CONTENIDO

1. Introducción

Dentro del gran abasto de técnicas analíticas que comprende la minería de datos, existe una gama de modelos que permiten realizar pronósticos de series temporales. En general la elección del método que se use para realizar el pronóstico de las series será determinada por las características generales de las series que buscamos pronosticar, pero sin olvidar nunca que siempre será complicado obtener un resultado exacto. El objetivo, al final de cuentas, es encontrar el mejor modelo que permita realizar un resultado lo más cercano posible a la realidad.

2. Características generales de una serie temporal

Una **serie de tiempo** como tal puede ser descrita como una colección de datos reunidos sobre la misma variable a lo largo de un período de tiempo establecido. Estas deberán de ser bajo las mismas condiciones durante el período y con intervalo de la misma medida. Esto significa que la información deberá de recopilarse en intervalos regulares, es decir:

- Diarios.
- Semanales.
- Mensuales.
- Trimestral, etc..

La notación que se suele ocupar para definir una serie de tiempo es la siguiente:

$$Y_t = f(Y_{t-1}, Y_{t-2}, \dots, Y_{t-p})$$

Donde:

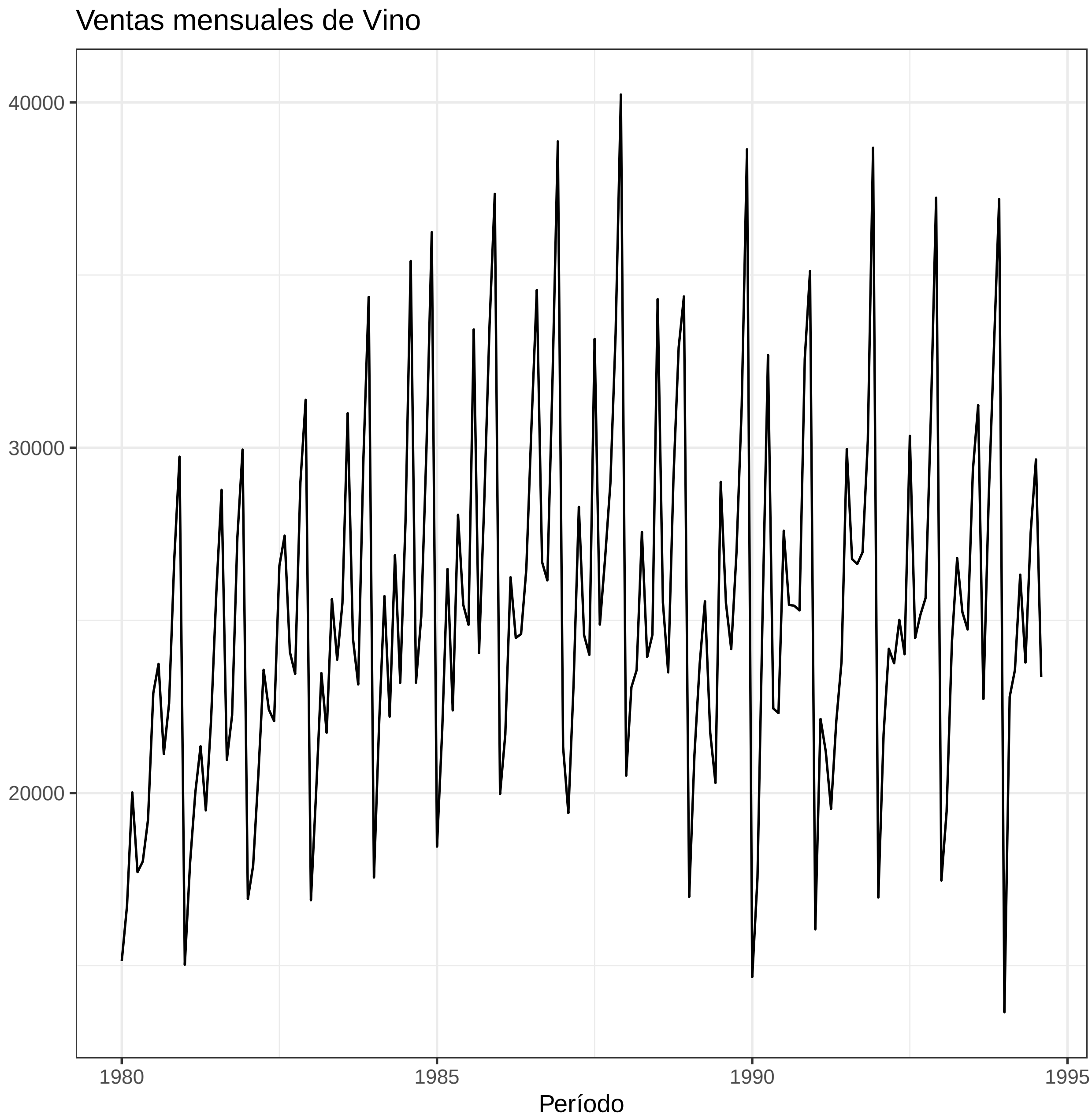
- Y_t es el valor de la serie temporal en el período t.
- Y_{t-1} es la variable en el período $t-1$, es decir, con un rezago.
- Y_{t-p} es el valor de la variable en el período $t-p$, es decir, con p rezagos.

Lo cual en su conjunto se refiere a que la variable Y_t esta en función de si misma, pero en períodos rezagados. La forma de poder visualizar una serie de tiempo por excelencia es utilizando gráficas históricas.

```
library(forecast)

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

library(ggplot2)
ventas <- forecast::wineind
autoplot(ventas)+labs(title = "Ventas mensuales de Vino", x="Período", y="")+theme_bw()
```



La función *autoplot()* de ggplot2 pide como requisito que para que pueda realizar la gráfica de tiempo de la serie esta deberá de estar guardada y cargada como una serie de tiempo.

```
class(ventas)

## [1] "ts"
```

Con la siguiente gráfica podemos observar con mayor detalle la definición de una serie tempora; un conjunto de observaciones de una variable (Ventas mensuales de Vino) que se mueven a través de un determinado período de tiempo (Con periodicidad mensual, de 1980 hasta 1995).

2.1. Componentes generales de una serie temporal.

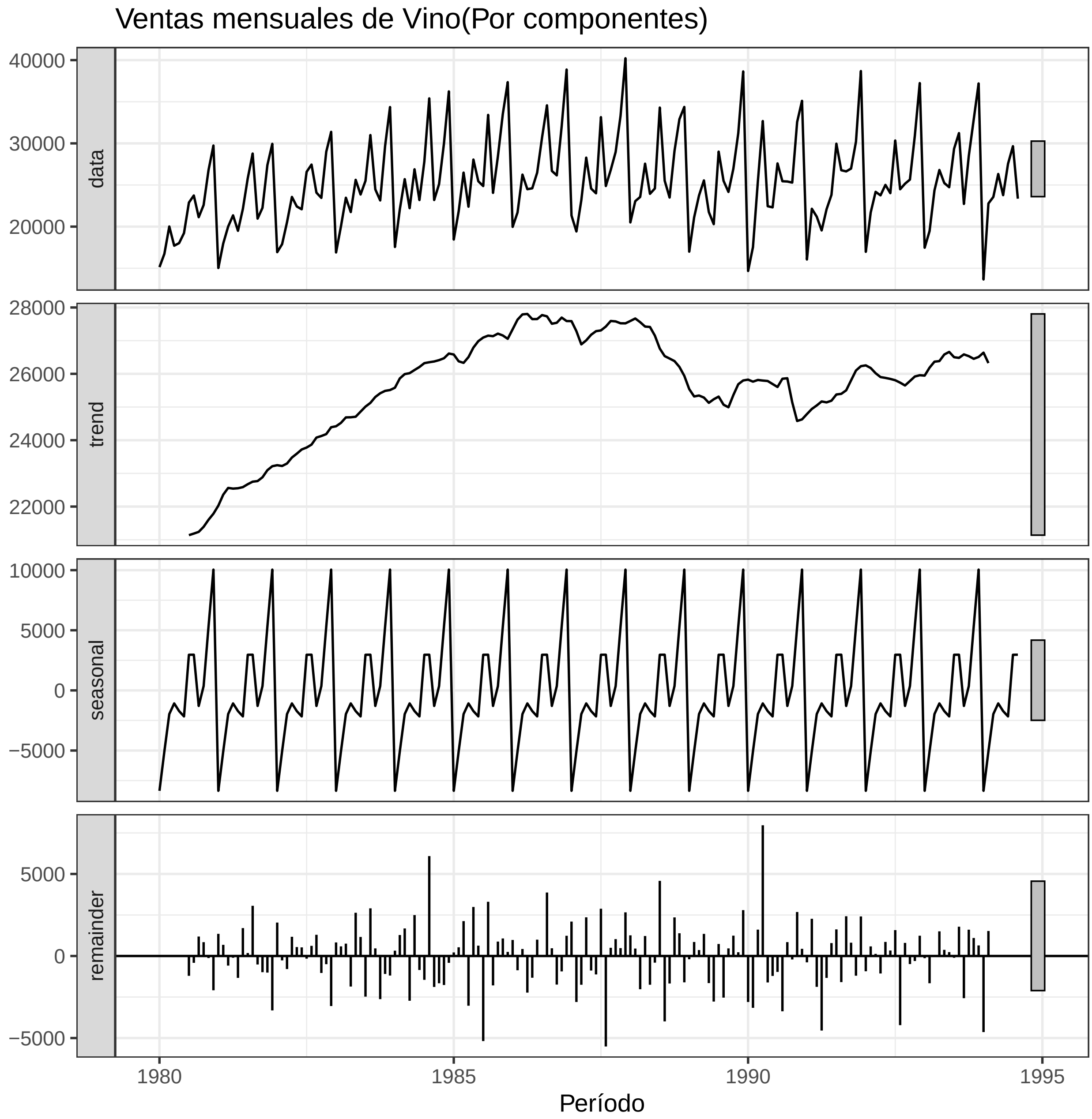
Una práctica común en el manejo de pronosticos es suponer que los datos de una serie de tiempo están integrados por componentes individuales que pueden ser ”eliminados” o ”calculados” para facilitar el análisis de las series. Estos componentes normalmente son:

