el PACF superando el error estándar, indica la incidencia de  $a_t$  sobre  $a_{t+3}$ . El modelo al que se ajusta la serie presenta la forma

$$(1 - B)r_t = c + (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3)a_t$$

Al determinar modelos considerando los criterios de AIC, BIC, y los indicadores RMSE, MAE y MAPE, del modelo ARIMA(0,1,4) son muy parecidos al del modelo ARIMA(0,1,3).

|              | AIC     | BIC      | RMSE    | MAE     | MAPE |
|--------------|---------|----------|---------|---------|------|
| ARIMA(0,1,3) | -141.1  | -131.421 | 0.09763 | 0.07594 | Inf  |
| ARIMA(0,1,4) | -142.65 | -130.56  | 0.00103 | 0.09545 | Inf  |

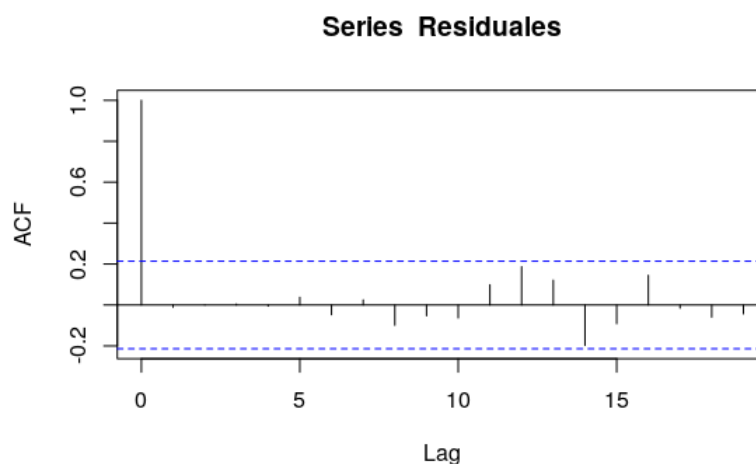
**Tabla 3-15.:** Comparación criterios del modelo ARIMA(0,1,3) vs el modelo ARIMA(0,1,4) para la predicción del mes de diciembre.

Aunque el lag-4 no supera y por mucho la línea significativa en ACF, en el PACF no se observa un decaimiento exponencial después del lag-3, lo que significa que, el ARIMA(0,1,3) es por poco el mejor modelo. Esta tendencia se presenta en los seis meses.

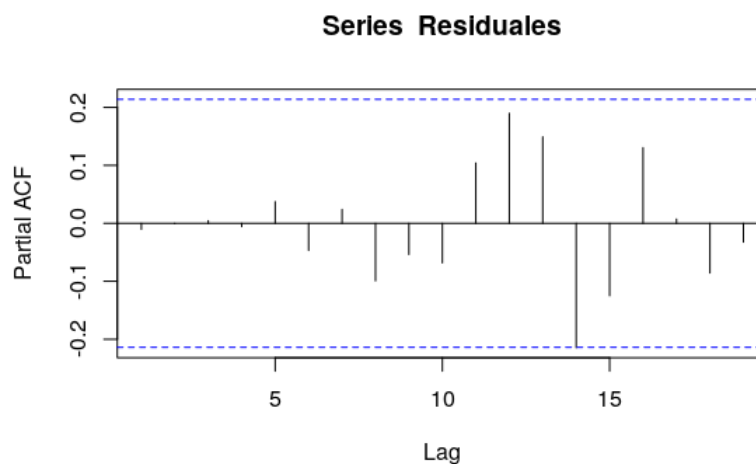
Modelo a tener en cuenta para el índice de vestuario que predice el mes de diciembre.

$$(1 - B)r_t = (1 + 0,3139B + 0,0160B^2 + 0,3476B^3)a_t$$

Procedemos a comprobar si los residuales del modelo son ruido blanco.



**Figura 3-31.:** ACF de los residuales del modelo ARIMA(0,1,3) para el índice de vestuario.



**Figura 3-32.:** PACF de los residuales del modelo ARIMA(0,1,3) para el índice de vestuario.

El ACF y la PACF muestran que los residuales del modelo ARIMA(0,1,3) en efecto son ruido blanco.

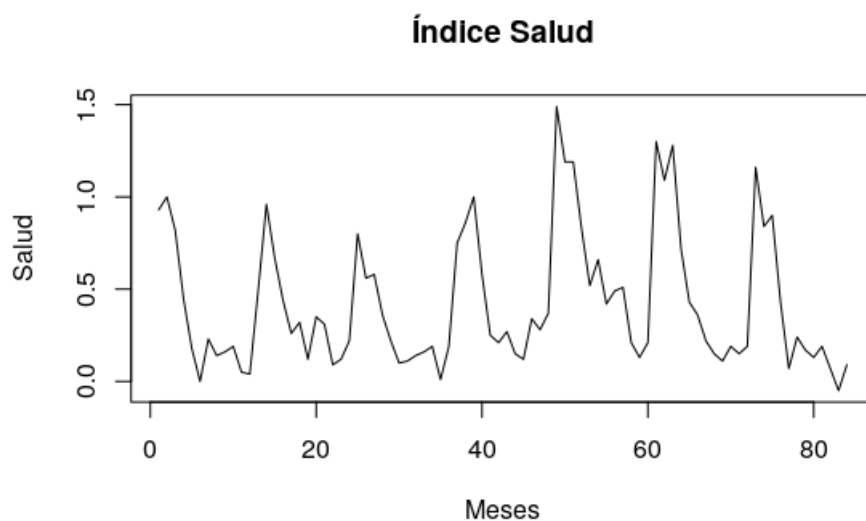
Significancia de los parámetros:

Parámetro MA1:  $p\text{-valor}=0,2254861$

Parámetro MA2:  $p\text{-valor}=0,0004842$

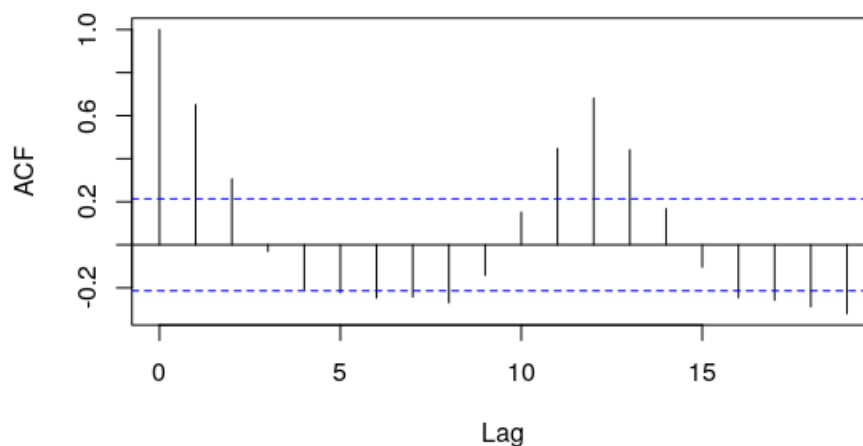
Parámetro MA3:  $p\text{-valor}=4,909 \times 10^{-5}$ .

### *Serie de tiempo Salud.*

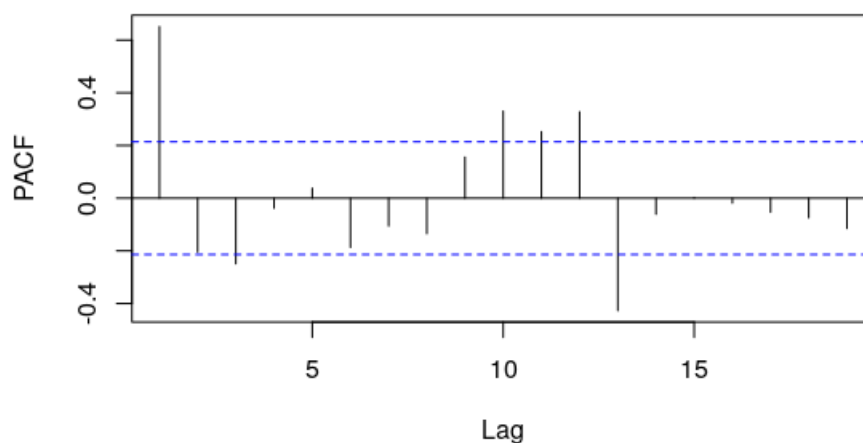


**Figura 3-33.:** Serie de tiempo del índice de salud, entre el 1 de Enero de 2012 y el 12 de Diciembre de 2018.

La serie de tiempo de salud evidencia un comportamiento estacional y la ACF de forma alterna presenta picos positivos y negativos. Estos picos no están decayendo a cero. Tal patrón es la firma del diagrama de autocorrelación de un modelo sinusoidal. (Kareem, 2005), y de posible raíces unitarias en los modelos  $AR(p)$ . Considerando el enfoque teórico planteado en este trabajo. Se tienen dos modelos AR para los cuales la prueba de ADF rechaza la hipótesis nula de no estacionariedad, la prueba ADF para el modelo  $AR(2)$  da un valor  $p$  de 0.02 y para el modelo  $AR(3)$  un valor  $p$  de 0.021.



