Curso

10

MACHINE LEARNING & BIG DATA

MODELOS ARIMA

Series Temporales

JOSÉ NELSON ZEPEDA DOÑO

Cluster de Estudio: Advanced Analytics

Este material es el resumen de muchos autores que por medio de sus libros y documentos nos ofrecen fuentes riquísimas de conocimiento sobre los temas de Big Data y Machine Learning.
Algunas citas, figuras y tablas pueden ser encontradas de forma textual tal como lo indica el autor en su material original.

Nelson Zepeda
MIP • V 1.0
San Salvador El Salvador
Phone 503 79074137 • @nelsonzepeda733

Tabla de Contenido

Capítulo 1

Conceptos de ARIMA[1]

A finales del siglo XX diversos investigadores comenzaron a desarrollar y aplicar nuevas propuestas para modelación de series de tiempo.

Las nuevas propuestas de análisis de series de tiempo se empezaron a implementar en un principio a problemas de contaminación, en la economía, a enfermedades epidemiológicas y en la actualidad a fenómenos físicos y sociales.

Específicamente los modelos desarrollados en las últimas dos décadas son los llamados autorregresivos (AR), de medias móviles (MA), integrados (I), así como sus posibles combinaciones (ARIMA).

La principal diferencia entre estos modelos y los clásicos es el enfoque estocástico que se le da a las series de tiempo, en vez de tratarla de forma determinística. Bajo este enfoque se concibe la serie de tiempo como un conjunto de valores de tipo aleatorio, generados a partir de un proceso totalmente desconocido.

Así mismo, derivado del conocimiento del proceso generador de datos, el objetivo de este enfoque es tratar de identificar el modelo probabilístico que represente las características principales del comportamiento de la serie.

El desarrollo y aplicación de este tipo de modelos de series de tiempo en un principio estuvieron gravemente limitados, fundamentalmente debido a razones de cómputo. Actualmente con la amplia disponibilidad de recursos, el uso de modelos ARIMA se ha tornado posible.

Procesos Estocásticos

Un proceso estocástico es una sucesión de variables aleatorias Yt ordenadas, pudiendo tomar t cualquier valor entre menos infinito e infinito.

En general las series de interés llevan asociados fenómenos aleatorios. Por este motivo el estudio de su comportamiento pasado solo permite acercarse a la estructura o modelo probabilístico para la predicción del futuro. La teoría de los procesos estocásticos se centra en el estudio y modelización de sistemas que evolucionan a lo largo del tiempo, o del espacio, de acuerdo a unas leyes no determinísticas, esto es, de carácter aleatorio y probabilístico.

Bajo este enfoque, una serie de tiempo se define como una observación particular sobre el estado dinámico de una variable de un proceso de naturaleza aleatoria. Simultáneamente se presupone que los procesos de series de

[1] Tomado de: http://www.ptolomeo.unam.mx:8080/xmlui/bitstream/handle/132.248.52.100/363/A7.pdf?sequence=7

tiempo siguen un comportamiento aleatorio y debido a la propia dinámica de dicho proceso, tales impulsos pueden generar un conjunto de datos serialmente dependientes.

Así, mediante la identificación de la estructura de dependencia latente existente entre las observaciones, se puede modelar la serie.

Una de las ventajas principales de este enfoque, es la gran flexibilidad que se logra para representar un buen número de fenómenos reales mediante una sola clase general de modelos. Otra ventaja es la facilidad y precisión para realizar pronósticos.

Tipos de Modelo

La estructura existente de dependencia entre los datos en un proceso estocástico, puede ajustarse a seis modelos que pueden describir una gran variedad de fenómenos reales. La principal característica de dichos modelos es que no involucran a las variables independientes en su construcción. En cambio, emplean la información que se encuentra en la serie misma para generar los pronósticos.

- Proceso de ruido blanco (a_t): es un proceso formado por una secuencia de variables aleatorias mutuamente independientes e idénticamente distribuidas.