- Tendencia.
- Estacionalidad.

- Ciclo (en algunos casos).
- Componentes Aleatorio

Para visualizar de mejor manera las diferencias entre los componentes que integran una serie de tiempo, realizaremos el siguiente ejemplo:

```
Descomposicion <- decompose(ventas)
autoplot(Descomposicion)+labs(title = "Ventas mensuales de Vino(Por componentes)", x="Período",
```



El proceso determinista para descomponer la serie con la función *descompose()* se basa principalmente en algunos métodos elementales:

- **Tendencia**: La tendencia se calcula como una media o promedio móvil.
- **Componente estacional**: El componente estacional se consigue promediando los valores de cada unidad de tiempo para todos los períodos (Como por ejemplo, calculando el promedio de los mese de abril si la serie es de periodicidad mensual) y finalmente centrando los datos resultantes.
- **Residuos**: Los residuos o componentes aleatorios se obtienen:

- Si tiene relación aditiva:

$$Ruido\ Blanco = data - tendencia - efecto\ estacional$$

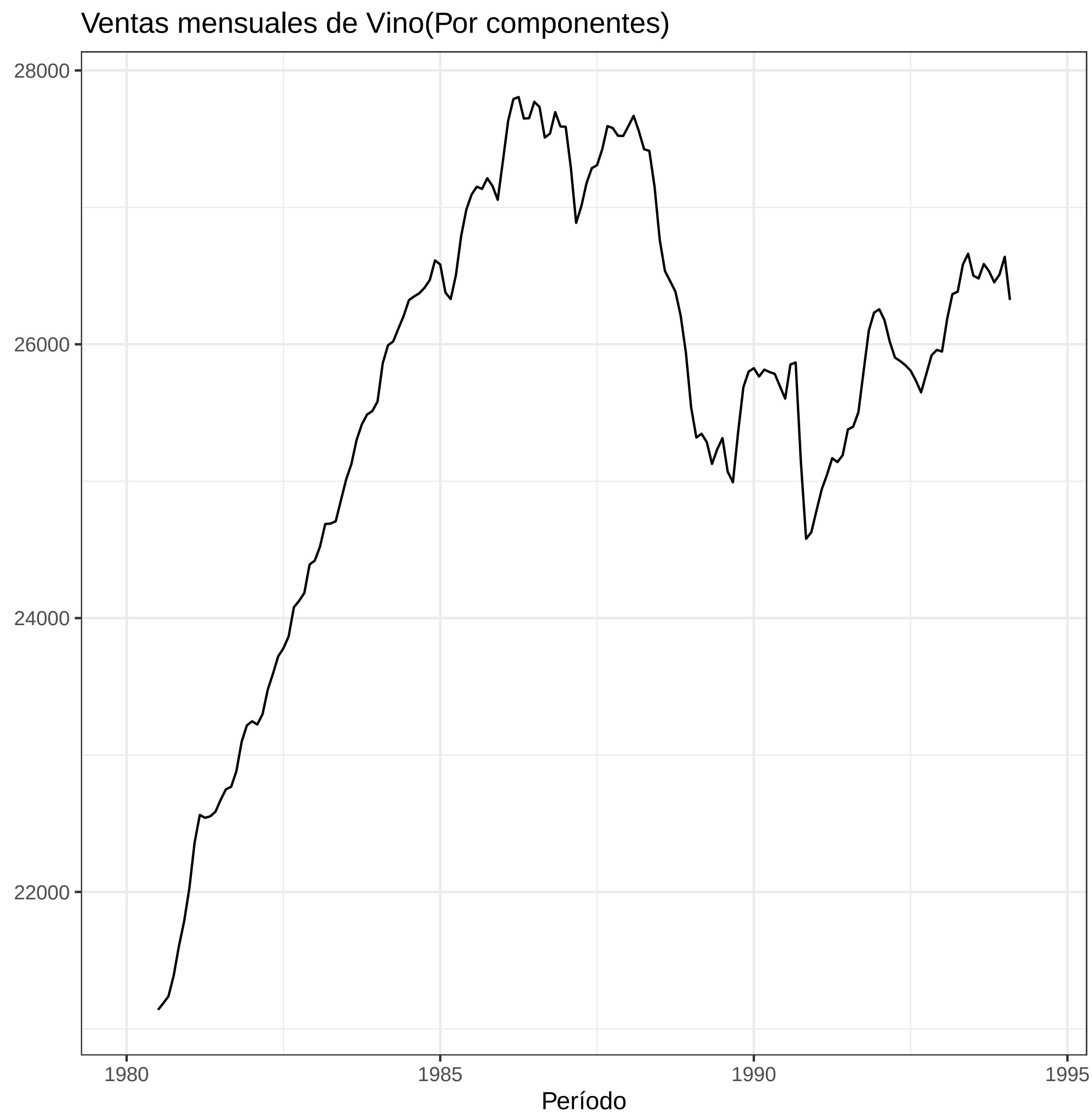
- Si tiene relación multiplicativa:

$$Ruido\ Blanco = \frac{data}{tendencia \cdot efecto\ estacional}$$

2.1.1. Tendencia

Podemos comprender a la **tendencia** de una serie de tiempo como el comportamiento a largo plazo de una serie.
Visualizando únicamente la tendencia, tendremos lo siguiente:

```
autoplot(Descomposicion$trend )+labs(title = "Ventas mensuales de Vino(Por componentes)", x
```