**Figura 3-34.:** ACF de la serie de salud

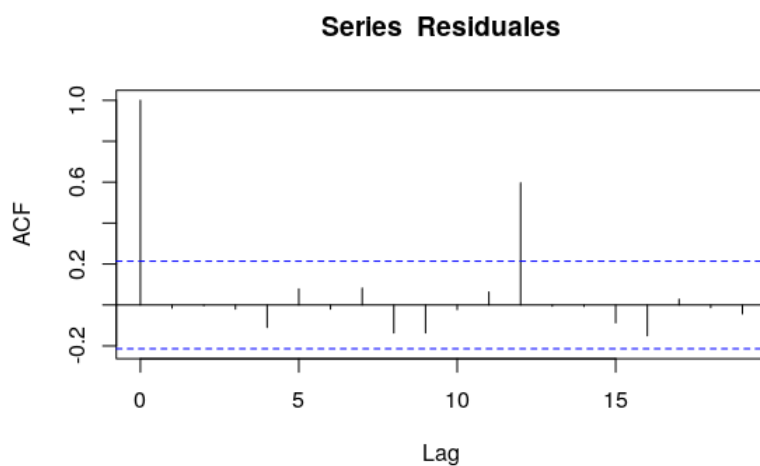


**Figura 3-35.:** PACF de la serie de salud

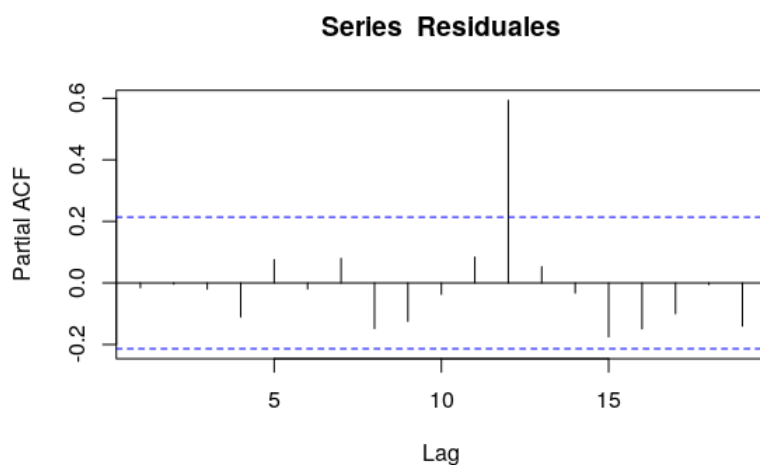
el decaimiento exponencial en los primeros lags, indica que un modelo AR(2) puede describir los datos, pero, en el PACF el valor del lag-3 indica que  $r_t$  tiene una influencia directa en  $r_{t+3}$ . Procedemos a comparar los modelos AR(2) y AR(3) para el mes de diciembre. El modelo ARIMA(3,0,0) supera al ARIMA(2,0,0) en los criterios e indicadores. Procedemos a hacer la ACF y PACF de los residuales del modelo AR(3).

|       | AIC   | BIC    | RMSE    | MAE       | MAPE |
|-------|-------|--------|---------|-----------|------|
| AR(2) | 24.69 | 34.408 | 0.2662  | 0.1900644 | Inf  |
| AR(3) | 20.88 | 33.03  | 0.25683 | 0.1808662 | Inf  |

**Tabla 3-16.:** Comparación criterios del modelo AR(2) vs el modelo AR(3) de salud en el mes de noviembre de 2018



**Figura 3-36.:** ACF de los residuales del modelo ARIMA(3,0,0).



**Figura 3-37.:** PACF de los residuales del modelo ARIMA(3,0,0).

El IPC-salud, que se deriva de los cálculos del DANE, mide la variación de los precios de los productos de salud representativos de los hogares colombianos. Como productos representati-

vos se entienden los comúnmente usados en los hogares, tales como anteojos, antidiabéticos, antiinflamatorios, preservativos, consultas médicas, seguros médicos, etc. Estos productos son parte del grupo salud, de la canasta básica familiar. La mayor incidencia el IPC-salud sucede a principios de cada año, en consecuencia con la ACF y PACF de residuales que evidencia el no considerar el regresor de orden 12 y la incidencia de  $r_t$  sobre  $r_{t+12}$ , escojemos para la serie salud el modelo ARIMA(3,0,0).

$$r_t = 0,4227 + 0,7386r_{t-1} + 0,0059r_{t-2} - 0,2608r_{t-3}$$

Significancia de los parámetros:

Intercepto:  $p\text{-valor}=1,412 \times 10^{-12}$

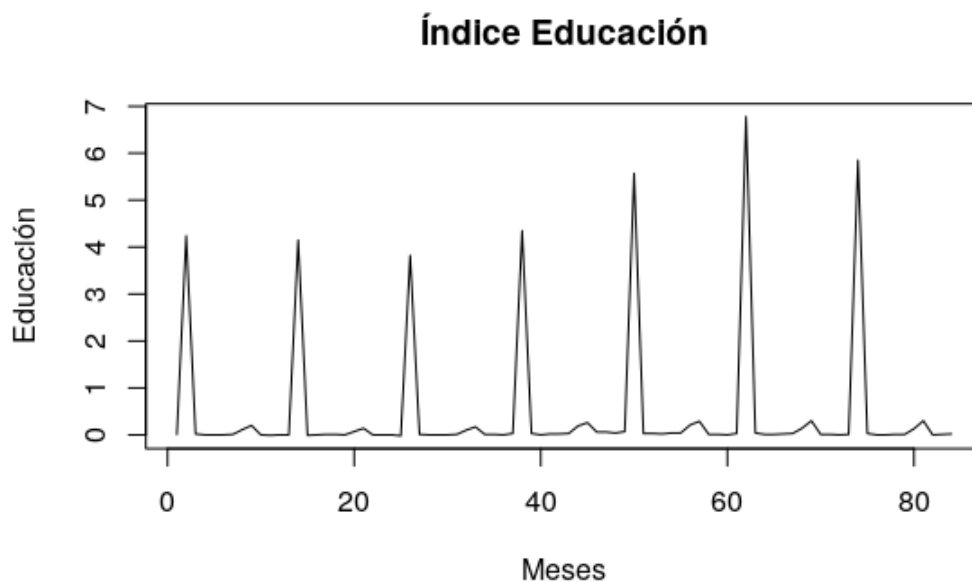
Parámetro AR1:  $p\text{-valor}=0,96427$

Parámetro AR2:  $p\text{-valor}=0,01393$

Parámetro AR3:  $p\text{-valor}=,175 \times 10^{-15}$

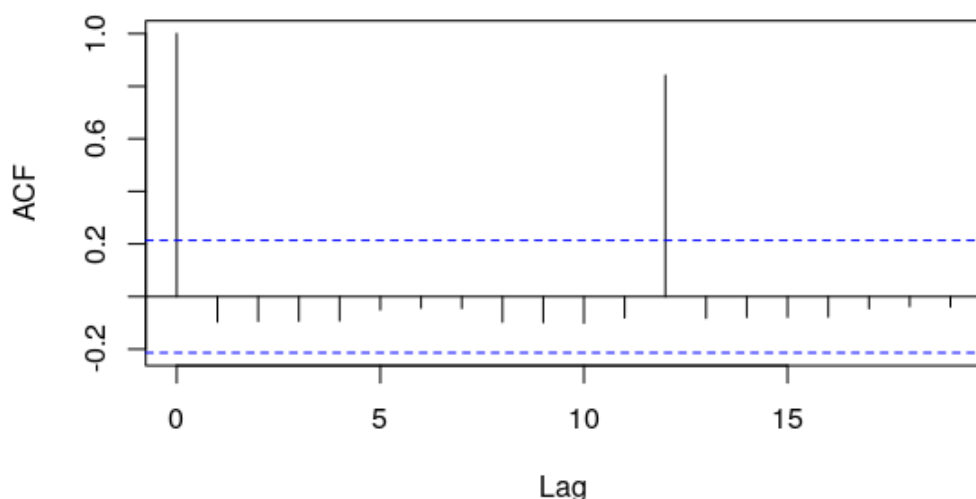
### ***Serie de tiempo Educación.***

La función de autocorrelación y autocorrelación parcial en la practica no presenta las condiciones ideales que sugieran un modelo definitivo, son la información que estás brindan se proponen modelos que después evaluamos para la selección de uno.



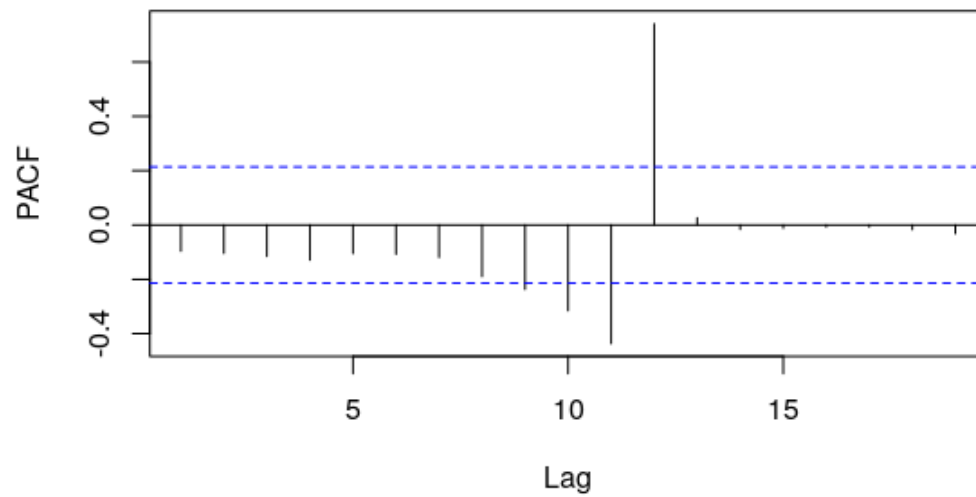
**Figura 3-38.:** Serie de tiempo índice de salud, entre el 1 de enero de 2012 y el 12 de diciembre de 2018

Los indicadores de educación generados por el DANE, reportan los establecimientos oficiales y privados de educación formal que existen en el país, identifica los factores que afectan la convivencia escolar en los estudiantes genera información para la caracterización de la población escolarizada en relación con sus actividades, relaciones familiares, comportamientos y actitudes sexuales (DANE, 2019). Estas variables tienen una incidencia sobre los costos de la educación, al iniciar el año el costo para la educación básica y media y superior tienen un crecimiento debido al ajuste que se hace a los precios, dado el crecimiento del IPC, y el calendario que está fijado anualmente. En Colombia se manejan dos calendarios para la educación básica y media, pero el impacto del precio para el calendario B no sube significativamente, igual pasa con la educación superior para el segundo semestre del año (DANE, 2019). Lo anterior se ve reflejado en la serie de tiempo de educación, ver Figura 3-38.