- Modelo no estacionario de corrido aleatorio (I): también conocido en inglés como Random Walk, corresponde a aquellas situaciones en las que los impulsos aleatorios tienden a sumarse o integrarse en el tiempo. La integración refleja la presencia de un componente de tendencia.

- Modelo autorregresivo de orden p AR(p): Se trata de un modelo en el que una determinada observación es predecible a partir de la observación anterior (modelo autorregresivo de primer orden) o a partir de las dos observaciones que les preceden (modelo autorregresivo de segundo orden). En este caso, la observación actual se define como la suma ponderada de una cantidad finita p de observaciones precedentes más un impulso aleatorio independiente.

  Matemáticamente, un modelo autorregresivo tiene la siguiente forma:

  $$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \varphi Y_{t-3} + \cdots + \varphi_p Y_{t-p} + a_t$$

  En este caso, las variables independientes son valores de la misma variable (de aquí el nombre auto), pero de periodos anteriores (t-1, t-2, t-3, t-p). at es el error, o termino residual, que representa perturbaciones aleatorias que no pueden ser explicadas por el modelo.

  El modelo se llama autorregresivo porque se asemeja a la ecuación de regresión:

  $$Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \cdots + b_p X_p + e$$

- Modelo de Medias Móviles de orden q MA(q): En este modelo, una determinada observación está condicionada por los impulsos aleatorios de las observaciones anteriores. De esta forma la observación actual se define como la suma del impulso actual y de los impulsos anteriores con un determinado peso

$$Y_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$$

Donde at es el residuo o error en el periodo t, la ecuación es semejante a la anterior, con la excepción de que implica que la variable dependiente Yt depende de los valores previos del término residual más que de la variable misma.

- Modelo autorregresivo de medias móviles de orden p,q ARMA(p,q): Este modelo es la combinación de las estructuras anteriores: modelo autorregresivo y modelo de medias móviles. Así, una observación está determinada tanto por observaciones anteriores así como por impulsos aleatorios o también llamados errores de observaciones pasadas.

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \varphi_3 Y_{t-3} + \cdots + \varphi_p Y_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$$

Se observa que la ecuación es una combinación de las ecuaciones del modelo de AR y MA.

Un modelo ARMA puede ajustarse a cualquier patrón de datos. Sin embargo, los valores de p y q se deben especificar.

Por ejemplo si p=1 y q=0, el modelo ARMA es $Y_t = \varphi_1 Y_{t-1} + a_t$ es decir es un ARMA(1,0).

Si p=2 y q=0, el modelo ARMA es $Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + a_t$ es decir es un ARMA(2,0)

Si p=0 y q=1 el modelo ARMA es $Y_t = a_t - \theta_1 a_{t-1}$ es decir es un ARMA(0,1)

Es importante destacar que la ecuación ARMA no es lineal y por lo tanto suficientemente general para describir una amplia variedad de patrones de datos.

Hay que mencionar que p y q pueden adoptar muchos valores, pero raramente son mayores a 2.

- Modelo autorregresivo integrado de medias móviles de orden p,d,q ARIMA(p,d,q): al igual que un modelo ARMA, es la combinación de los modelos autorregresivo y el de medias móviles, con la particularidad de incluir un proceso de restablecimiento el cual se denomina integración. La forma general de un modelo ARIMA es semejante a la de un modelo ARMA:

$$Y'_t = \varphi_1 Y'_{t-1} + \varphi_2 Y'_{t-2} + \varphi_3 Y'_{t-3} + \cdots + \varphi_p Y'_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$$

Donde Y'_t: *Es la serie inducida a la estabilidad* .

