EGADE Business School
Tecnológico de Monterrey

Aplicaciones de analítica de datos a los negocios II

PROF: JUAN C. BUSTAMANTE

JUCBUSTAM@TEC.MX

Normas para la conexión síncrona:

1. La clases tiene un back-up garantizado (Grabación disponible en la nube de Zoom).
2. Ingresar a la clase con la cámara del equipo de computo encendida.
3. La **cámara deberá permanecer encendida a lo largo de la clase**.
4. Al ingresar a la clase deben silenciar el micrófono del equipo de computo.
5. Levantar la mano es una opción cuando se quiere preguntar algo durante la sesión de clase, pero les recomiendo que mejor hagamos uso intensivo del chat del canal general para hacer preguntas**.**
6. En caso de necesitar hacer una pregunta, puede interrumpir la clase sin problema, activando el micrófono de vuestro equipo de computo, luego de la pregunta desactívelo nuevamente**.**
7. Para una buena clase online es indispensable **el debate, así que foméntelo!!!**.
8. Toda la información se gestiona en CANVAS LMS.

EGADE Business School
Tecnológico de Monterrey

Cronograma de trabajo:

Sesiones	Contenidos	Actividad		Fecha
1	Información general del curso	Utility of classification algorithms		Martes 18/04
2	Algoritmo de regresión logística	Ejecutar script	**Solución caso:** Retention modelling at Scholastic Travel Company (A) and (B)	Martes 25/04
3	Algoritmo Naïve Bayes	Ejecutar script		Martes 02/05
4	Algoritmo k-nearest-neighbors (KNN)	Ejecutar script		Martes 09/05
5	Algoritmo Support vector machine	Ejecutar script		Martes 16/05
6	Algoritmo Decision Trees	Ejecutar script		Martes 23/05
7	Algoritmo Random Forest	Ejecutar script		Martes 30/05
8	Modelo RFM	Ejecutar script	**Solución caso**: CD Now	Martes 06/06
9	Modelo valor de vida del cliente (I)	Ejecutar script		Martes 13/06
10	Modelo valor de vida del cliente (II)	Ejecutar script		Martes 20/06
11	Análisis de series de tiempo	Ejecutar script		Martes 27/06
12	**Proyecto final**	Presentación en equipos		Martes 04/07
	Evaluación final			

1

〈 Series de tiempo

Series de tiempo

Por definición, una serie temporal es una sucesión de observaciones de una variable realizadas a intervalos regulares de tiempo

Llamamos Serie de Tiempo a un conjunto de mediciones de cierto fenómeno o experimento registradas secuencialmente en el tiempo. Estas observaciones serán denotadas por $\{x(t1), x(t2), ..., x(tn)\} = \{x(t) : t \hat{I} T \acute{I} R\}$ con x(ti) el valor de la variable x en el instante ti. Si T = Z se dece que la serie de tiempo es discreta y si T = R se dice que la serie de tiempo es continua.

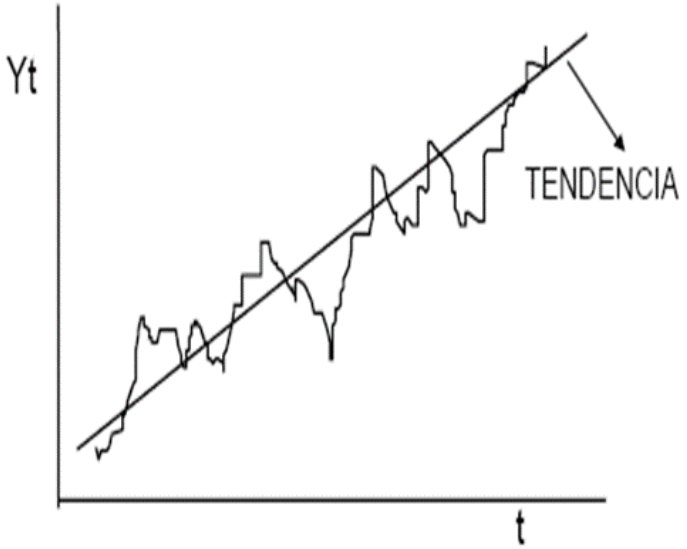
Pronósticos

El pronóstico es la estimación anticipada del valor de una variable, en un lapso de tiempo determinado. Por ejemplo, la demanda de un producto durante el 2024.