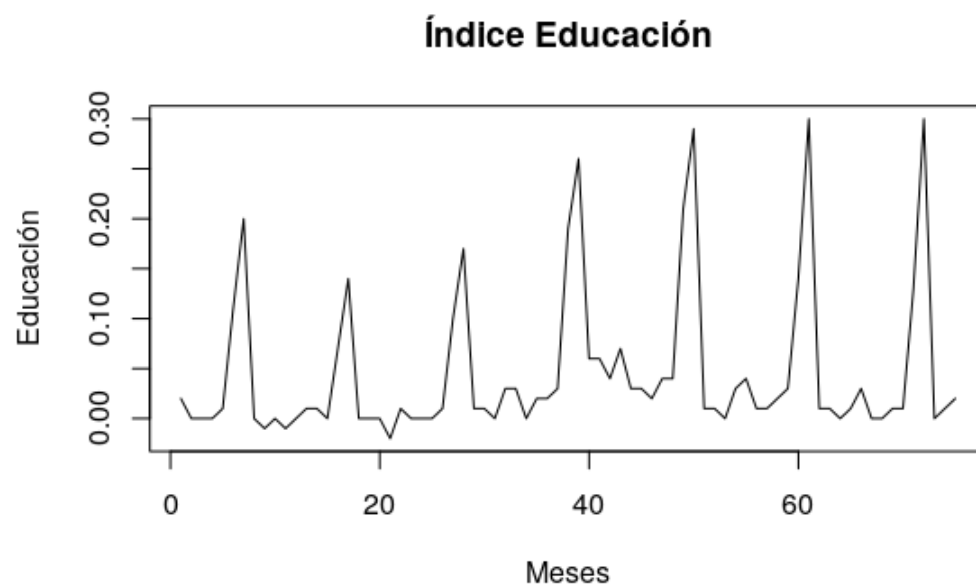


**Figura 3-39.:** ACF serie de Educación

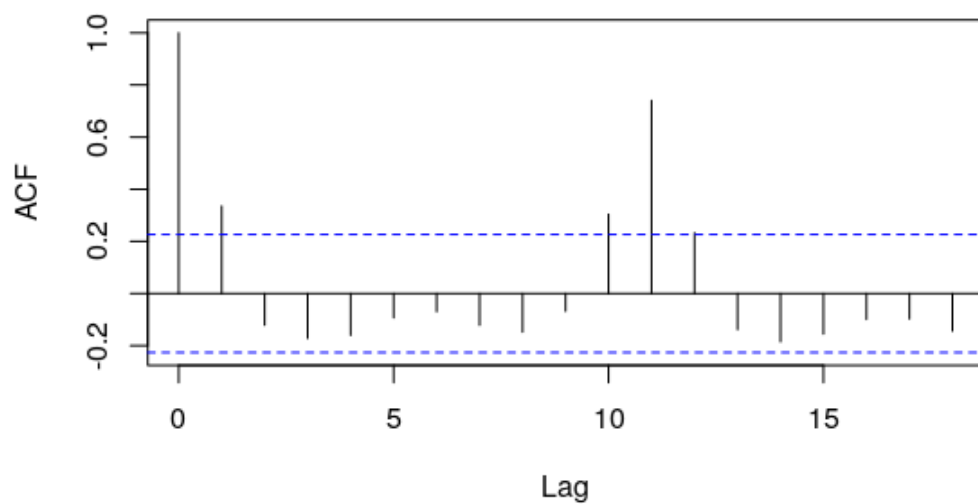
La ACF y PACF (ver figuras 3-39 y 3-40) reafirman la correlación del término autorregresivo 12, y el impacto directo de  $r_t$  sobre  $r_{t+12}$  dando como candidato a modelo un  $ARIMA(12,0,0)$ , pero esta crecimiento abrupto de los datos es causa del reajuste del salario mínimo y subida del IPC que se da a principio de año, opacando correlaciones que se pueden dar en otros ordenes, procedemos a hacer un análisis para evidenciar si dicha correlación se da.



**Figura 3-40.:** PACF serie de Educación

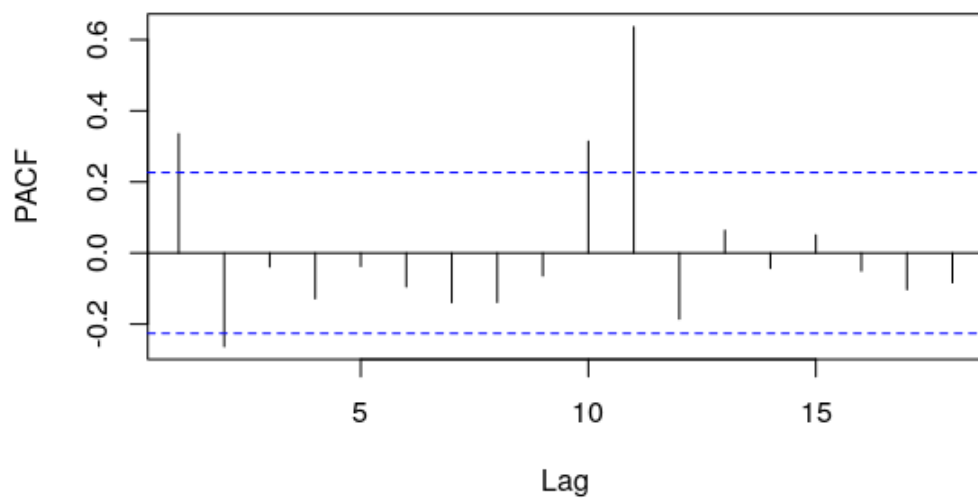


**Figura 3-41.:** Serie de tiempo de educación quitando el índice del mes de febrero.



**Figura 3-42.:** ACF serie educación quitando los datos de febrero.

Las regularidades intermedias son imperceptibles por el índice tan elevado en cada mes de febrero, pero al quitar este dato por cada año notamos que hay correlaciones,

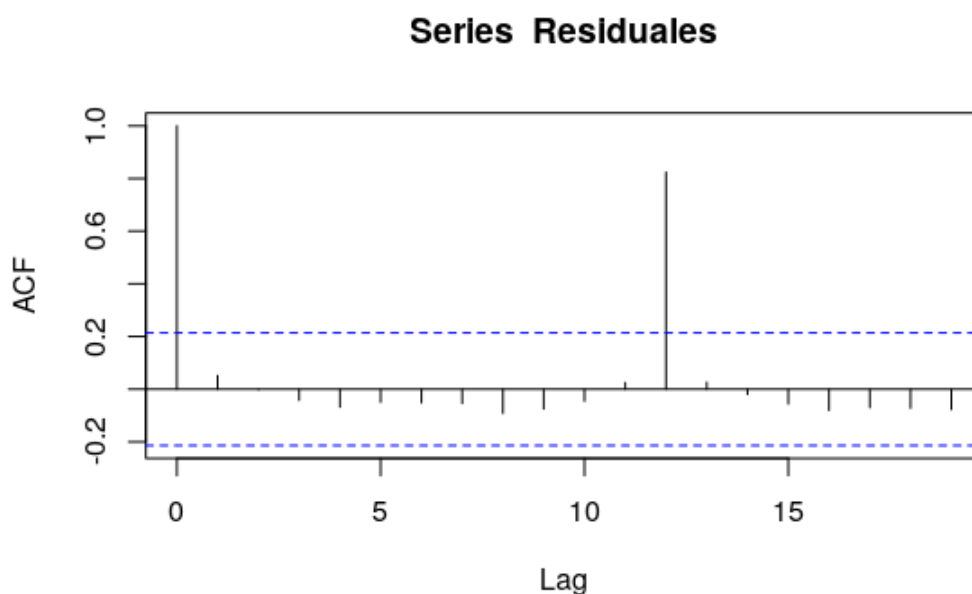


**Figura 3-43.:** PACF serie educación quitando los datos de febrero.

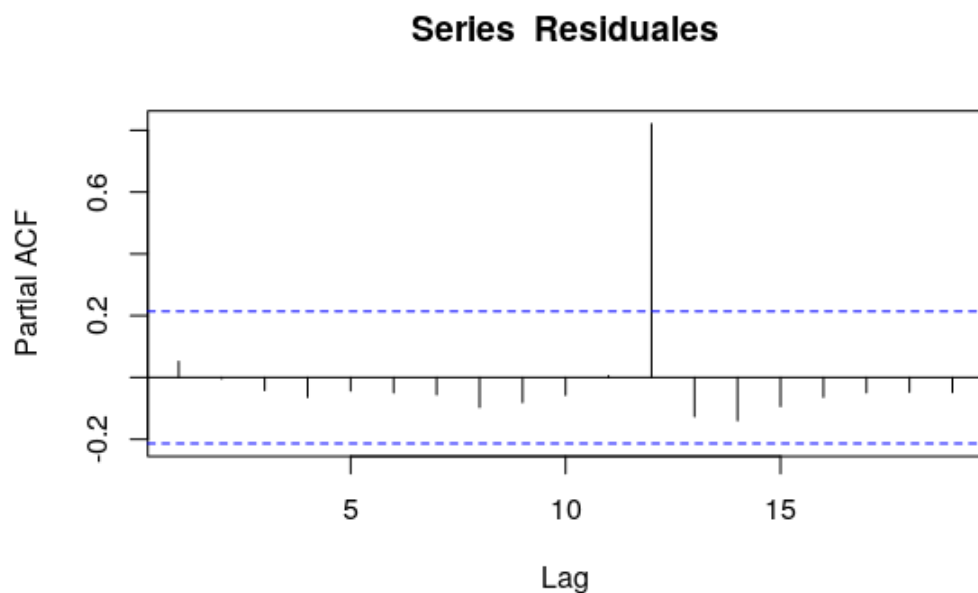
el valor del lag-1 que supera el error estándar, en el caso de que el decrecimiento en el PACF fuera decreciente con los dos lags superando el umbral del error estándar. Para el calculo del modelo se consideran todos los datos, pero teniendo en cuenta las correlaciones que se dan si se omite el dato de febrero, se consideran los modelos ARIMA(12,0,0), ARIMA(1,0,0), ARIMA(1,0,1), y ARIMA(1,0,2).