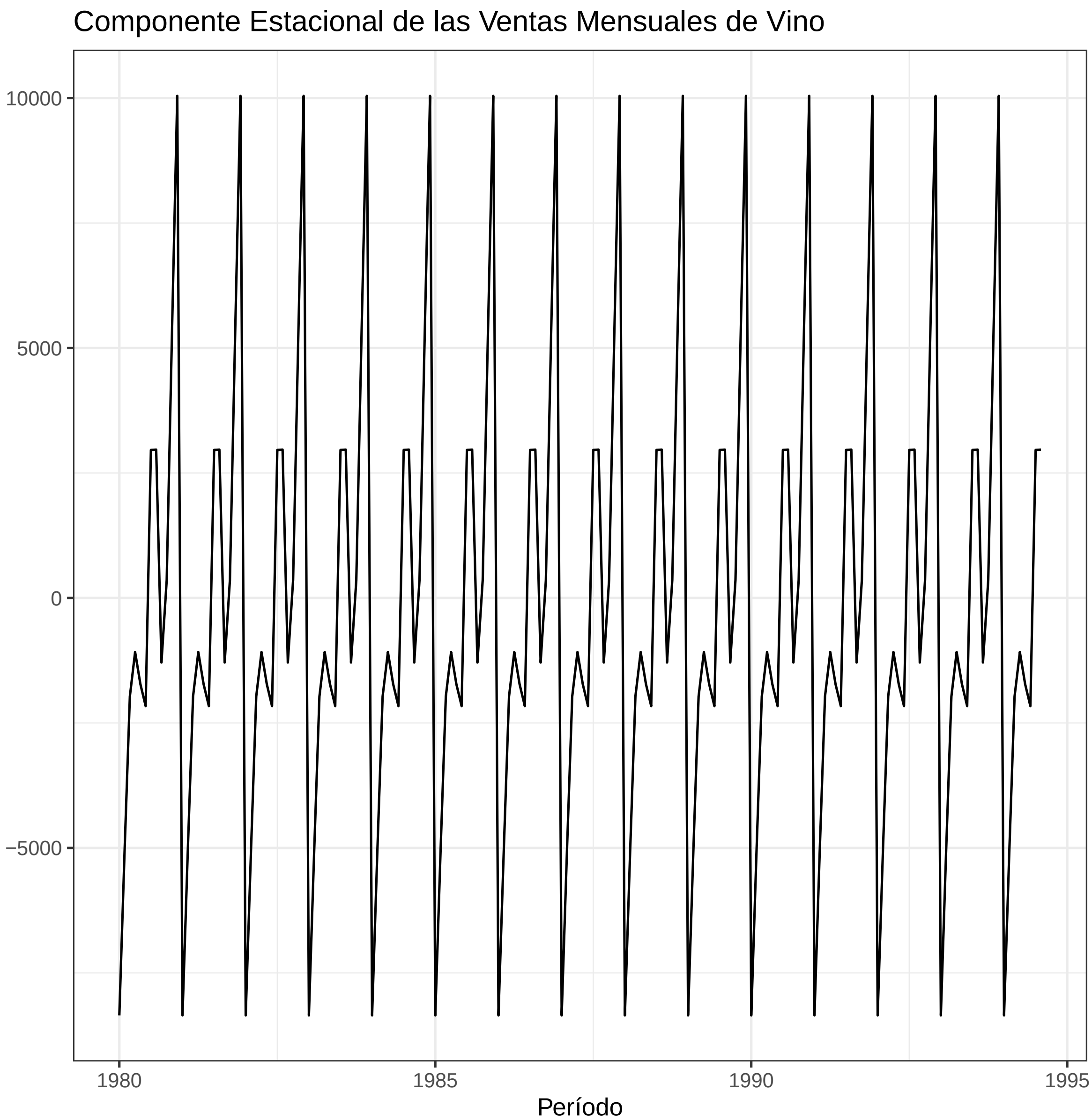
Podemos observar que la tendencia tiene un efecto fuerte sobre la serie de ventas, y que a pesar de tener un comportamiento sostenido en la gran parte del período, esta tambipen tuvo un importante cambio alrededor de 1986 a 1991, dode prácticamente paso de tener un comportamiente creciente a uno decreciente.

Es importante mencionar que el utilizar un **Método NO lineal** (como por ejemplo, utilizando *promedios móviles* como en el ejercicio) para estimar el comportamiento de tendencia por lo regular tiene mejores resultados que estimar la tendencia de **Manera Lineal** (como por ejemplo, utilizando una *regresión lineal*), ya que permite captar los cambios estructurales importantes de la tendencia.

2.1.2. Componente estacional

El **componente estacional** de una serie de tiempo representa, en su esencia más básica, la variabilidad presente en los datos de la serie ocasionada por influencia de las estaciones o períodos de tiempo en específico. Esta variación relacionada a los movimientos de la serie ocurre cada determinado período de tiempo; por ejemplo, año tras año en los mismos meses con prácticamente la misma intensidad. Gráficando el componente estacional de nuestra base de datos, tenemos lo siguiente:

```
autoplot(Descomposicion$seasonal )+labs(title = "Componente Estacional de las Ventas Mensua
```



Podemos confirmar que, en efecto, las variaciones estacionales de la serie son bastantes intensas con la particularidad de que se repiten exactamente en el mismo período de tiempo.

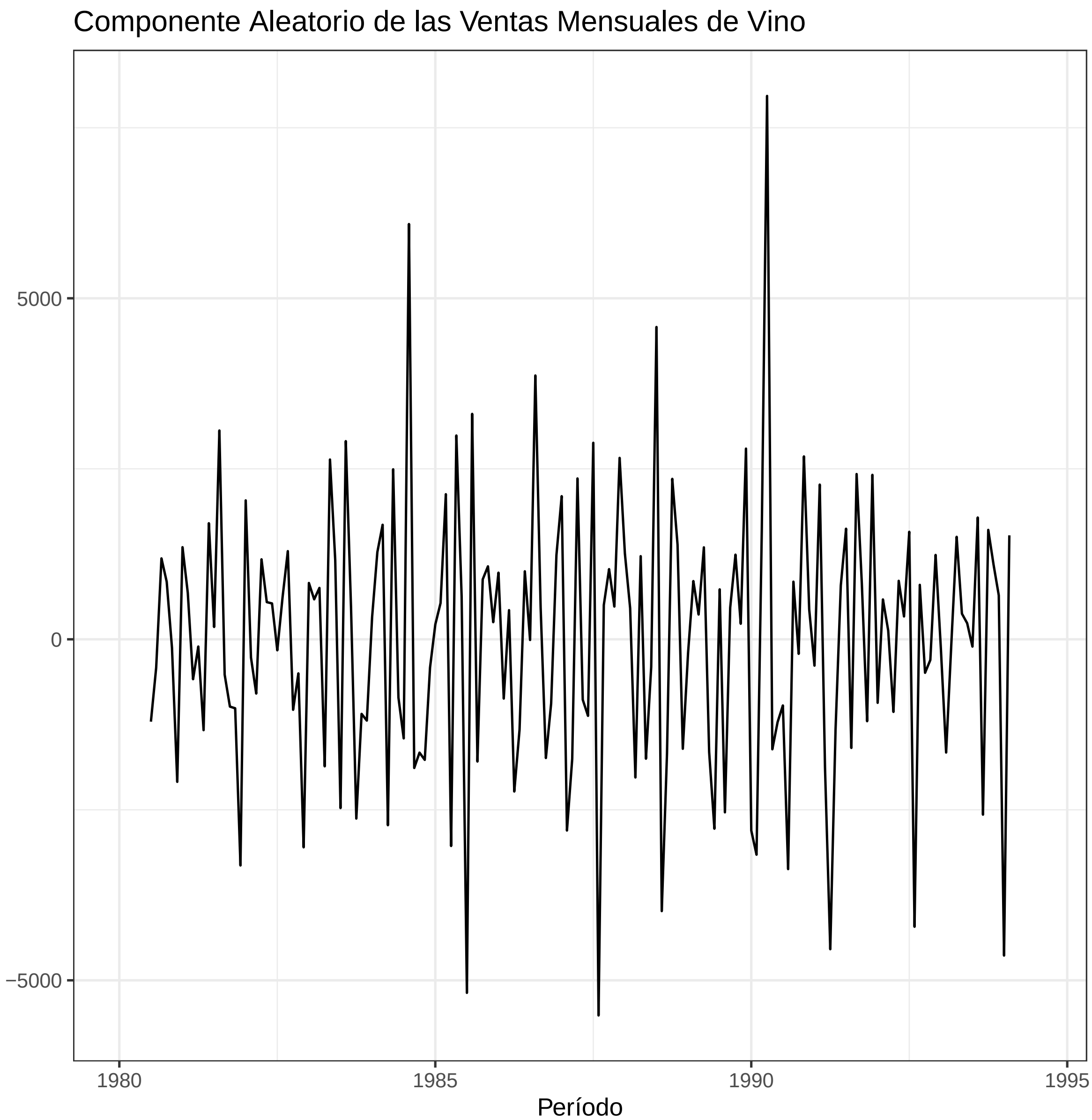
2.1.3. Componente Aleatorio o irregular

Este es causado principalmente por factores a corto plazo los cuales son a su vez imprevisibles y no recurrentes. El cálculo de este componente ayuda a explicar la variabilidad aleatoria de la serie, ya que no se puede esperar predecir su impacto sobre la serie de tiempo.
En general existen dos tipos de variaciones irregulares en las series de tiempo:

1. Aquellos movimientos provocados por acontecimientos especiales, como lo son las elecciones, las inundaciones o fenómenos naturales importantes, huelgas, o cualquier evento que afectan a la serie.
2. Variaciones aleatorias, las cuales presentan razones que no se pueden señalar, pero que sin problemas logran estabilizarse a través del período.