Exploración de los patrones de datos y selección de la técnica de pronóstico

Al seleccionar un método de pronósticos adecuado para los datos de series de tiempo, es vital considerar las distintas clases de patrones de datos. Existen cuatro tipos generales: **horizontales o estacionarias, tendencias, estacionales y cíclicos.**

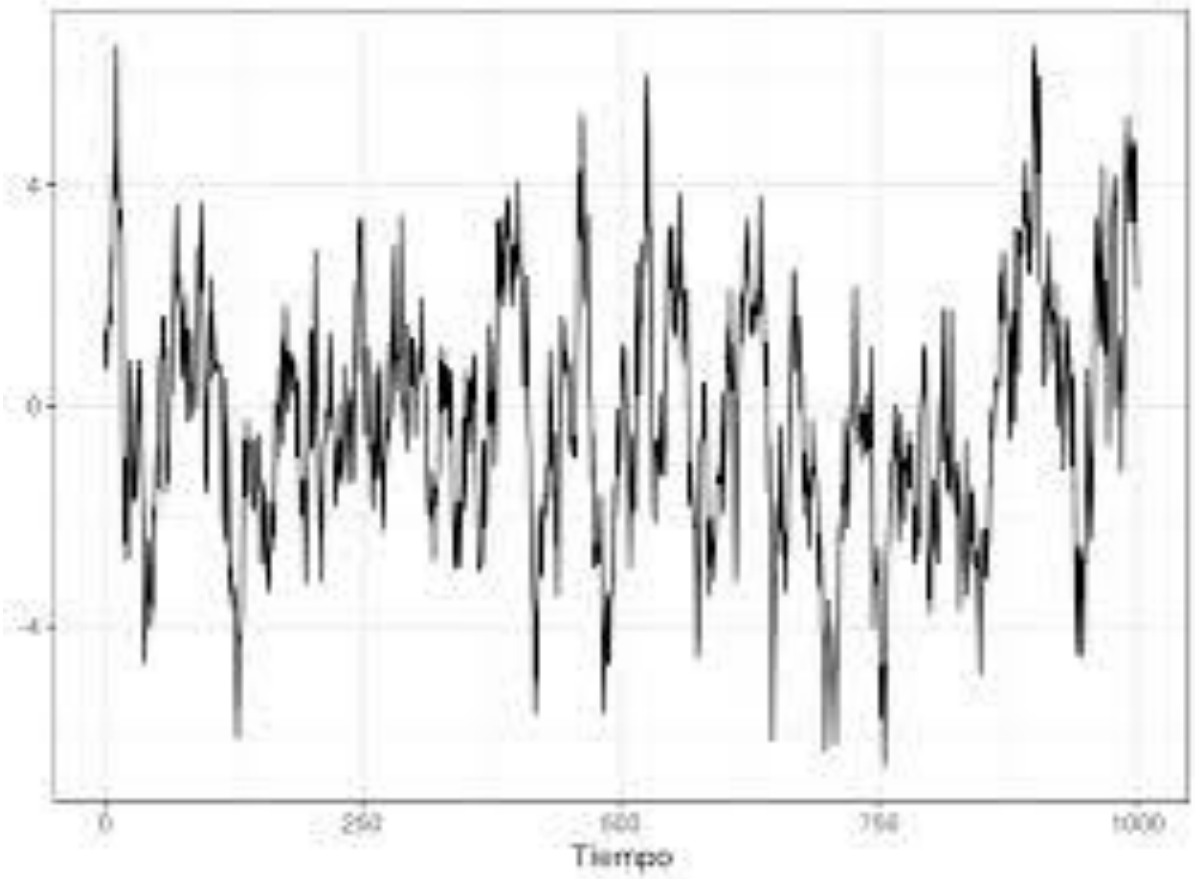
1) COMPONENTE TENDENCIAL (SECULAR): COMPONENTE DE LARGO PLAZO, QUE REPRESENTA EL CRECIMIENTO O DECLINACIÓN DE LA SERIE EN EL TIEMPO.