El mejor modelo considerando los criterios de determinación y de información es el modelo educación, ARIMA(1,0,1) o ARMA(1,1):  $r_t = a_t + \phi_1 r_{t-1} + \theta_1 a_{t-1}$  Los codigos en el repositorio permiten hacer la comparación de todos los modelos propuestos. El modelo explicito del último mes es:  $r_t = 0,4575 + 0,6545r_{t-1} - 0,9207a_{t-1}$ .

Pasamos a hacer el análisis de los residuales.



**Figura 3-44.:** ACF de los residuales del modelo ARIMA(1,0,1)



**Figura 3-45.:** PACF de los residuales del modelo ARIMA(1,0,1)

La ACF y PACF son un reflejo de no considerar el impacto que tiene los costes de la educación al inicio de año, aún así, el resto de correlaciones es casi cero, evidenciando que el modelo ARIMA(1,0,1) es un buen modelo.

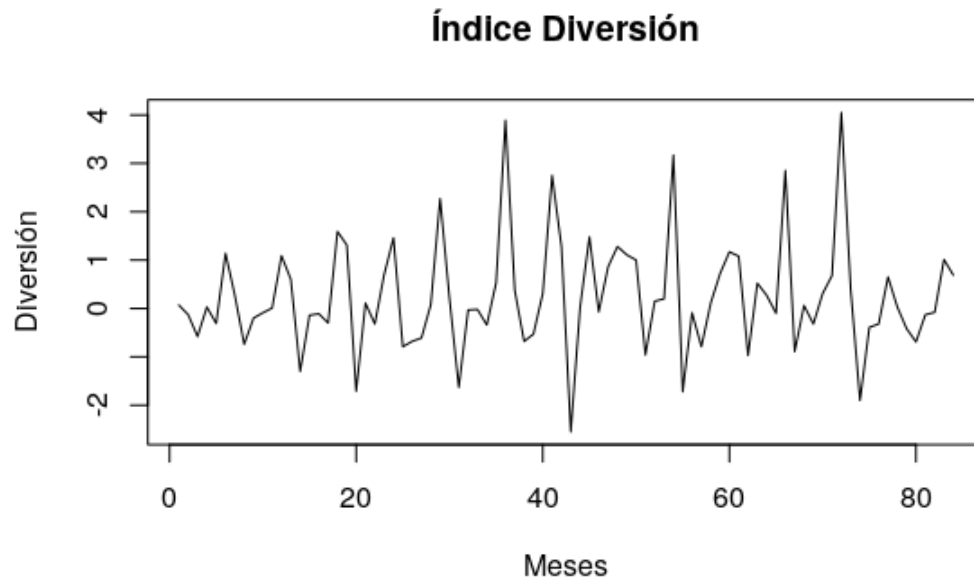
Significancia de los parámetros:

Intercepto:  $p\text{-valor}=2,2 \times 10^{-16}$

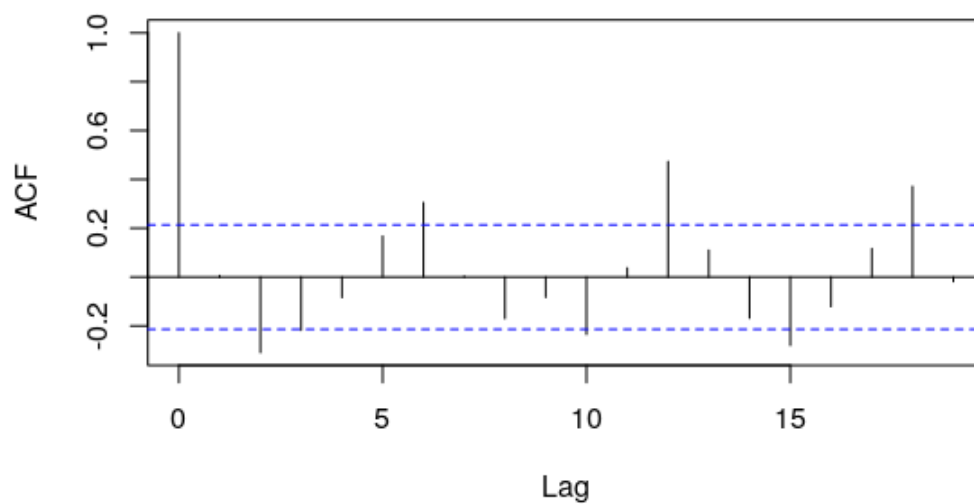
Parámetro AR1:  $p\text{-valor}=1,743 \times 10^{-6}$

Parámetro MA1:  $p\text{-valor}= 2,2 \times 10^{-16}$

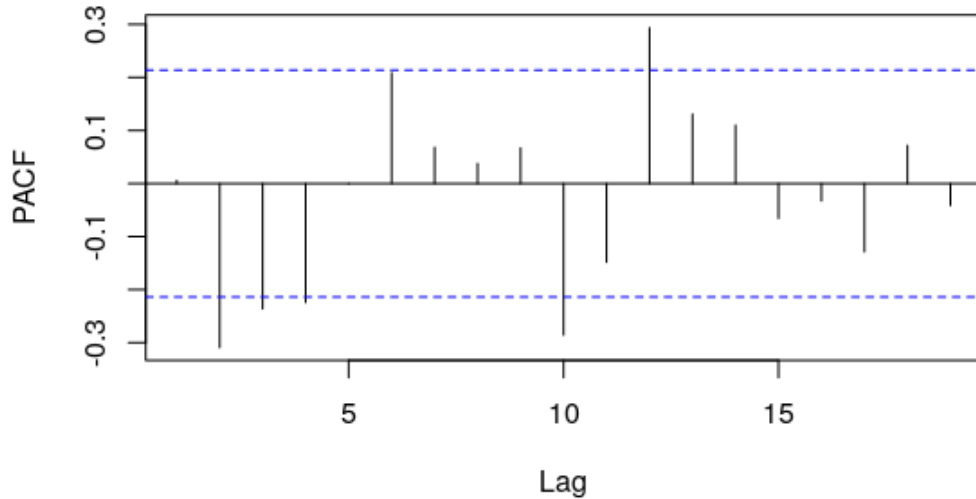
*Serie de tiempo Diversión.*



**Figura 3-46.:** Serie de tiempo índice de diversión, entre el 1 de enero de 2012 y el 12 de diciembre de 2018



**Figura 3-47.:** ACF serie Diversión.



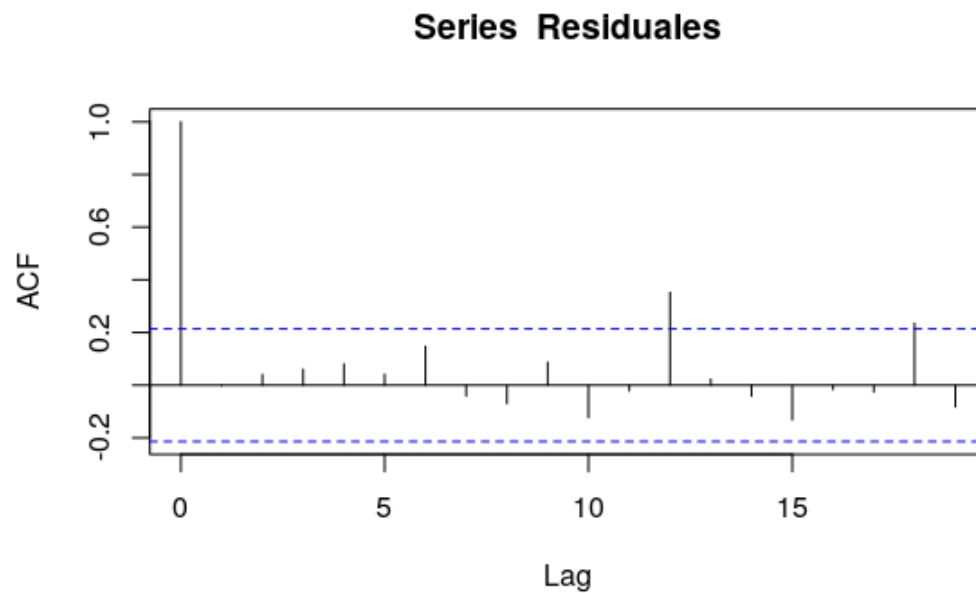
**Figura 3-48.:** PACF serie Diversión.