Metodología Box-Jenkins

Existen distintas metodologías que ayudan a elegir el modelo que mejor se acople a la serie de tiempo de entre los diversos modelos existentes. La metodología más utilizada y difundida es la que propusieron los profesores G.E.P. Box y G.M. Jenkins en la década de los años 70, en la cual lograron grandes avances en identificar, ajustar y verificar los modelos ARIMA adecuados. Comúnmente se le conoce como Metodología Box-Jenkins. Se especifica en esta metodología únicamente el uso de los modelos ARIMA ya que los modelos AR y MA se dicen que son las formas particulares de la clase general de modelos ARMA y este a su vez del modelo ARIMA.

Esta metodología se basa en tratar de determinar cuál es el modelo probabilístico que rige el comportamiento del proceso a lo largo del tiempo. Cabe apuntar en este momento la diferencia entre proceso y modelo. Se puede decir que un proceso es lo real, es decir, el fenómeno en sí, del cual se desconoce su mecanismo generador. Por otro lado un modelo es solo la imitación o representación del proceso.

Consideraciones importantes en la modelación ARIMA:

- Los modelos ARIMA aplican tanto para datos discretos como continuos.

- Aunque la metodología ARIMA trata tanto con datos discretos como continuos, solo se puede aplicar a datos espaciados equidistantemente en el tiempo, en intervalos discretos de tiempo. Los datos medidos en intervalos discretos de tiempo pueden clasificarse en dos tipos: 1) datos que son producto de la acumulación durante un periodo de tiempo, por ejemplo los ahorros de una persona en un mes. 2) datos que son producto de la medición instantánea periódicamente, por ejemplo la medición de la presión en una tubería en intervalos de una hora.

- Para elaborar un modelo ARIMA se requiere una cierta cantidad de datos mínimos. Los profesores Box y Jenkins sugieren un mínimo de 50 observaciones. Un modelo ARIMA se puede aplicar a una serie de menor tamaño, realizando con mucho cuidado su interpretación. Para series con patrones estacionales se aconseja una serie con un gran número de muestras observadas.

- Los modelos ARIMA son especialmente útiles en el tratamiento de series que presentan patrones estacionales.

- El método Box-Jenkins aplica a series estacionarias y no estacionarias. Una serie estacionaria es aquella cuya media, varianza y función de autocorrelación permanecen constantes en el tiempo.

- Se asume que las perturbaciones aleatorias presentes en la serie son independientes entre sí, no existe correlación entre ellas, por lo tanto ningún patrón modelable.

La metodología Box-Jenkins se compone de 5 etapas:

1. Estacionariedad

2. Identificación

3. Estimación

4. Evaluación

5. Pronostico

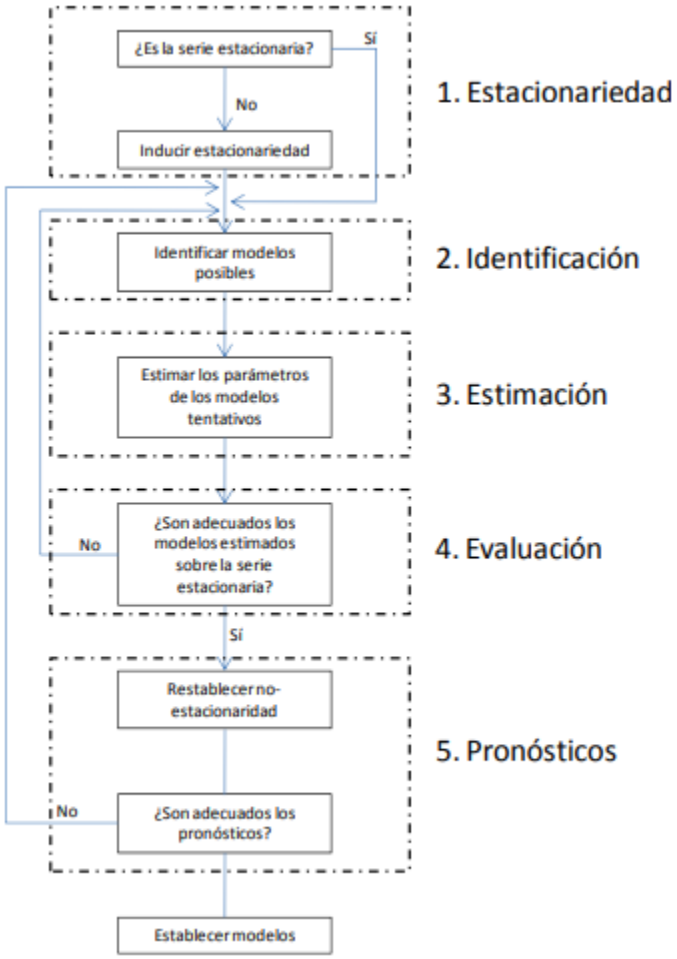

Fig 1-1 Metodología Box-Jenkins

Estacionariedad

Los procesos estocásticos se clasifican entre estacionarios y no estacionarios. La idea de la Estacionariedad está relacionada con la estabilidad de la serie. Un proceso estacionario se describe como una secuencia de datos o valores que no presentan cambio en la media ni cambio en la varianza, por lo cual se dice que la serie es estable.

En un proceso no estacionario, la serie de datos es inestable en el tiempo.

Dicho lo anterior y en palabras prácticas, se puede afirmar que una serie cuyos valores fluctúan respecto a una media constante, es decir sin tendencia, es una serie estacionaria.

En caso de que la serie de tiempo no sea estable en el tiempo, es necesario aplicar una transformación a la serie para inducirla a ello.

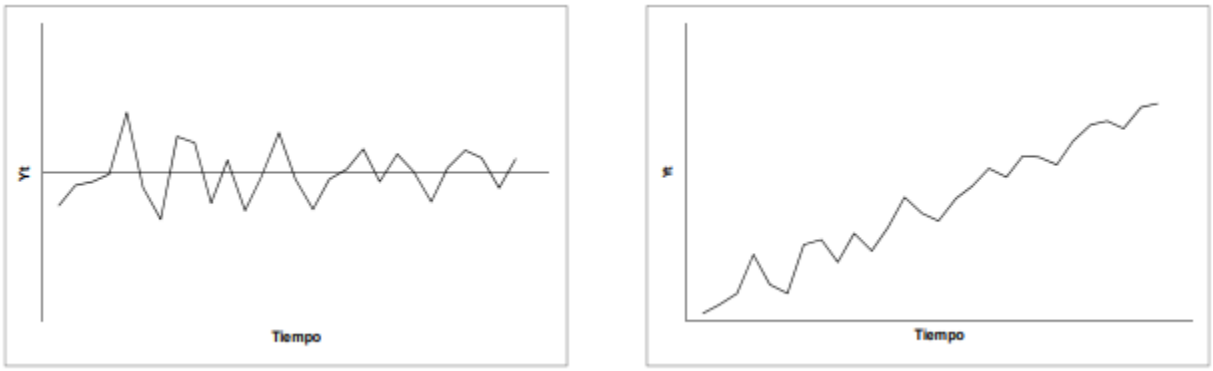

Fig 1-2 Serie de tiempo estacionaria vs no estacionaria

Adicional a la estrategia gráfica, es usual recurrir a la función de autocorrelación simple (FAS), en especial aquellas series con tendencias poco remarcada.

A continuación se muestran los diferentes escenarios que se pueden obtener con la FAS:

- La FAS se puede cortar o truncar

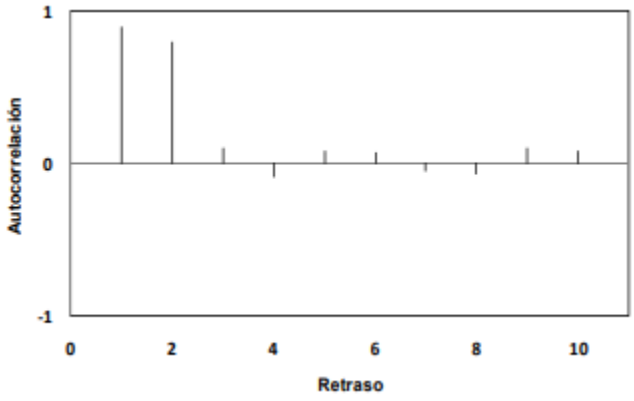

Fig 1-3 FAS se trunca

- La FAS decae ya sea de forma exponencial, amortiguada o bien una combinación de estas.