Este componente nos ayuda a garantizar que la serie esta "lo más limpia" posible, por lo que cualquier otra variación o comportamiento solo responde a sucesos completamente aleatorios y no a un factor o variable especifica. Realizamos la gráfica del componente aleatorio de la siguiente manera:

```
autoplot(Descomposicion$random )+labs(title = "Componente Aleatorio de las Ventas Mensuales
```



Gracias a la gráfica podemos observar algunas características interesantes de los residuos:

- Presenta importantes caídas y subidas.
- Siempre varia entre valores positivos y negativos.
- Se mantiene constante a partir de una media igual a 0.
- Sus variaciones se mantienen dentro de los intervalos de valores de -5000 a 5000 con algunas excepciones.

Todas estas características nos ayudaran a especificar de mejor manera si este componente es ”realmente aleatorio” o presenta algún comportamiento que no hemos modelado, ¿ Cómo?, comparándolo con otra serie que llamaremos ”Ruido Blanco’.

3. Metodología Box-Jenkins para la estimación de pronósticos.

3.1. Ruido Blanco

El análisis de los residuos de una serie nos permiten contrastar si un modelo para pronosticar es útil o no para predecir el futuro de la serie.

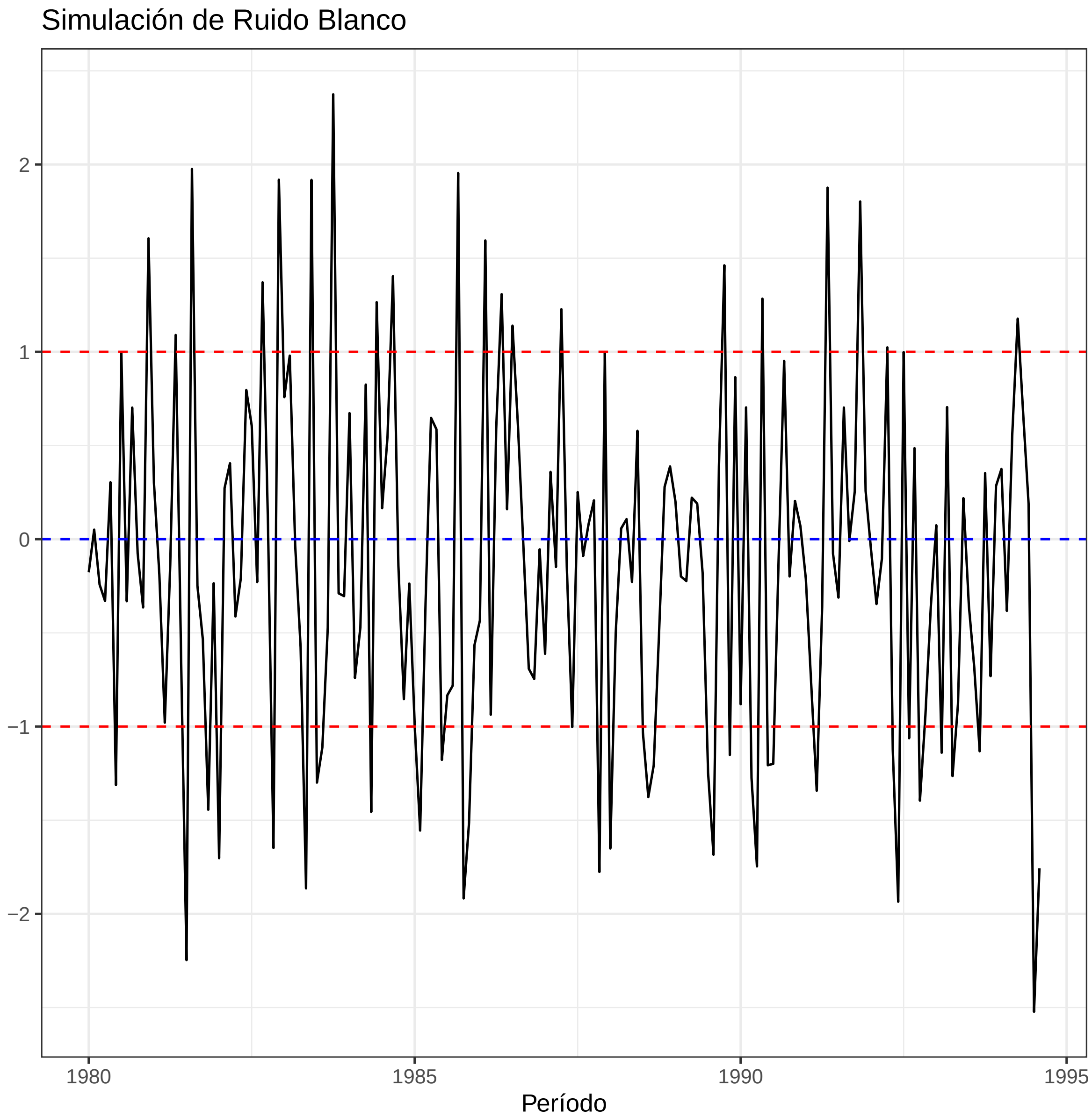
$$Residuos\ de\ una\ serie = Valor\ real\ de\ los\ datos - Predicción\ de\ los\ datos$$

En notación matemática tendremos que se cumplirá lo siguiente:

$$\begin{gathered} E[e_t] = E[e_{t+h}] = \mu_t = 0 \\ Var[e_t] = Var[e_{t+h}] = \sigma^2 \\ Cov(e_t, e_{t+h}) = Cov[e_{t+j}, e_{t+j+h}] = 0 \end{gathered}$$

En el siguiente ejemplo. buscaremos realizar la simulación de un Ruido Blanco de tipo Gaussiano (sigue una distribución normal) utilizando la funcion *rnomr()*.

```
library(tseries)
Ruido <- rnorm(176, mean = 0, sd =1)
autoplot(ts(Ruido, start=c(1980,1), frequency = 12))+
geom_hline(yintercept = 0,lty =2, color = "blue")+
geom_hline(yintercept = 1,lty =2, color = "red")+
geom_hline(yintercept = -1,lty =2, color = "red")+
labs(title = "Simulación de Ruido Blanco", x = "Período", y ="" )+theme_bw()
```



Observemos que la simulación de Ruido Blanco que creamos presenta, tanto una media que se mantiene a un nivel de 0, y una desviaci+on estándar igual a 1, la cual no varía ya que se mantiene en los rangos de -1 a 1, característcas qque definen a un Ruido Blanco Estricto.
Las lineas rojas en la gráfica son las bandas de desviación estándar que simulamos, de tal manera que la simulación no supera en ningún punto ese intervalo, por lo que es evidencia suficiente para contrastar que la serie es estacionaria en Varianza, ya que presenta una desviación estándar constante igual a 1;y que además es estacionaria en media, ya que la linea azul, la cual es la media de la serie simulada, es igual a 0, confirmando el supuesto de que la serie presenta una media constante que es igual a 0

3.2. Análisis con Correlogramas

El análisis del correlograma es una opción que nos ayudará bastante a analizar la interdependencia de los valores de observación. En cuestiones de la metodología Box-Jekins que estamos analizando, el análisis con correlogramas nos ayudará a encontrar el tipo de proceso estocástico que mejor represente a una serie temporal, de tal forma que nos permitirán especificar el tipo de modelo a utilizar sobre la serie propuesta.
La idea con la que debemos conservar al manejar correlogtamas es que sirven para describir la presencia o ausencia de correlación en los datos de las series temporales, indicando si las observaciones pasadas influyen en las actuales. A grandes rasgos, se puede decir que la autocorrelación hace referencia cuando los valores que toman una variable en el tiempo son independientes entre