La serie presenta un comportamiento típico según los ACF y PACF de modelos sinusoidales que no desarrollaremos, se propone un ARIMA(3,0,0), ARIMA(4,0,0), ARIMA(6,0,0) considerando las correlaciones en los lags, se descartan modelos de orden altos AR debido al sobreajuste, se escoge el modelo en base a los criterios de determinación e indicadores dando como resultado que el mejor modelos es un ARIMA(4,0,0) o AR(4)

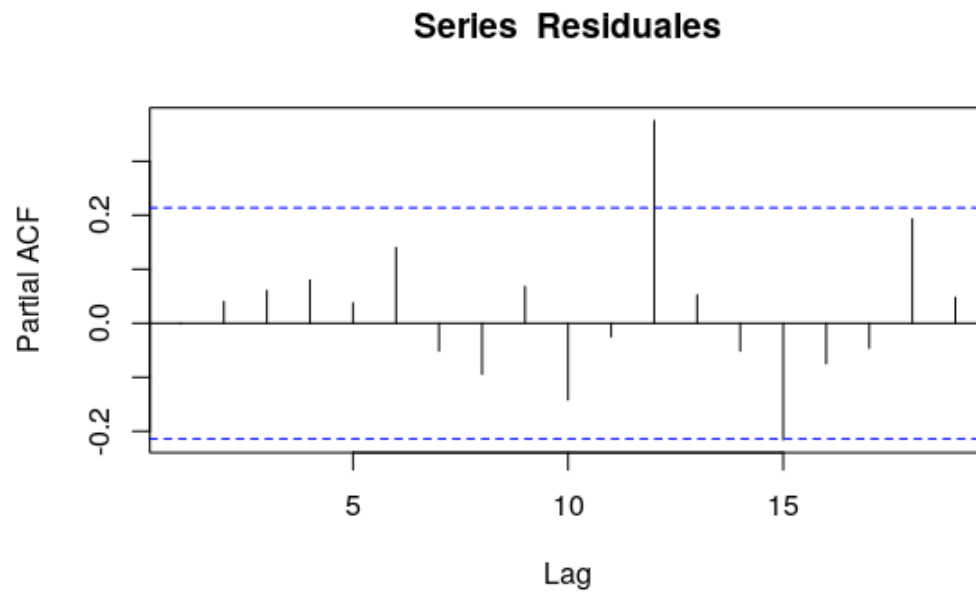
Modelo diversión, ARIMA(4,0,0) que predice el dato del mes de diciembre,

$$r_t = 0,2423 - 0,1224r_{t-1} - 0,3740r_{t-2} - 0,2506r_{t-3} - 0,2216r_{t-4}.$$

El lag-12 por encima del umbral del error estándar es un patrón que se mantiene, al igual que es notable que el resto de lags en los residuales quedan por debajo de este. Procedemos a verificar si los residuales son ruido blanco,



**Figura 3-49.:** ACF de los residuales del modelo ARIMA(4,0,0) del IPC de diversión.



**Figura 3-50.:** PACF de los residuales del modelo ARIMA(4,0,0) del IPC de diversión.

El lag-12 por encima del umbral del error estándar es un patrón que se mantiene en varias series desagregadas del IPC, al igual que es notable que el resto de lags en la ACF de los

residuales quedan por debajo del error estándar.

Significancia de los parámetros:

Intercepto:  $p\text{-valor}=2,973 \times 10^{-5}$

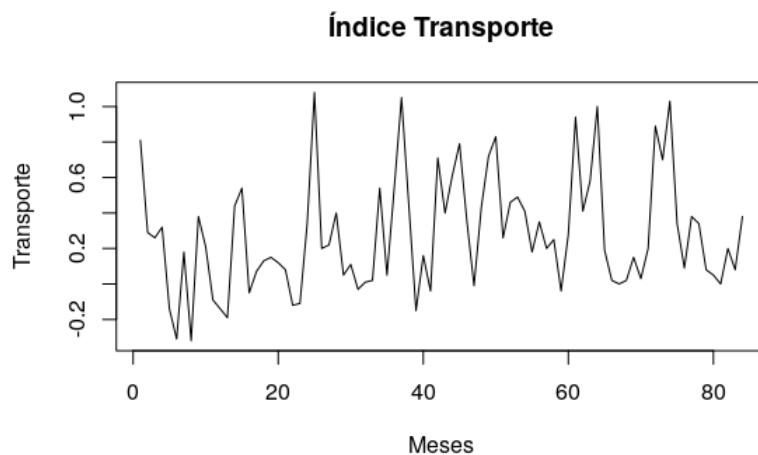
Parámetro AR1:  $p\text{-valor}=0,2485927$

Parámetro AR2:  $p\text{-valor}=0,0002777$

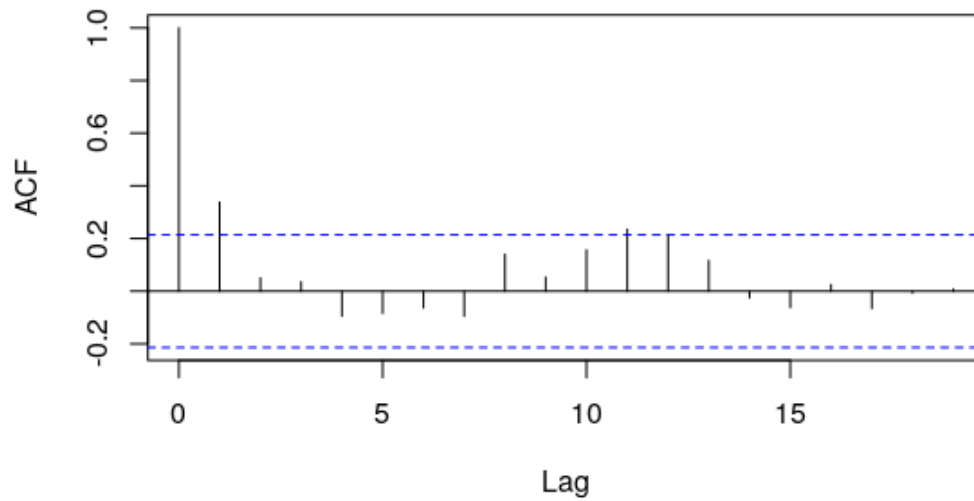
Parámetro AR3:  $p\text{-valor}=0,0142459$

Parámetro AR4:  $p\text{-valor}=0,0339424$

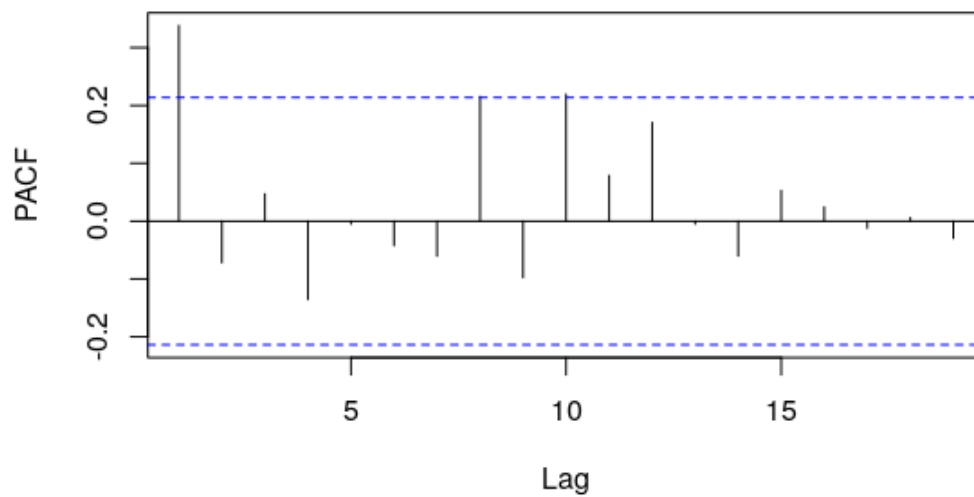
### *Serie de tiempo Transporte*



**Figura 3-51.:** Serie de tiempo índice de transporte, entre el 1 de enero de 2012 y el 12 de diciembre de 2018



**Figura 3-52.:** ACF serie transporte.



**Figura 3-53.:** PACF serie transporte.

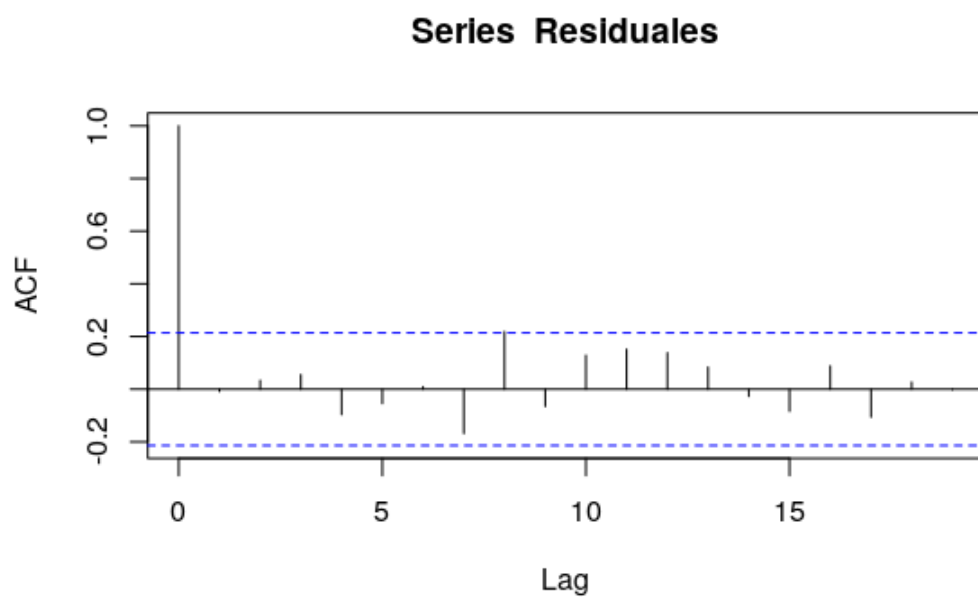
Un ACF con lag-1 significativo, pero correlación no significativa para lags más altos indica un posible modelo MA(1) o ARIMA(0,0,1). Se considera un AR(12) por el decremento del

lag-11 al lag-14 en la ACF, y el lag-11 en la PACF. Al comparar los criterios de determinación e indicadores, de forma notable tenemos que el modelo ARIMA(0,0,1) o MA(1) es mejor.