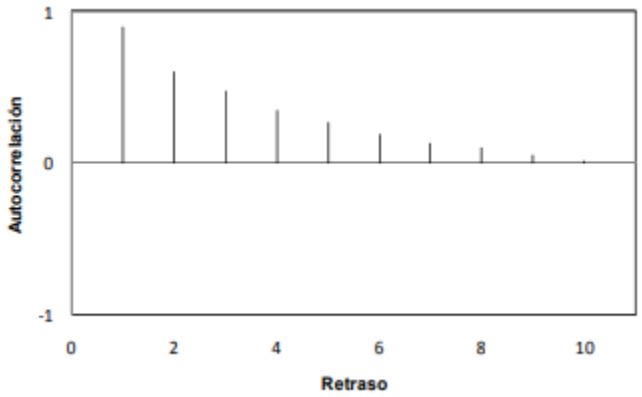

Fig 1-4 FAS cae en forma exponencial

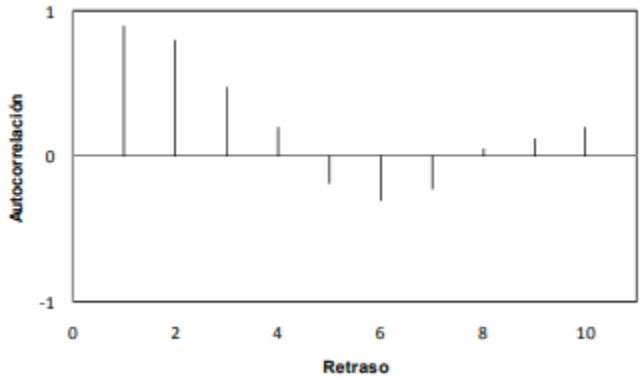

Fig 1-5 FAS se extingue en forma de seno-amortiguada

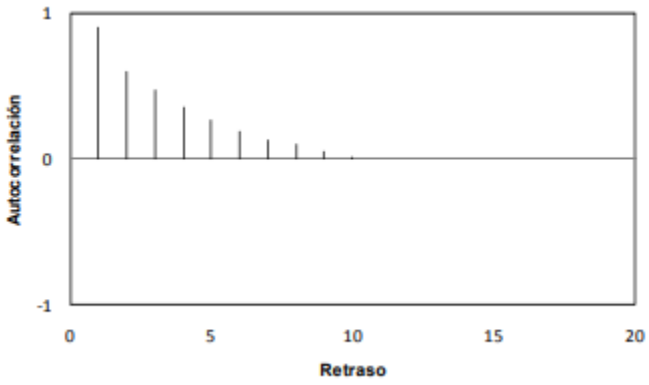

Fig 1-6 se extingue con rapidez

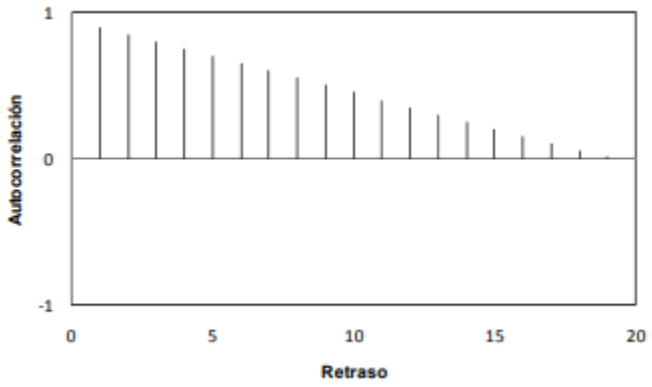

Fig 1-7 se extingue con lentitud extrema

Ahora bien, dada cada forma de las gráficas se puede inferir lo siguiente:

- Si la FAS se corta con rapidez, entonces se debe considerar que los valores de la serie temporal son estacionarios.