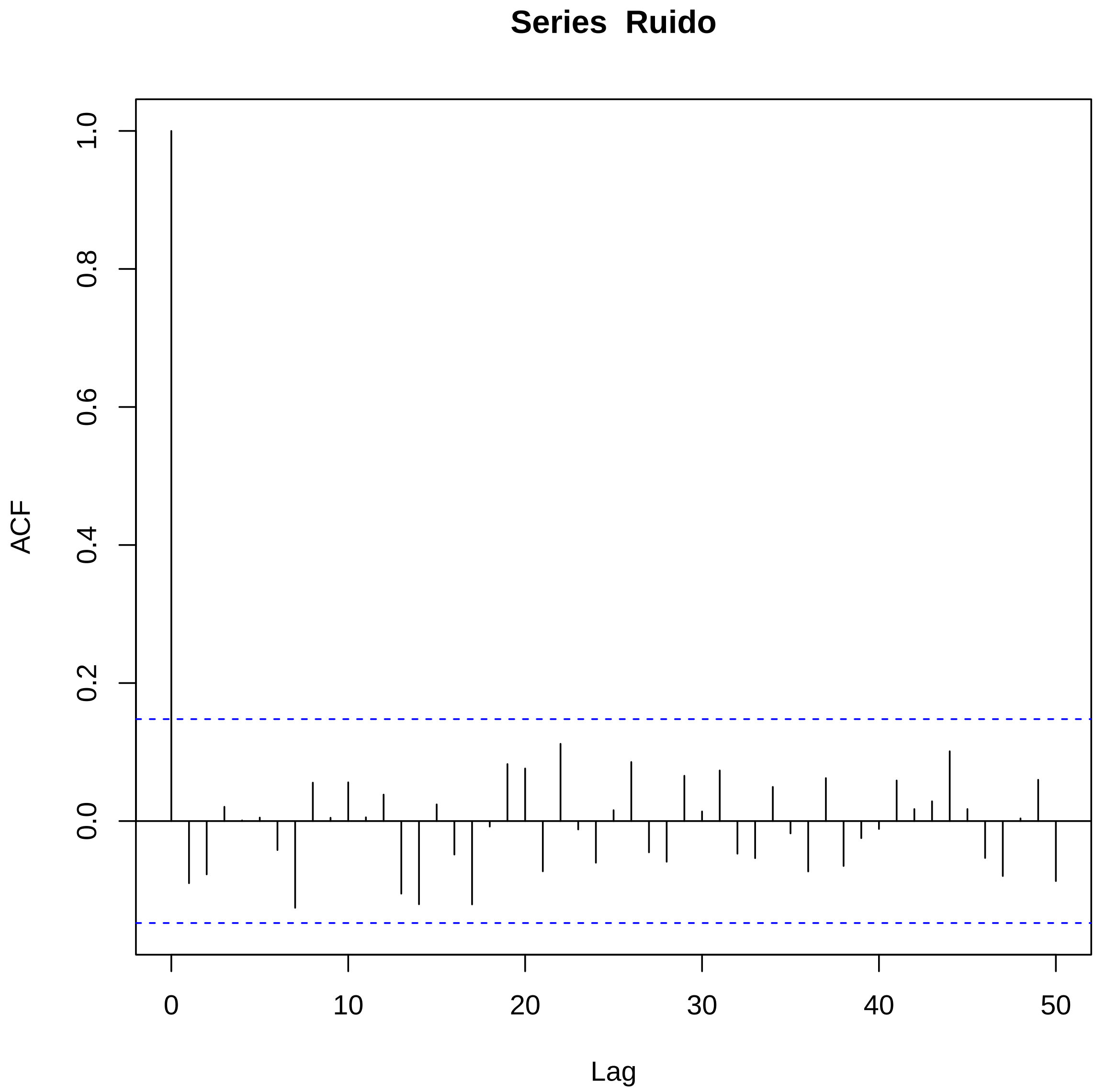
sí en diferentes períodos, sino que un valor determinado en el futuro depende de los valores anteriores de la misma variable. Las dos funciones de **auto-correlación** que podemos manejar en los pronósticos de Seires de Tiempo son:

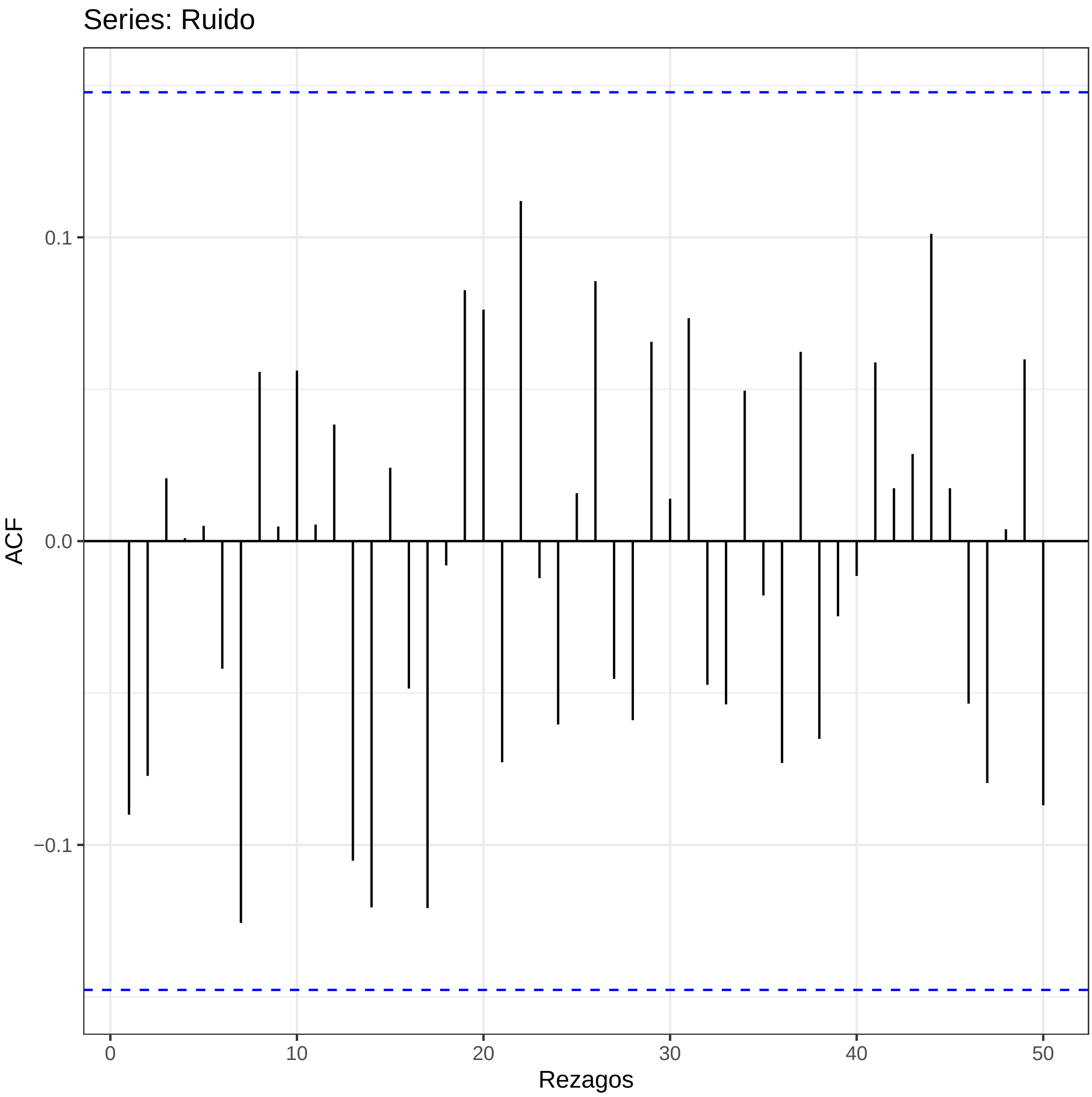
- **La función de auto-correlación simple:** Mide la correlación entre dos variables separadas por k períodos, o, en otras palabras, el grado de asociación lineal que existe entre dos variables del mismo proceso aleatorio.
- **La función de auto-correlación parcial:** Mide la correlación entre dos variables separadas por k períodos cuando no se considera la presencia de la dependencia creada por los retardos intermedios que hay entre ambas. En otras palabras, esta encuentra la auto-correlación que existe entre dos variables separadas por k períodos descontando los posibles efectos debidos a variables intermedias.