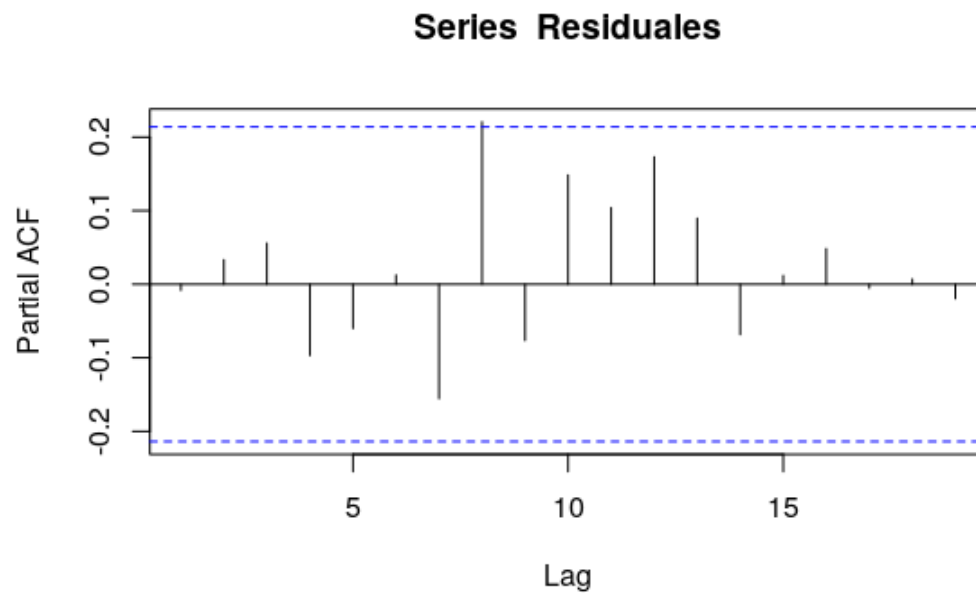
Modelo transporte ARIMA(0,0,1). Modelo que predice el dato del mes de diciembre,

$$r_t = 0,2805 + 0,3862a_{t-1}.$$

Comprobamos con los correlogramas que los residuales son ruido blanco.



**Figura 3-54.:** ACF de los residuales del modelo ARIMA(0,0,1) de transporte.



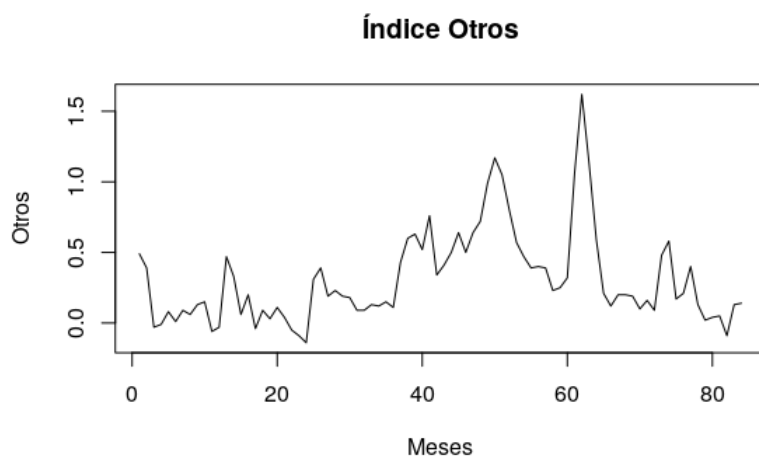
**Figura 3-55.:** PACF de los residuales del modelo ARIMA(0,0,1) de transporte.

En el ACF y la PACF de los residuales del modelo ARIMA(0,0,1) se observa el lag-6 sobre la línea del error estándar siendo el ajuste del modelo aceptable.

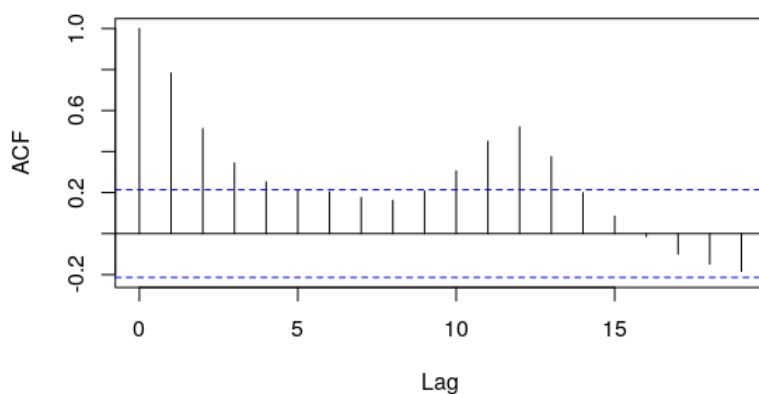
Significancia de los parámetros:

Intercepto:  $p\text{-valor}=5,302 \times 10^{-10}$

Parámetro MA1:  $p\text{-valor}=0,0002287$

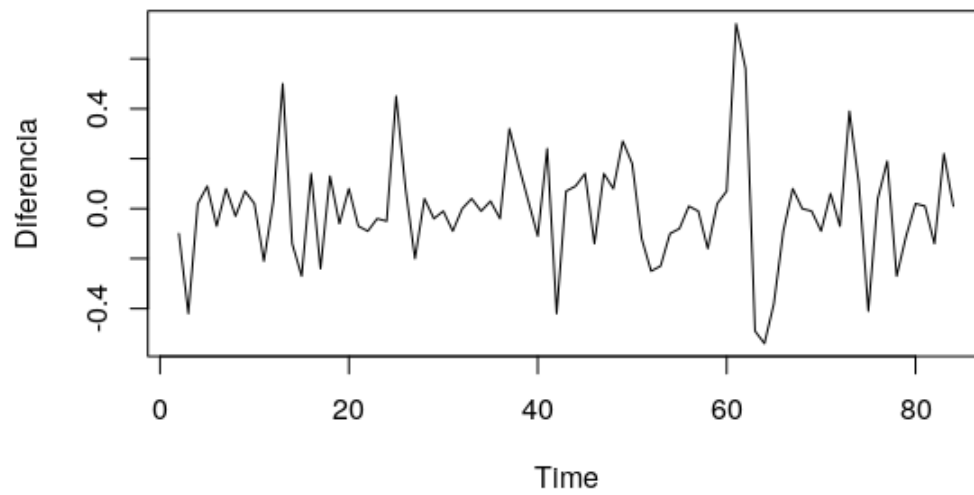
*Serie de tiempo Otros*

**Figura 3-56.:** Serie de tiempo índice otros, entre el 1 de enero de 2012 y el 12 de diciembre de 2018

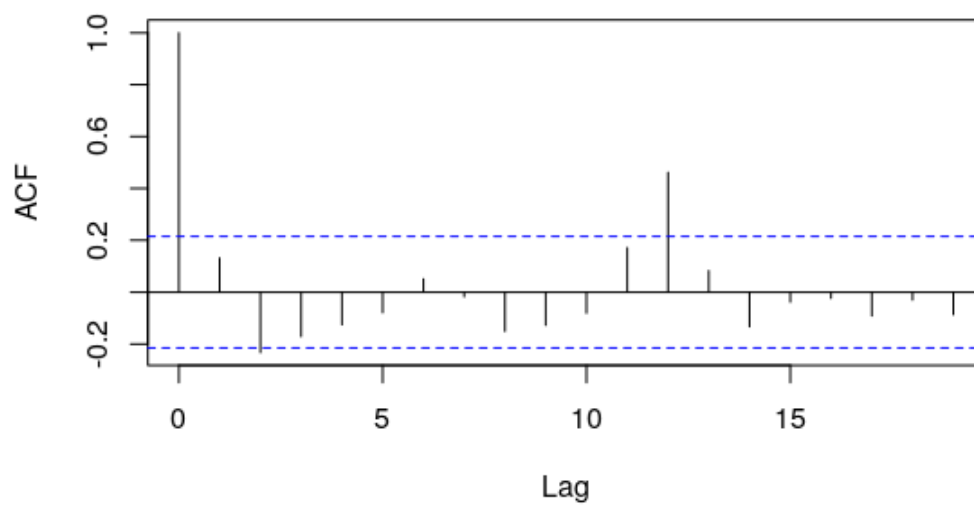


**Figura 3-57.:** ACF serie otros.

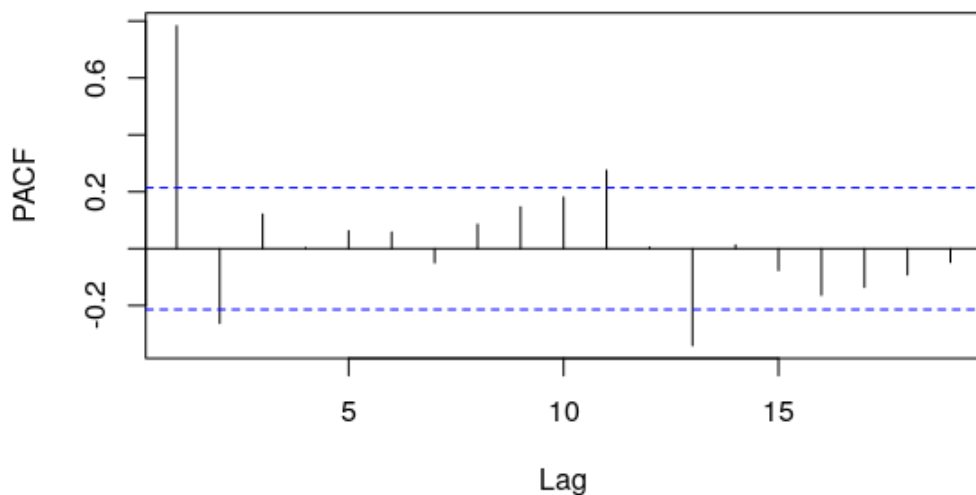
La serie de tiempo de otros presenta tendencia entre el dato 20 y el dato 60, alcanzando su pico máximo en el dato 60 que después decae, junto con un ACF que evidencia correlación en varios lags, es indicio de que hacer una diferencia puede volver estacionaria la serie.



**Figura 3-58.:** Primera diferencia de la serie otros.



**Figura 3-59.:** ACF de la diferencia de la serie otros.



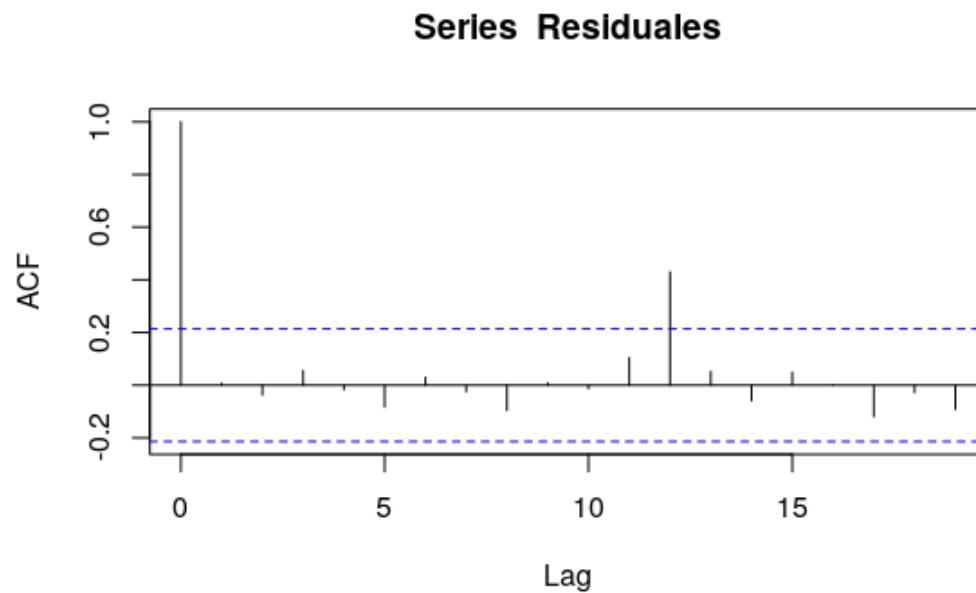
**Figura 3-60.:** PACF de diferencia de la serie otros.

Excepto por el índice entre el dato 60 y 70, la diferencia es estacionaria alrededor del valor 0. La ACF y la PACF nos permiten proponer varios modelos, el lag 1 de la PACF muestra que  $r_t$  impacta de forma directa a  $r_{t+1}$ , pero en el ACF el lag-1 no es significativo, esto quiere decir que el error tiene incidencia en un retraso, por otro lado el lag-2 de la ACF supera el error estándar, es decir, hay correlación entre  $r_t$  y  $r_{t+2}$ , siendo el mejor modelo un ARIMA(2,1,1). Se hizo una comparación con otros modelos, teniendo que los indicadores y criterios de determinación mostraban que el ARIMA(2,1,1) era en efecto el mejor modelo.

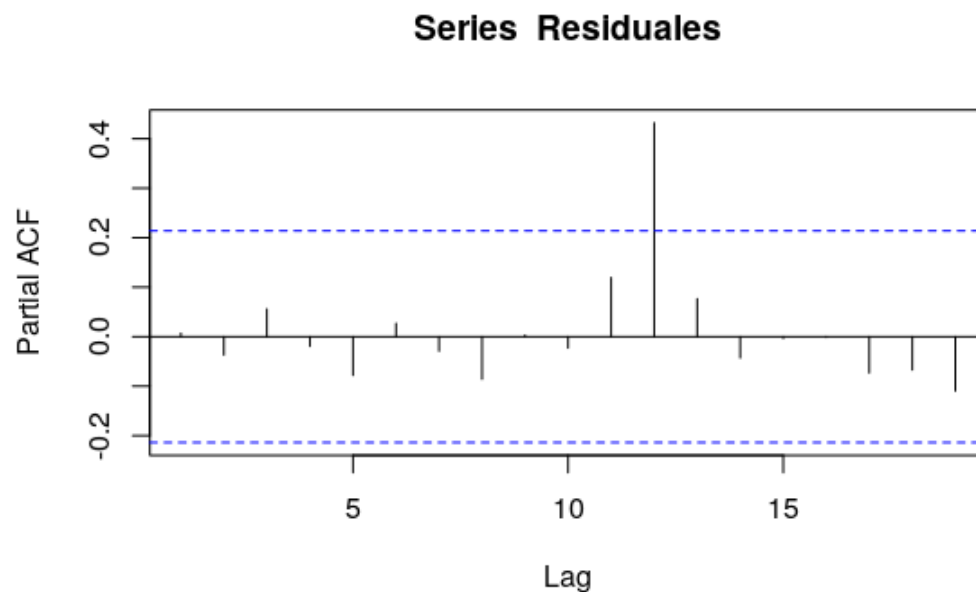
Se muestra el modelo de otros para el mes de noviembre, ARIMA(2,1,1),

$$(1 - 0,8725B + 0,3389B^2)(1 - B)r_t = (1 - 0,8425B)a_t$$

ACF y PACF de los residuales del modelo ARIMA(2,1,1).



**Figura 3-61.:** ACF de los residuales del modelo ARIMA(2,1,1) de otros.



**Figura 3-62.:** PACF de los residuales del modelo ARIMA(2,1,1) de otros.

Unicamente el lag-12 supera el umbral estándar, consideramos entonces el modelo ARIMA(2,1,1).

Significancia de los parámetros:

Parámetro AR1:  $p\text{-valor}=6,282 \times 10^{-13}$

Parámetro AR2:  $p\text{-valor}=0,001436$

Parámetro MA1:  $p\text{-valor}=2,2 \times 10^{-16}$

### 3.2. Modelo desagregado del IPC vs modelo total del IPC

Con la intención de comparar el resultado, calculamos el mejor modelo ARIMA para la serie de tiempo del índice de precios al consumidor total, para ello consideramos el AIC, el BIC, y los indicadores de determinación, en la siguiente tabla se muestran los modelos que presentan mejores indicadores.

|           | AIC   | BIC   | RMSE    | MAE     | MAPE     |
|-----------|-------|-------|---------|---------|----------|
| MA(2)     | 1.67  | 11.39 | 0.23193 | 0.17577 | 105.0582 |
| MA(1)     | 13.93 | 21.21 | 0.25302 | 0.18939 | 117.7051 |
| ARMA(2,1) | 4.09  | 16.24 | 0.2326  | 0.17522 | 104.8477 |

**Tabla 3-17.:** Comparación criterios para el modelo del IPC total.

estos son los mejores modelos considerando los criterios establecidos, como se aprecia en la tabla, el modelo MA(2) es el mejor modelo.

Modelo la serie IPC total para la predicción del dato de diciembre:

$$r_t = 0,328174 - 0,8238a_t - 0,4070a_{t-1}$$

Significancia de los parámetros:

Intercepto:  $p\text{-valor}=4,533 \times 10^{-9}$

Parámetro MA1:  $p\text{-valor}=4,073 \times 10^{-16}$

Parámetro MA2:  $P\text{-valor}=1,620 \times 10^{-5}$

Los modelos fueron establecidos considerando los datos hasta diciembre de 2018, los valores de la predicción fueron calculados de la siguiente manera: Para la predicción del mes de julio se consideran las series hasta el mes de junio, para la predicción del mes de agosto se consideran las series hasta el mes de julio, y así hasta el mes de diciembre. La razón de considerar solo este periodo para la predicción es porque los modelos en estos seis meses no

cambian.