Exploración de los patrones de datos y selección de la técnica de pronóstico

Al seleccionar un método de pronósticos adecuado para los datos de series de tiempo, es vital considerar las distintas clases de patrones de datos. Existen cuatro tipos generales: **horizontales o estacionarias, tendencias, estacionales y cíclicos.**

Componente estacionario: se refiere a que las propiedades de la serie no varían con respecto al tiempo. En otras palabras, significa que su **variación** (la forma en la que cambia) **no cambia en función del tiempo**.



Exploración de los patrones de datos y selección de la técnica de pronóstico

Al seleccionar un método de pronósticos adecuado para los datos de series de tiempo, es vital considerar las distintas clases de patrones de datos. Existen cuatro tipos generales: **horizontales o estacionarias, tendencias, estacionales y cíclicos.**

3) COMPONENTE ESTACIONAL: PATRÓN DE CAMBIO REGULAR INTRAANUAL QUE SE REPITE EN MANERA SIMILAR EN EL MISMO PERIODO DEL AÑO (TÍPICO DE DATOS FRACCIONADOS EN MENOS DE UN AÑO: CAMBIOS CLIMATICOS, DATOS BASADOS EN CALENDARIOS)

Exploración de patrones de datos mediante análisis de autocorrelación

Cuando se mide una variable a través del tiempo, con frecuencia está correlacionada consigo misma cuando se desfasa uno o más periodos. Esta correlación se mide mediante el coeficiente de autocorrelación. Por tanto, la autocorrelación es la correlación que existe entre una variable retrasada uno o más períodos consigo misma.

La ecuación 1 contiene la fórmula para calcular el coeficiente de autocorrelación (r_k) entre las observaciones Y_t y Y_{t-k}, las cuales se encuentran a k periodos de distancia.

$$r_k = \frac{\sum_{t=k+1} (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2} \qquad k = 0, 1, 2, \ldots$$

Tiempo t	Mes	Datos originales Y_t	Y retrasada un periodo Y_{t-1}	Y retrasada dos periodos Y_{t-2}
1	Enero	123		
2	Febrero	130	123	
3	Marzo	125	130	123
4	Abril	138	125	130
5	Mayo	145	138	125
6	Junio	142	145	138
7	Julio	141	142	145
8	Agosto	146	141	142
9	Septiembre	147	146	141
10	Octubre	157	147	146
11	Noviembre	150	157	147
12	Diciembre	160	150	157

EXPLORACIÓN DE PATRONES DE DATOS MEDIANTE ANÁLISIS DE AUTOCORRELACIÓN

(SERIE ESTACIONARIA)