- Si la FAS se corta con lentitud extrema, entonces se debe considerar que los valores de la serie temporal son no estacionarios.

Las series de tiempo normalmente no se comportan establemente, es decir no son procesos estocásticos estacionarios. Cuando se presenta esa situación es necesario aplicar una transformación a fin de tener una serie temporal estacionaria.

La estrategia más utilizada para aplicar esta transformación es la de construcción de diferencias, este método consiste, como su nombre lo indica, en obtener diferencias entre los mismos valores de la serie con el fin de remover cualquier patrón de tendencia.

Serie de datos	Primeras diferencias	Nueva serie
2	4-2 = 2	2
4	6-4 = 2	2
6	8-6 = 2	2
8	10-8 = 2	2
10	12-10 = 2	2
12	14-12 = 2	2
14		-

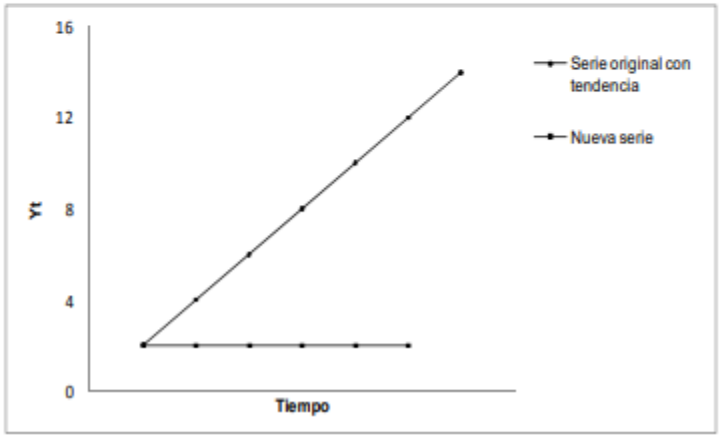

Fig 1-8 ejemplo de transformación mediante construcción de diferencias

La aplicación de esta transformación nos dará como resultado una serie estacionaria tal como se mencionó al inicio de esta sección, lo abarcado hasta este momento nos permitirá establecer la diferencia principal entre ARMA y ARIMA: *un modelo ARMA es capaz de operar únicamente sobre series estacionarias, mientras que un modelo ARIMA es capaz de operar tanto series estacionarias como no estacionarias.*

Identificación

Una vez que se asegura que la serie es estable en el tiempo, el siguiente paso en la metodología Box-Jenkins es la identificación del modelo que rige el proceso de la serie de tiempo.

La identificación del modelo se realiza por medio de la comparación de las FAS

Modelo	FAS
AR	Decae a cero
MA	Se trunca o se corta (después del retraso q)
ARMA	Decae a cero

Estimación

En esta etapa se estiman los coeficientes del modelo escogido en el paso anterior, el criterio más utilizado es el de mínimos cuadrados o Least Squares (LS en inglés).

La estimación de los parámetros se realiza por medio de minimizar la suma de los cuadrados de los residuales SSR (sum of squared residuals).

Adicional a la búsqueda de parámetros dentro del paso de estimación también se tiene en cuenta:

- El modelo debe tener parsimonia, el objetivo en la modelación ARIMA no es encontrar el modelo exacto que represente al proceso generador de las observaciones, es más bien encontrar el modelo que se aproxime al verdadero proceso y que explique el comportamiento de la variable en forma práctica.

- El modelo debe ser estacionario e inversible.