Los pesos se distribuyen de la siguiente forma, alimentos 28.21 %, vivienda 30.1 %, vestuario 5.16 %, salud 2.43 %, educación 5.73 %, diversión 3.1 %, transporte 15.19 %, comunicaciones 3.72 %, y otros 6.35, ver DANE (2019b).

| Serie          | Modelo       |
|----------------|--------------|
| Alimentos      | AR(1)        |
| Vivienda       | ARIMA(0,0,0) |
| Vestuario      | ARIMA(0,1,3) |
| Salud          | AR(3)        |
| Educación      | ARMA(1,1)    |
| Diversión      | AR(4)        |
| Transporte     | MA(1)        |
| Comunicaciones | ARIMA(0,0,0) |
| Otros          | ARIMA(2,1,1) |

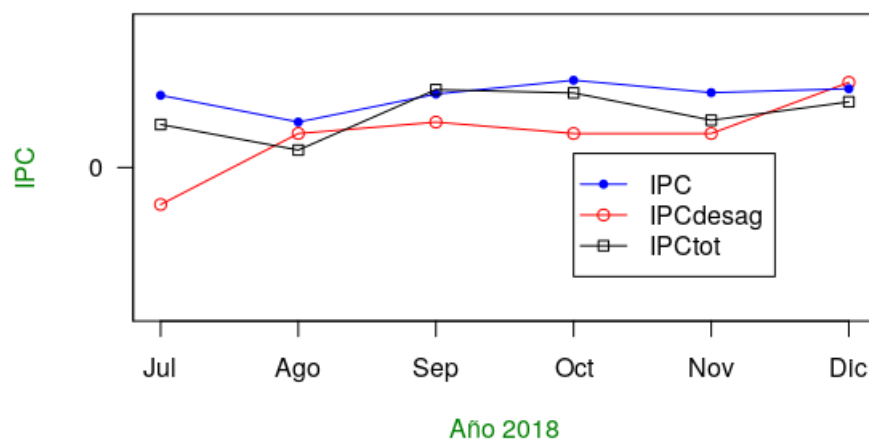
**Tabla 3-18.:** Modelos IPC desagregado.

comparamos el dato real vs la predicción del IPC desagregado y el IPC total.

|                 | Jul      | Ago      | Sep      | Oct      | Nov      | Dic      |
|-----------------|----------|----------|----------|----------|----------|----------|
| IPC desagregado | -0.13    | 0.12     | 0.16     | 0.12     | 0.12     | 0.30     |
| IPC real        | 0.25433  | 0.16042  | 0.25888  | 0.30742  | 0.26390  | 0.27769  |
| IPC total       | 0.151651 | 0.061751 | 0.274469 | 0.262759 | 0.166863 | 0.231547 |

**Tabla 3-19.:** Comparación IPC real vs predicción IPC desagregado y predicción IPC total.

No es evidente que el modelo IPC desagregado es mejor que el modelo del IPC total, en la siguiente gráfica observamos la comparación de los modelos contra el valor real.



**Figura 3-63.:** Línea azul IPC real, línea roja, predicción por los datos desagregados, línea negra, predicción por el IPC total.

Se hace un ajuste lineal para comparar los modelos estimados con el valor real del IPC total. Al comparar el modelo del IPC desagregado con el IPC total, tenemos un R-cuadrado de 0.01611 y un error estándar de estimación 0.1539. Para la comparación del modelo del IPC total con el IPC real, tenemos, un R-cuadrado de 0.1337 y un error estándar de estimación 0.1445.

Con una diferencia de los R-cuadrados de 0.11759 a favor del modelo desagregado y en los errores estándar de estimación de 0.0094 a favor del modelo del IPC total, muestra que el modelo del IPC desagregado es mejor levemente que el modelo del IPC total.

## 4. Conclusiones

### 4.1. Conclusiones

Este trabajo permite evidenciar que para el caso de la inflación mensual en Colombia, y el período de estudio seleccionado, emplear un método considerando los datos desagregados para pronosticar la inflación del siguiente mes tiene un mejor comportamiento que emplear un modelo ARIMA del IPC total. En especial encontramos que partiendo de los índices de precios desagregados al grupo de bienes y servicios a nivel nacional se puede construir un mejor pronóstico para la inflación mensual en lugar de emplear un modelo univariado para el IPC nacional total.

De esta manera, nuestros resultados muestran que, para pronosticar la inflación en el corto plazo, es útil emplear información desagregada que incorpora de forma diferente las innovaciones (modelos MA) y diferencias (modelos ARIMA). Estas innovaciones y diferencias pueden ser enmascarados en el comportamiento agregado. Así un método como el empleado puede incorporar mejor toda la información disponible.

El peso del índice de alimentos dentro del grupo de bienes y servicios del 28 %, es considerable. Estudiar la relación directa con modelos VAR entre la serie de alimentos y el IPC total puede que mejore el modelo, esto debido a que es el índice que más varía dentro del grupo.

Si no consideramos el valor del lag-12 de la ACF y PACF de los residuales, junto con la significancia de los parámetros escogidos, los modelos cumplirían las necesidades teóricas para establecer que son buenos modelos. El efecto de lags de corto alcance es de mayor relevancia a la hora de predecir valores en los meses de julio a diciembre, razón por la cual optamos por modelos de ordenes bajos, partir de la no normalidad de los datos permitirá capturar la volatilidad presente que se evidencia en estos lags y establecer correlaciones de impacto, consideraciones que el modelo ARIMA no alcanza a capturar. Por ello para trabajos futuros considerar modelos como los ARCH o GARCH bajo el supuesto de la no normalidad podrían dar como resultado la no correlación de los residuales. El primer modelo considera la información de las perturbaciones para modelar la volatilidad de la serie mientras que el segundo lo hace de forma generalizada al incluir la volatilidad pasada como variable explicativa (Tsay, 2010. pp. 115-145).

# Bibliografía

- [1] DANE. (2006). Manual de Índices de Precios al Consumidor Teoría y Practica. Recuperado de: [https://www.dane.gov.co/files/investigaciones/fichas/cpi\\_sp.pdf](https://www.dane.gov.co/files/investigaciones/fichas/cpi_sp.pdf)
- [2] González, M.P. (2009). Análisis de series temporales: Modelos ARIMA. Editorial UPV/EHU. España. 19-38.
- [3] Box, G., Jenkins G., & Reinsel, G. (1994). Time Series Analysis Forecasting and Control. 4 Ed. JHON WILEY & SON, INC. United States. 55-57.
- [4] Dickey, D., & Fuller, W. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. Journal of the American Statistical Association 74, 427-431.
- [5] Guerrero, V.M. (2003). Análisis Estadístico de Series de Tiempo Económicas. International Thomson Editores. Ciudad de México.
- [6] Salazar, N., Steiner, R., Becerra, A., Ramírez, J. (2013). Los efectos del precio del suelo sobre el precio de la vivienda para Colombia. Ensayos Sobre Política Económica 31, 18-66.
- [7] DANE. (2019). Boletín Técnico Índice de Precios de la Vivienda Nueva (IPVN) I Trimestre 2019.
- [8] Tsay, R. (2010). Analysis of Financial Time Series. 3 Ed. JOHN WILEY & SONS, INC. United States.
- [9] Paul, H., & Brani, V. (2007). Nonparametric Statistics with Applications to Science and Engineering. JOHN WILEY & SONS, INC. United States. 90-91.
- [10] Willmoot, C., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assesing average model performance. CLIMATE RESEARCH 30, 79-82.
- [11] Jalil, M., & Misas, M. (2006). Evaluación de pronósticos del tipo de cambio utilizando redes neuronales y funciones de pérdida asimétricas. Technical report 376. Sugerencia de estudios económicos-Banco Central de Colombia.

- [12] Banco de la Republica de Colombia. (2008). Índice representativo del mercado de deuda pública interna: IDXTES (488). Recuperado de: <http://repositorio.banrep.gov.co/handle/20.500.12134/5505>.
- [13] Santana, J. C. (2006). Predicción de series temporales con redes neuronales: una aplicación a la inflación colombiana. *Revista Colombiana de Estadística*, 29, 77-92.
- [14] Gallego, J. P. (2010). Aplicación de la teoría de caos para el análisis y pronóstico de series de tiempo financieras en Colombia. Medellín: Universidad Nacional de Colombia. Tesis de Maestría.
- [15] Espinal, T., Giraldo, P., Peña, N., & Lotero, L. (2016). Análisis, modelizado y pronóstico de la serie de índices de precios al consumidor en Medellín 2010-2015. Encuentro de Investigación Formativa-Memorias VII versión. Universidad Pontificia Bolivariana. Colombia.
- [16] Huwiler, M., & Kaufmann, D. (2013). Combining disaggregate forecasts for inflation: The SNB's ARIMA model. *Swiss National Bank Economic Studies*.
- [17] Zhang, G. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* 50, 159-175.
- [18] Bierens, H. (1996). *Topics in Advanced Econometrics: Estimation, Testing, and Specification of Cross-Section and Time Series Models*. Cambridge University Press. United States. 136-138
- [19] Botchkarev, A. (2019). A NEW TYPOLOGY DESIGN OF PERFORMANCE METRICS TO MEASURE ERRORS IN MACHINE LEARNING REGRESSION ALGORITHMS. *Interdisciplinary Journal of Information, Knowledge, and Management* 14, 45-76.
- [20] DANE. (2019b). Estadísticas: Índice de precios al consumidor (Base 2018). Recuperado de <https://www.dane.gov.co/index.php/estadisticas-por-tema/precios-y-costos/indice-de-precios-al-consumidor-ipc/ipc-historicobase-2008>
- [21] Silviu, G., & Abe, S. (1985). The Principle of Maximum Entropy. *THE MATHEMATICAL INTELLIGENCER* 7, 42-48.
- [22] Pedrosa, I., Juarros-Basterretxea, J., Robles-Fernández, A., Basteiro, J., & García Cueto, E. (2015). Pruebas de bondad de ajuste en distribuciones simétricas, ¿qué estadístico utilizar? *Universitas Psychologica*, 14(1), 245-254.

## **A. Anexo: Códigos utilizados para cada modelo.**

### **A.0.1. Repositorio GitHub**

En el repositorio se encuentran los códigos usados para cada modelo.

<https://github.com/Harvyncho/ModelosArimaIPC.git>