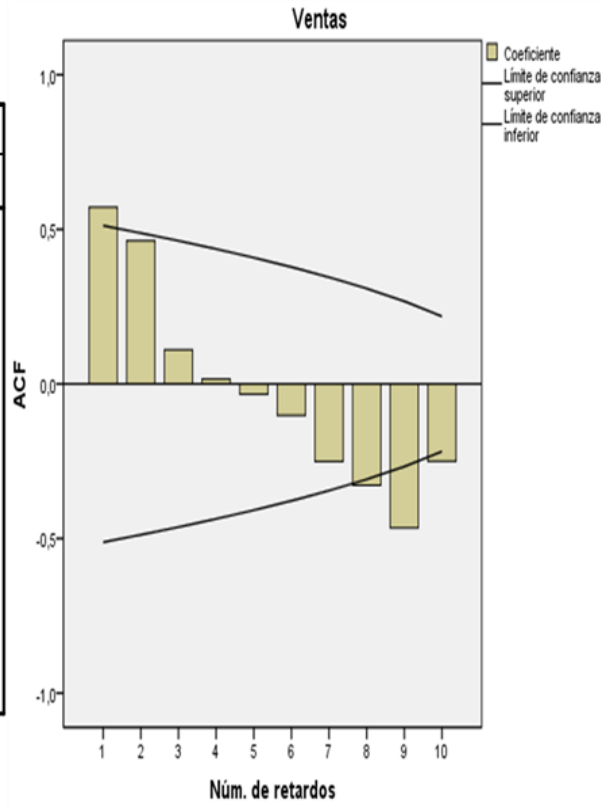
Autocorrelaciones

Serie:Ventas

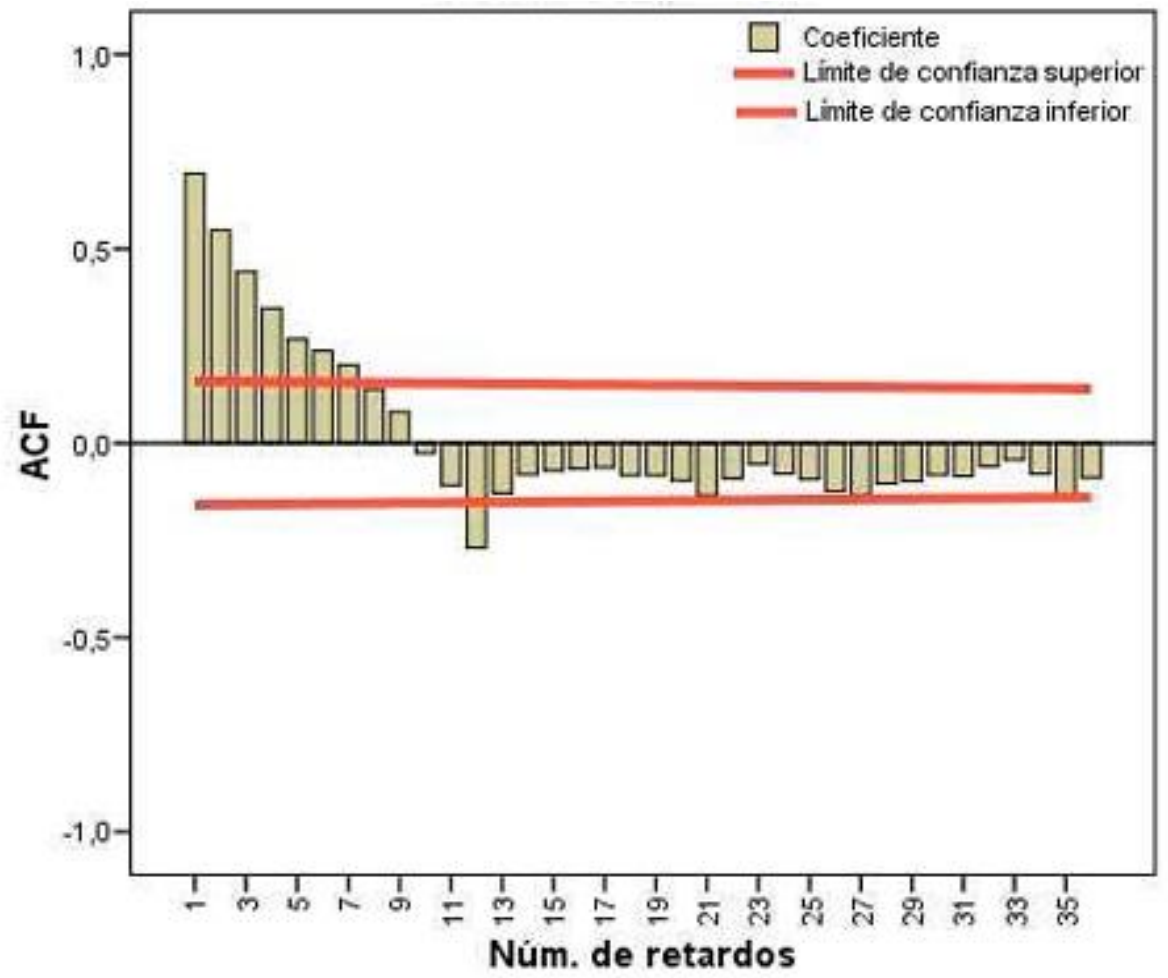
Retardo	Autocorrelación	Tip. Error[a]	Estadístico de Box-Ljung		
			Valor	gl	Sig.[b]
1	,572	,256	4,995	1	,025
2	,463	,244	8,592	2	,014
3	,111	,231	8,820	3	,032
4	,016	,218	8,825	4	,066
5	-,033	,204	8,852	5	,115
6	-,102	,189	9,142	6	,166
7	-,250	,173	11,248	7	,128
8	-,328	,154	15,757	8	,046
9	-,466	,134	27,922	9	,001
10	-,250	,109	33,158	10	,000