3.2.1. Auto-Corelación Simple

Para realizar el análisis de correlogramas de una manera rápida y sencilla podemos utilizar la función *acf()*. Esta función te permite seleccionr el número de rezagos que quieres analizar utilizando la opción *lag*. Para visualizar el correlograma utilizaremos, de igual manera, la función *autoplot()*. Para realizar el correlación sobre la simulacion de Ruido Blanco, utilizamos el siguiente código:

```
autoplot(acf(Ruido, lag.max =50))+
labs(titiltle = "Correlograma de la simulación de Ruido Blanco", x="Rezagos")+theme_bw()
```


Series Ruido
ACF
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
Lag

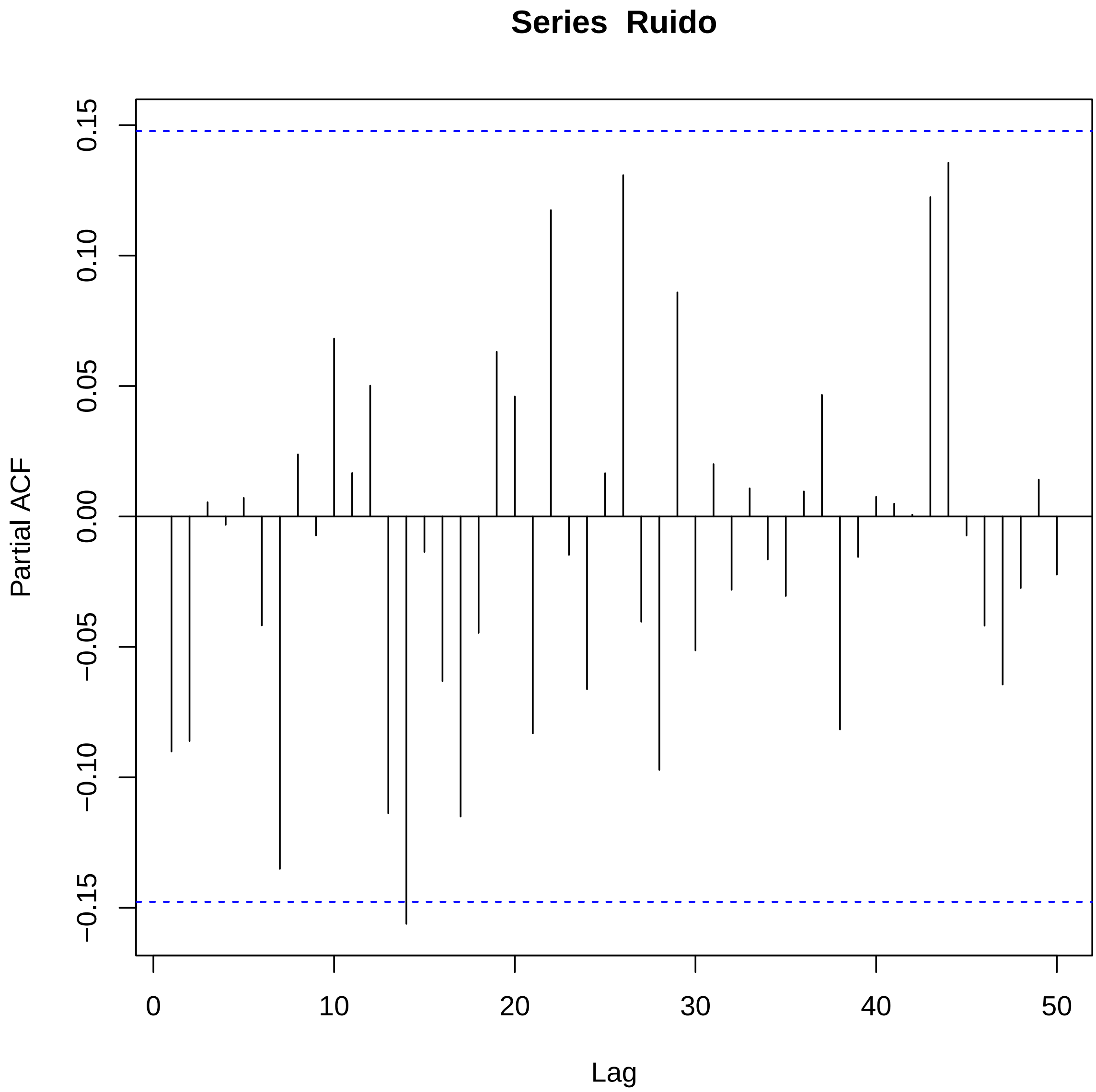


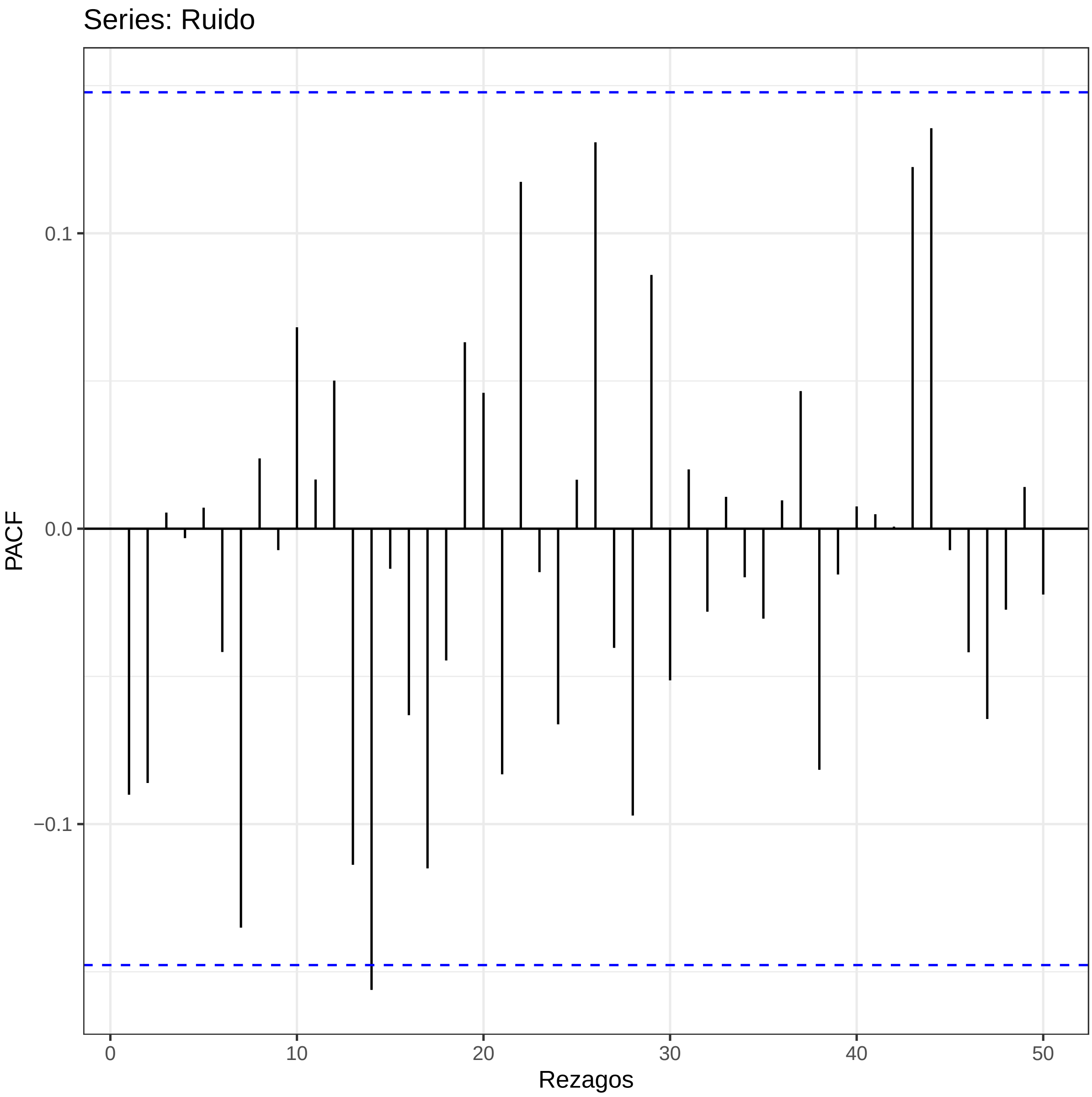
Podemos observar que al utilizar el cálculo del correlograma simple en el Ruido Blanco, prácticamente ninguno de los rezagos de la serie es mayor al intervalo de confianza que se creó, de tal manera que la serie no presenta problemas de auto-correlación entre las distintas observaciones de la serie.

3.2.2. Auto-Corelación Parcial

Para proceder a realizar el correlograma parcial de la serie podemos ocupar la función *pacf()*, ocupando el siguiente código:

```
autoplot(pacf(Ruido, lag=50))+
labs(titile = "Correlograma Parcial de la simulación de Ruido Blanco", x="Rezagos")+theme_
```


Series Ruido
Partial ACF
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
0
10
20
30
40
50
Lag



Al igual que el correlograma normal, el correlograma parcial del Ruido Blanco muestra que no existe una importante correlación parcial de ninguno de los rezagos de la serie. Estos resultado son congruentes con el Ruido Blanco que simulamos, ya que las observaciones pasadas de la serie no afectan a las observaciones furutas, de tal forma que no existe correlación serial entre los rezagos de la variable.