- Los coeficientes del modelo deben ser estadísticamente significativos, por ejemplo el valor t, como regla practica se establece que si el valor absoluto de t es mayor que 2, el parámetro que estamos evaluando es bueno, caso contrario lo más recomendable es excluir ese parámetro, por otro lado también se puede evaluar el valor de p, cuanto más pequeño es el valor de p, más fiable es el resultado, en la practica un valor de $p<0.05$ indica que el resultado es importante.

Evaluación del modelo

En este paso se comprueba la eficiencia del modelo y se evalúa si es estadísticamente adecuado.

La función de autocorrelación simple de los residuales es el instrumento que se utiliza para determinar si el modelo es estadísticamente adecuado. Si los residuales muestran estar correlacionados entre sí, significa que existe un patrón que aún no ha sido tomado en cuenta por los términos autorregresivos y/o medias móviles del modelo propuesto, por lo tanto se debe buscar otro modelo cuyos residuales sean completamente aleatorios.

Por medio de un gráfico y sus bandas se puede establecer si el valor de un coeficiente es significativo. Si uno o más valores sobrepasan las bandas de dos desviaciones estándar, significa que los residuales no son independientes entre sí y que el modelo no es estadísticamente adecuado.

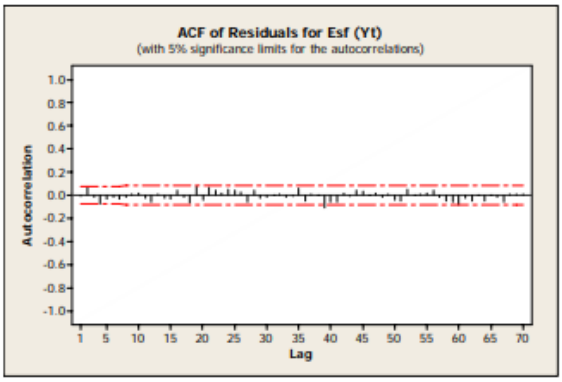

Fig 1-8 Evaluación de los residuales.

Por otro lado no podemos dejar fuera los índices AIC y BIC (Criterio de información de Akaike y criterio de información bayesiano. Dadas dos modelos estimados, el modelo con el menor valor de BIC es el que se prefiere, de la misma forma se evalúa el AIC.

Pronostico

Después que se ha encontrado el modelo adecuado, se pueden llevar a cabo los pronósticos para un periodo, o varios en el futuro.

Si el patrón de la serie parece cambiar en el tiempo, los nuevos datos podrán usarse para volver a estimar los parámetros del modelo o, de ser necesario, desarrollar un modelo completamente nuevo.

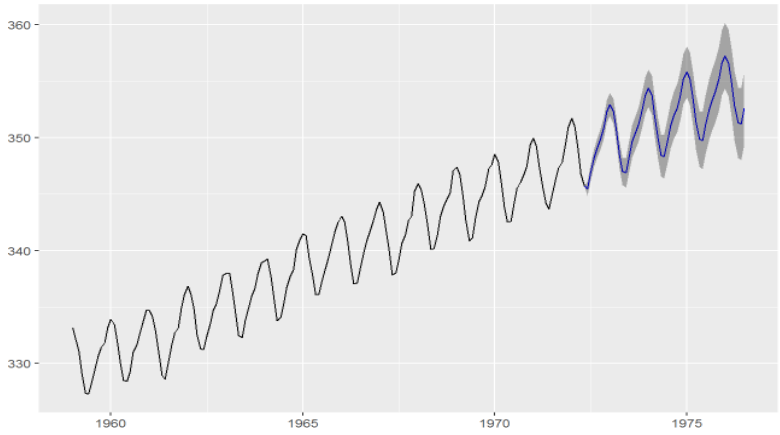

Fig 1-9 Pronóstico.

Bibliografía

- Machine Learning for Beginners
  By Ken Richards, 2017

- R Data Analysis Cookbook
  by Kuntal Ganguly,2017

- Estadística Descriptiva: Series Temporales
  by Santiago de la Fuente Fernández