a. El proceso subyacente asumido es la independencia (ruido blanco).

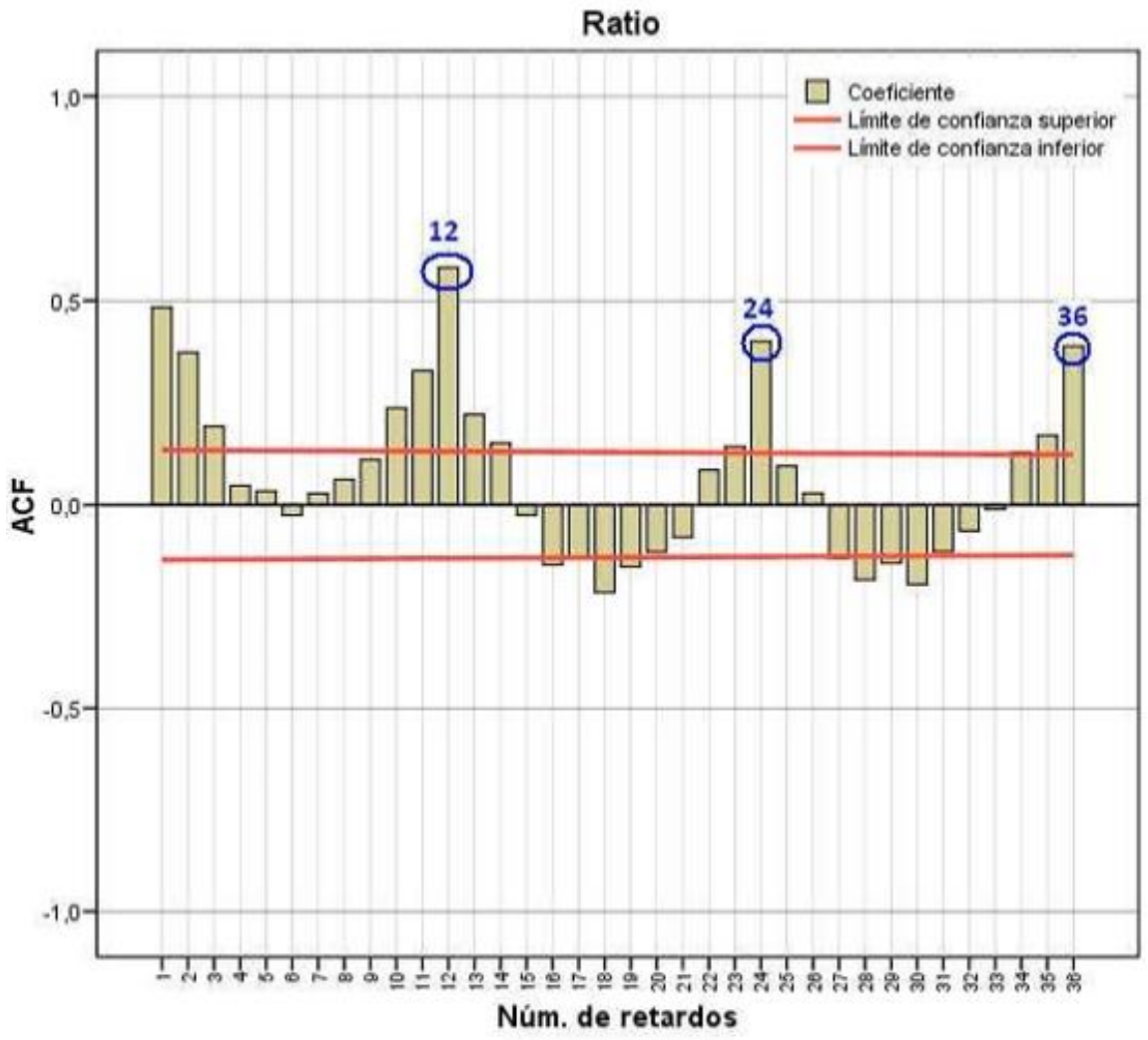
b. Basado en la aproximación chi cuadrado asintótica.