3.3. ARIMA

Los modelos autorregresivos integrados de promedio móvil (ARIMA) son una clase de modelos lineales que son capaces de representar tanto series de tiempo *estacionarias* como *NO estacionarias.* Los modelos ARIMA no implican variables independientes en su construcción. En vez de elllos, utilizan la información de la serie misma para generar pronósticos. Por ejemplo, un modelo ARIMA para ventas mensuales proyectaría el patrón de ventas históricas para generar un pronóstico para las ventas del siguiente mes. Los modelos ARIMA dependen mucho de los patrones de autocorrelación en los datos.

La *metodología Box-Jenkins* se refiere a un conjunto de procedimientos para identificar, ajustar y verificar modelos ARIMA con los datos de la serie de tiempo. Los pronósticos se derivan directamente de la forma de un modelo ajustado.

Al seleccionar un modelo, recuerde que las autocorrelaciones calculadas de los datos no serán exactamente iguales a las autocorrelaciones teóricas asociadas con un modelo ARIMA. Las autocorrelaciones calculadas de los datos están sujetas a la variación de la muestra.

El nombre genérico ARIMA de estos modelos hace referencia a sus tres componentes principales:

1. Auto-regresivo (AR).
2. Integrado (I)
3. Medias Móviles (MA)

La metodología Box-Jenkins se basa en el proceso de modelizar una serie emporal que consiste en derivar un modelo ARIMA que se ajuste al conjunto de datos dado, para lo cual, se requiere realizar un proceso de análisis de las características esenciales de las series de tiempo como la tendencia, estacionalidad, estaciones cíclicas, funciones de auto-correlación y el análisis de los residuos. Un modelo ARIMA convencional, el cual surge de incorporar dos procesos estocásticos AR(p) y MA(q).
El método ARIMA (auto-regresivo de medias móviles integrado) es una clase de modelo que explican una serie de tiempo dada basado en su pasado (tanto en valores como en errores cometidos), de modo que su ecuación puede ser usada para predecir el futuro.
Estos modelos se caracterizan con tres parámetros: p, d y q donde:

- p es el orden de la parte auto-regresiva (AR(p)).
- d es el número de diferenciaciones requeridas para tener una serie estacionaria (I).
- q es el orden de la parte de medias móviles (MA(q)).

Observación: No funciona bien si los datos NO son estacionarios.

3.3.1. Modelos Autorregresivos AR(p)

Los modelos auto-regresivos predicen el valor actual de nuestra serie temporal en función de los valores pasados.
Un proceso AR de orden p se formula de la siguiente manera:

$$Y_t = \Phi_0 + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \cdots + \Phi_p Y_{t-p} + \varepsilon_t$$

Donde:

- Y_t es la variable de interés.
- $\Phi_0, \Phi_1, \ldots, \Phi_p$ son los parámetros para estimar el modelo de orden p , es decir, son coeficientes que serán estimados
- Y_{t-p} es la variable rezagada en p períodos de tiempo.
- ε_t es el término de error en el tiempo t, el cual presenta los efectos de variables no explicadas por el modelo; los supuestos acerca del término de error son los mismos que los del modelo de regresión estándar.

NOTA: Como queremos un modelo eficiente solo queremos considerar aquellos retrasos que tengan un efecto directo y significativo sobre el período presente. Por lo tanto, debemos examinar la PACF (Auto-Correlación Parcial) antes de construir un modelo con demasiados coeficientes de retraso.

Los modelos autorregresivos son modelos adecuados para series estacionarias y el coeficiente Φ_0 está relacionado con el nivel constante de la serie. Si los datos varían alrededor de cero o se expresan como desviaciones de la media $Y_t - \bar{Y}$, no se requiere el coeficiente Φ_0.

3.3.2. Modelos de Promedios Móviles MA(q)

Los modelos de medias móviles predicen el valor actual de nuestra serie temporal en función de los residuos pasados.
Un modelo de promedio móvil de q-ésimo orden tiene la forma:

$$Y_t = \mu + \varepsilon_t - \omega_1 \varepsilon_{t-1} - \omega_2 \varepsilon_{t-2} - \cdots - \omega_q \varepsilon_{t-q}$$

Donde:

- Y_t = Variable de respuesta (dependiente) en el tiempo t.
- μ= Promedio constante en el proceso.
- $\omega_1, \omega_2, \ldots, \omega_q$= Coeficientes que se estimarán.
- ε_t=es el término de error en el tiempo t, el cual presenta los efectos de variables no explicadas por el modelo; los supuestos acerca del término de error son los mismos que los del modelo de regresión estándar.
- $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-q}$= Errores en períodos anteriores que, para el tiempo t, se incorporan a la respuesta, Y_t

De lo anterior, podemos concluir lo siguiente:

- **Los modelos de promedio móvil (MA)** permiten hacer prónosticos de Y_t con base en una combinación lineal de un número finito de errores pasados.
- Los pesos $\omega_1, \omega_2, \ldots, \omega_q$ no necesariamente suman 1 y pueden ser positivos o negativos, aun cuando estén precedidos de un signo negativo en la especificación del modelo.
- **Los modelos autorregresivos (AR)** pronostican Y_t como una función lineal de un número finito de valores pasados de Y_t

3.3.3. Modelos de promedio móvil autorregresivos (ARMA(p,q))

Un modelo con términos autorregresivos se puede combinar con un modelo que tenga términos de promedio móvil para obtener un modelo "mixto" de promedio móvil autorregresivo. Para representar estos modelos, es conveniente utilizar la notación ARMA (p,q), donde p es el orden de la parte autorregresiva . Para representar estos modelos, es conveniente utilizar la notación ARMA (p,q) donde p es el orden de la parte autorregresiva y q es el orden de la parte del promedio móvil. Un modelo ARMA (p,q) tiene la forma general:

$$Y_t = \Phi_0 + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \cdots + \Phi_p Y_{t-p} + \varepsilon_t - \omega_1 \varepsilon_{t-1} - \omega_2 \varepsilon_{t-2} - \cdots - \omega_q \varepsilon_{t-q}$$

Los modelos ARMA (p,q) pueden describir una variedad de comportamientos de las series de tiempo estacionarias. Los pronósticos generados por un modelo ARMA (p,q) dependen de los valores actuales y pasados de la respuesta, Y, así como de los valores actuales y pasados de los errores (residuos).

Los número de términos autorregresivos y de promedio móvil (orden p y orden q) en un modelo ARMA están determinados por los patrones de las autocorrelaciones y autocorrelaciones parciales de la muestra y los valores de los criterios de selección del modelo. En la práctica, los valores de p y q rara vez exceden de 2.

¿ Cómo conseguimos una serie estacionaria?

- La manera más usual es "diferenciandola". Esto significa restar el valor previo del valor actual. A veces, dependiendo de la complejidad de la serie, se requiere una o más diferenciaciones.

- El valor d es, por lo tanto, el mínimo número de diferenciaciones que se equieren para obtener una serie estacionaria. Si la serie original ya era estacionaria, entonces $d = 0$
- Un modelo ARIMA es aquel donde la serie original ha sido diferenciada al menos una vez para volverla estacionaria, y luego se le combinan las partes autorregresivas y de medias móviles, de modo que satisface la ecuación anterior.

¿ Cómo encontrar d?

- El orden correcto de diferenciación es la diferenciación mínima requerida para obtener una serie casi estacionaria que oscila alrededor de una media bien definida y la gráfica de la función de correlación llega a cero con bastante rapidez.
- Si las autocorrelaciones son positivas para muchos retraso (10 o más), entonces la serie necesita una mayor diferenciación. Por otro lado, si la autocorrelación del retardo 1 en sí es demasiado negativa, entonces la serie probablemente esté sobrediferenciada.
- Si es el caso en que no puede realmente decidir entre dos órdenes de diferenciación, elija el orden que dé la menor desviación estándar en la serie diferenciada.

3.3.4. Construcción del modelo ARIMA

El método de Box-Jenkins se basa en una estrategia iterativa para la construcción del modelo que consiste en:

1. Selección del modelo inicial (identificación del modelo).
2. Estimación de los coeficientes del modelo (estimación de parámetros).
3. Análisis de los residuos(Verificación del modelo)
4. Elaboración de pronósticos con el modelo.

Si es necesario, el modelo inicial se modifica y el proceso se repite hasta que los residuos indiquen que ya no es necesaria otra modificación. En este punto, estamos en condiciones a utilizar el modelo ajustado para pronosticar.

Pasos de la estrategia

1. Identificación del modelo:

   a) El primer paso en la identificación del modelo es determinar si la serie es estacionaria, es decir, si la serie de tiempo parece variar alrededor de un nivel fijo. Si la serie es NO estacionaria, a menudo se puede convertir en una serie estacionaria por diferenciación. Es decir, la serie original se sustituye por una serie de las diferencias. Se específica entonces un modelo ARMA para la serie diferenciada. En efecto, el analista está modelando cambios en vez de niveles.
   Por ejemplo, suponga que la serie original, Y_t, se incrementa por lo general con el tiempo, pero las primeras diferencias, $\Delta Y_t = Y_t - T_{t-1}$, varían alrededor de un nivel fijo. Sería recomendable modelar las diferencias estacionarias empleando un modelo ARMA de, digamos, orden $p = 1$ y $q = 1$. En este caso, el modelo es:

   $$\Delta Y_t = \Phi_1 \Delta Y_{t-1} + \varepsilon_t - \omega_1 \varepsilon_{t-1}$$
   $$Y_t - Y_{t-1} = \Phi_1 (Y_{t-1} - Y_{t-2}) + \varepsilon_t - \omega_1 \varepsilon_{t-1}$$

   La diferenciación se hace hasta que una gráfica de los datos indica que la serie varía alrededor de un nivel fijo de las autocorrelaciones de la muestra desaparecen con rapidez. El número de diferencias requeridas para alcanzar la estacionalidad se denota por d.
   Los modelos de las series **NO estacionarias** se llaman modelos de promedio móvil integrados autorregresivos y se denotan por ARIMA(p, d, q). Donde:

   - p indica el orden de la parte autorregresiva.
   - d indica el número de diferenciaciones.
   - q indica el orden de la parte de promedio móvil.

   Si la serie original es **estacionaria**, entonces $d = 0$, y los modelos ARIMA se reducen a modelos ARMA. Por consiguiente, de ahora en adelante, la notación ARIMA(p,d,q) se usará para identificar modelos tanto de series **estacionarias** ($d = 0$) como **NO estacionarias** ($d > 0$).

   b) Una vez que se ha obtenido una serie estacionaria, el analista debe identificar la forma del modelo que se utilizará.
   La identificación de la forma del modelo se lleva a cabo comparando las autocorrelaciones y las autocorrelaciones parciales calculadas con los datos de las autocorrelaciones y las autocorrelaciones parciales teóricas de los diferentes modelos ARIMA.
   Es posible que exista alguna ambigüedad en la identificación de un modelo ARIMA adecuado a partir de los patrones de las autocorrelaciones y las autocorrelaciones parciales. Por ello, la selección del modelo incial debería considerarse como tentativa. Se deben realizar análisis durante los pasos 2 y 3 para determinar si el modelo es adecuado. Recuerde que:

- SI las autocorrelaciones de la muestra se desvanecen exponencialmente hacia cero y las autocorrelaciones parciales de la muestra se cortan de forma abrupta, el modelo requerirá términos autorregresivos.
- Si las autocorrelaciones de la muestra se cortan manera abrupta y las autocorrelaciones parciales de la muestra se desvanecen, el modelo requerirá términos de promedio móvil.
- Si tanto las autocorrelaciones de la muestra como las autocorrelaciones parciales de la muestra se desvanecen, tanto los términos autoregresivos como los términos de promedio móvil están indicados.
- Al contar el número de autocorrelaciones y autocorrelaciones parciales significativas de la muestra, pueden determinarse los órdenes de las partes MA y AR.
- Para juzgar su significancia, tanto las autocorrelaciones de la muestra como las autocorrelaciones parciales de la muestra se comparan generalmente con $\pm\frac{2}{\sqrt{n}}$ donde n es el número de observaciones en la serie de tiempo

El *Principio de Parsimonia* se refiere a la preferencia por los modelos sencillos por encima de los modelos complejos. El objetivo es desarrollar el modelo más sencillo que ofrezca una descripción adecuada de las principales características de los datos.

2. Estimación del modelo:

   a) Una vez que se ha seleccionado un modelo tentativo, se deben estimar los parámetros para ese modelo.
   Los Parámetros de los modelos ARIMA se estiman minimizando la suma de cuadrados de los errores de ajuste. Estos mínimos cuadrados estimados deben obtenerse, en general, usando un procedimiento NO lineal de mínimos cuadrados. Un procedimiento NO lineal de mínimos cuadrados es simplemente un algoritmo que obtiene el mínimo de la función de la suma de los errores al cuadrado. Una vez que se determinan las estimaciones de los mínimos cuadrados y sus errores estándar, se pueden construir e interpretar los valores t del modo habitual.
   - Los parámetros que son considerados significativamente diferentes de 0 se conservan en el modelo ajustado.
   - Los modelos que no son significativos se eliminan del modelo.

   b) Se calcula el error cuadrático medio de los residuos, una estimación de la varianza del error, ε_t.
   El error cuadrático medio de los residuos se define como:

$$s^2 = \frac{\sum_{t=1}^{n} \varepsilon_t^2}{n-r} = \frac{\sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2}{n-r}$$

   Donde:

$$\varepsilon_t = Y_t - \hat{Y}_t = \textit{Residuo para el tiempo } t$$
$$n = \textit{número de residuos.}$$
$$r = \textit{número total de parámetros estimados}$$

   c) El error cuadrático medio de los residuos es útil para evaluar el ajuste y comparar modelos diferentes. También es útil para calcular los límites de error de pronóstico.

3. Verificación del modelo:
   Antes de utilizar el modelo para pronosticar, debe verificarse que tan adecuado es. En esencia un modelo es adecuado si los residuos no se pueden usar para mejorar los pronósticos. Es decir, los residuos deben ser aleatorios.

   a) Muchas gráficas residuales que son útiles en el análisis de regresión se pueden desarrollar para los residuos de un modelo ARIMA. Un histograma y una gráfica de probabilidad norma (para verificar la normalidad), así como una gráfica de la secuencia del tiempo (para verificar los valores atípicos) son particularmente útiles.

b) Las autocorrelaciones residuales individuales deben ser pequeñas y generalmente dentro de $\pm\frac{2}{\sqrt{n}}$ de cero. Las autocorrelaciones residuales significativas en retrasos cortos o retrasos estacionales sugieren que el modelo es inadecuado y que debe seleccionarse un nuevo modelo o modificarse.

c) Las autocorrelaciones residuales como un grupo deben ser congruentes con aquellas producidas por los errores aleatorios
.Una verificación general de la **idoneidad** del modelo que se realiza mediante una prueba de distribución Chi Cuadrada (χ^2) con base en el estadístico Q de Ljung-Box. Esta prueba considera los ta,años de las autocorrelaciones residuales como un grupo. El estadístico de prueba Q es:

$$Q = n(n+2)\sum_{k=1}^{m}\frac{r_k^2(\varepsilon)}{n-k}$$

el cual esta distribuido aproximadamente como una variable aleatoria chi cuadrada con $m-r$ grados de libertad, donde r es el número total de parámetros estimados en el modelo ARIMA. Donde:

- $r_k(\varepsilon)$= autocorrelación residual para el retraso k.
- n= número de residuos.
- k=retraso de tiempo.
- m= número de retrasos de tiempo que van a ser evaluados.

Si el valor p asociado con el estadístico Q es pequeño (digamos, el valor $p < 0.05$), el modelo se considera inadecuado. El analista debe considerar un nuevo modelo o uno modificado y continuar el análisis hasta encontrr un modelo satisfactorio.

4. Elaboración de pronósticos con el modelo:

a) Una vez que se ha encontrado un modelo adecuado, es factible elaborar los pronósticos de uno o varios períodos futuros.
Con base en los pronósticos también se pueden construir intervalos de predicción. En general, para un nivel de confianza dado, cuanto más largo sea el tiempo guía del pronóstico, mayor será el intervalo de predicción. Esto es razonable, puesto que se espera que la incertidumbre sea mayor para el pronóstico de un valor distante que para el pronóstico de, digamos, la siguiente observación.

b) Conforme más datos están disponibles, se puede usar el mismo modelo ARIMA para generar pronósticos modificados de otro origen de tiempo.

c) Si el patrón de la serie parece estar cambiando con el tiempo, los datos nuevos se pueden usar para recalcular los parámetros del modelo o, si es necesario, desarrollar un modelo completamente nuevo.