EXPLORACIÓN DE PATRONES DE DATOS MEDIANTE ANÁLISIS DE AUTOCORRELACIÓN (SERIE CON TENDENCIA)



EXPLORACIÓN DE PATRONES DE DATOS MEDIANTE ANÁLISIS DE AUTOCORRELACIÓN (SERIE CON ESTACIONALIDAD)



Selección de la técnica de pronóstico

Tipo de datos	Caracterización	Técnica a utilizar
Datos estacionarios	- Las fuerzas que generan una serie se han estabilizado y el medio en el que existe la serie permanece relativamente sin cambios - Se requiere modelo muy sencillo por falta de datos - Correcciones sencillas que estabilizan el proceso - La serie se puede transformar a una serie estable	- Métodos no formales - Métodos de promedio simple - Métodos de promedios móviles - Atenuación exponencial - Box – Jenkins

Selección de la técnica de pronóstico

Tipo de datos	Caracterización	Técnica a utilizar
Datos con tendencia	- Una productividad creciente y la nueva tecnología conducen a cambios en el estilo de vida - El incremento de la población provoca un incremento en la demanda de bienes - El poder de compra del Bolívar afecta las variables económicas por causa de la inflación - Aumenta la aceptación en el mercado	- Promedio móvil lineal - Atenuación exponencial lineal de Brown - Atenuación exponencial lineal de Holt - Atenuación exponencial cuadrática de Brown - Regresión simple - Modelo de Gompertz - Curvas de crecimiento - Modelos exponenciales

Selección de la técnica de pronóstico

Tipo de datos	Caracterización	Técnica a utilizar
Datos estacionales	- El clima influye en la variable de interés. - El año calendario influye en la variable de interés.	- Descomposición clásica - Atenuación exponencial de Winter - Regresión múltiple de series de tiempo - Métodos Box - Jenkins

Tipo de datos	Caracterización	Técnica a utilizar
Cíclicos	- El ciclo del negocio influye sobre la variable de interés - Se presentan cambios en el gusto popular - Se presentan cambios en la población - Se presentan cambios en el ciclo de vida del producto.	- Descomposición clásica - Indicadores económicos - Modelos econométricos - Regresión múltiple - Regresión logarítmica - Regresión cúbica - Box – Jenkins.

Medición del error de pronóstico

Para calcular el error de pronóstico o residual de cada periodo pronosticado se utiliza la siguiente expresión:

$$\mathbf{e_t} = Y_t - \hat{Y}_t$$

La desviación absoluta media (**MAD**, del inglés *mean absolute deviation*) mide la precisión del pronóstico al promediar las magnitudes de los errores de pronóstico (valores absolutos de cada error). MAD es más útil cuando el analista quiere medir el error de pronóstico en las mismas unidades que la serie original.

$$MAD = \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \hat{Y}_t\right|$$

El error cuadrático medio (**MSE**, del inglés mean squared error) es otro método para evaluar una técnica de pronóstico. Cada error de pronóstico se eleva al cuadrado; luego, se suman y se dividen entre el número de observaciones. Este método penaliza los errores grandes de pronóstico debido a que los errores se elevan al cuadrado, lo cual es importante; una técnica que produce errores moderados podría ser preferible a una que, por lo general tiene errores pequeños, pero que en ocasiones produce errores muy grandes.

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2$$

Cuando es necesario determinar si un método de pronóstico tiene sesgo (produce pronósticos más altos o más bajos de manera sistemática). En estos casos se usa el error porcentual medio (**MPE**, del inglés mean porcentage error). Si el método de pronóstico no tiene sesgo, el MPE producirá un número cercano a cero. Si el resultado es un alto porcentaje negativo, el método sobreestima de forma consistente, y si el resultado es un porcentaje alto positivo, el método subestima consistentemente.

$$MPE = \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)/Y_t$$

El error porcentual absoluto medio (**MAPE**, del inglés mean absolute porcentage error) se calcula al encontrar el error absoluto en cada periodo, dividiéndolo entre el valor real observado para ese periodo y luego promediando los errores porcentuales absolutos. Este método es útil cuando el tamaño o magnitud de la variable del pronóstico es importante para evaluar la precisión del pronóstico. El MAPE proporciona una indicación de cuán grandes son los errores de pronóstico en comparación con los valores reales de la serie.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \hat{Y}_t\right|/Y_t$$

MODELO ARIMA(p, d, q) $(P, D, Q)_s$

Se han analizado las series temporales desde un punto de vista determinista o clásico. A partir de ahora se estudian desde un punto de vista estocástico o moderno, que utiliza métodos más complejos y su aplicación requiere series más largas.

Box y Jenkins han desarrollado modelos estadísticos para series temporales que tienen en cuenta la dependencia existente entre los datos, esto es, cada observación en un momento dado es modelada en función de los valores anteriores. Los análisis se basan en un modelo explícito. Los modelos se conocen con el nombre genérico de **ARIMA** (*AutoRegresive Integrated Moving Average*), que deriva de sus tres componentes **AR** (Autoregresivo), **I**(Integrado) y **MA** (Medias Móviles).

$$y_t^* = \overbrace{\Delta^d y_t}^{\mathbf{I}} \quad \text{serie}$$

$$y_t^* = \mu + \underbrace{\sum_{i=1}^{p} \phi_i y_{t-i}^*}_{\text{AR}} + \underbrace{\sum_{i=1}^{q} \theta_i \epsilon_{t-i}}_{\text{MA}} + \epsilon_t$$

serie diferenciada — constante — error

La metodología de Box y Jenkins se resume en cuatro fases:

- La ***primera fase*** consiste en identificar el posible modelo **ARIMA** que sigue la serie, lo que requiere:
  - Decidir qué transformaciones aplicar para convertir la serie observada en una serie estacionaria.
  - Determinar un modelo **ARMA** para la serie estacionaria, es decir, los órdenes p y q de su estructura autorregresiva y de media móvil.
- La ***segunda fase***: Seleccionado provisionalmente un modelo para la serie estacionaria, se pasa a la segunda etapa de estimación, donde los parámetros AR y MA del modelo se estiman por máxima verosimilitud y se obtienen sus errores estándar y los residuos del modelo.
- La ***tercera fase*** es el diagnostico, donde se comprueba que los residuos no tienen estructura de dependencia y siguen un proceso de ruido blanco. Si los residuos muestran estructura se modifica el modelo para incorporarla y se repiten las etapas anteriores hasta obtener un modelo adecuado.
- La ***cuarta fase*** es la predicción, una vez que se ha obtenido un modelo adecuado se realizan predicciones con el mismo.

Bibliotecas

Manipulación y análisis de datos

Creación de vectores y matrices
Colección de funciones matemáticas

Generación de gráficos

Biblioteca de visualización basada en Matplotlib

Biblioteca de aprendizaje automático

Biblioteca para estimar modelos estadísticos