Universidad del Rosario

Analisis Avanzado de Datos

W16. SARIMA, SARIMAX, VA, VARIMA

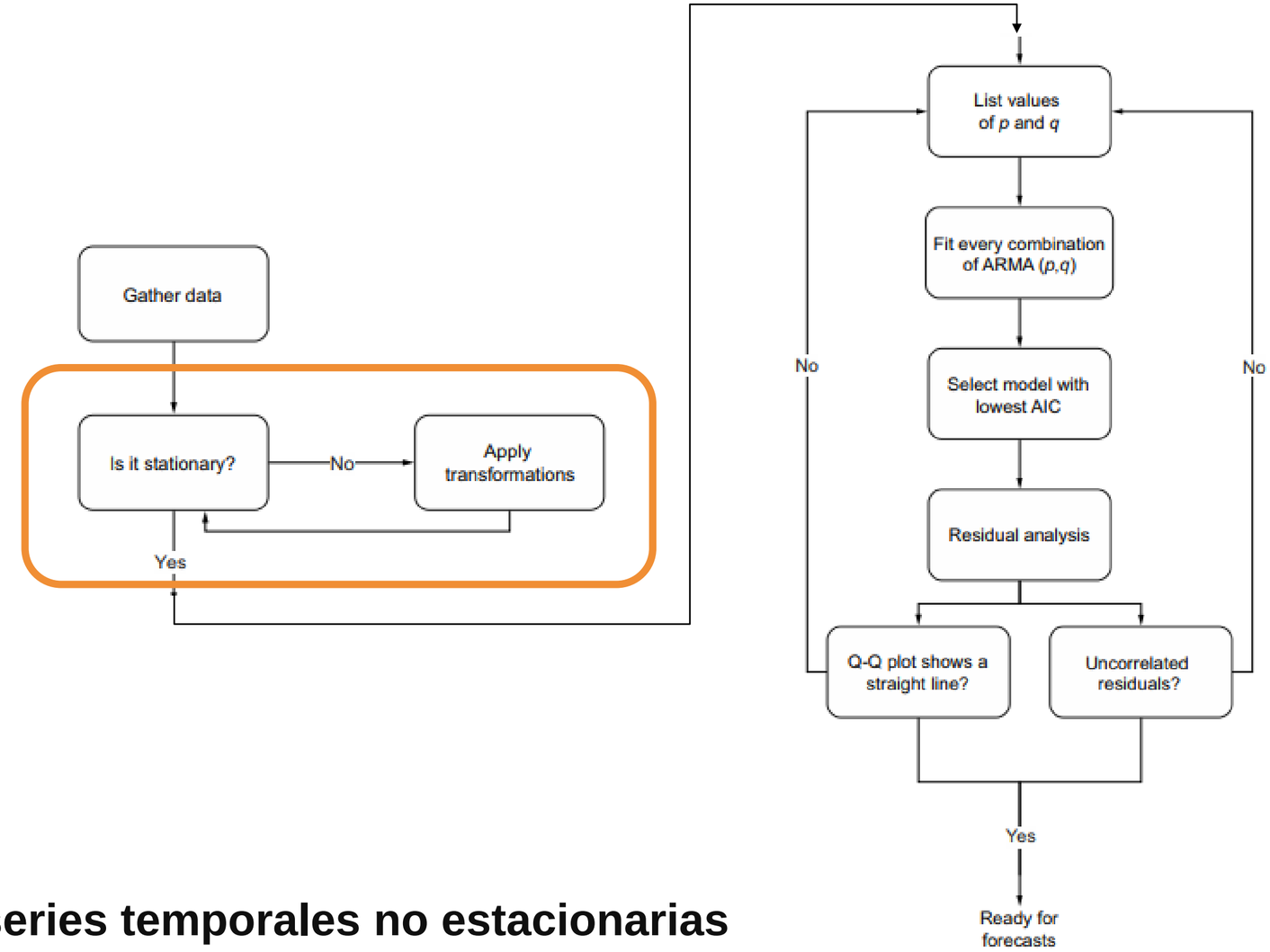
FERNEY ALBERTO BELTRAN MOLINA
Escuela de Ingeniería, Ciencia y Tecnología
Matemáticas Aplicadas y Ciencias de la Computación

ARIMA

Enfoque: Este modelo se centra en describir las autocorrelaciones en los datos de series temporales.

Componentes: Incorpora tres componentes principales: autoregresivo (AR), integrado (I), y promedio móvil (MA).

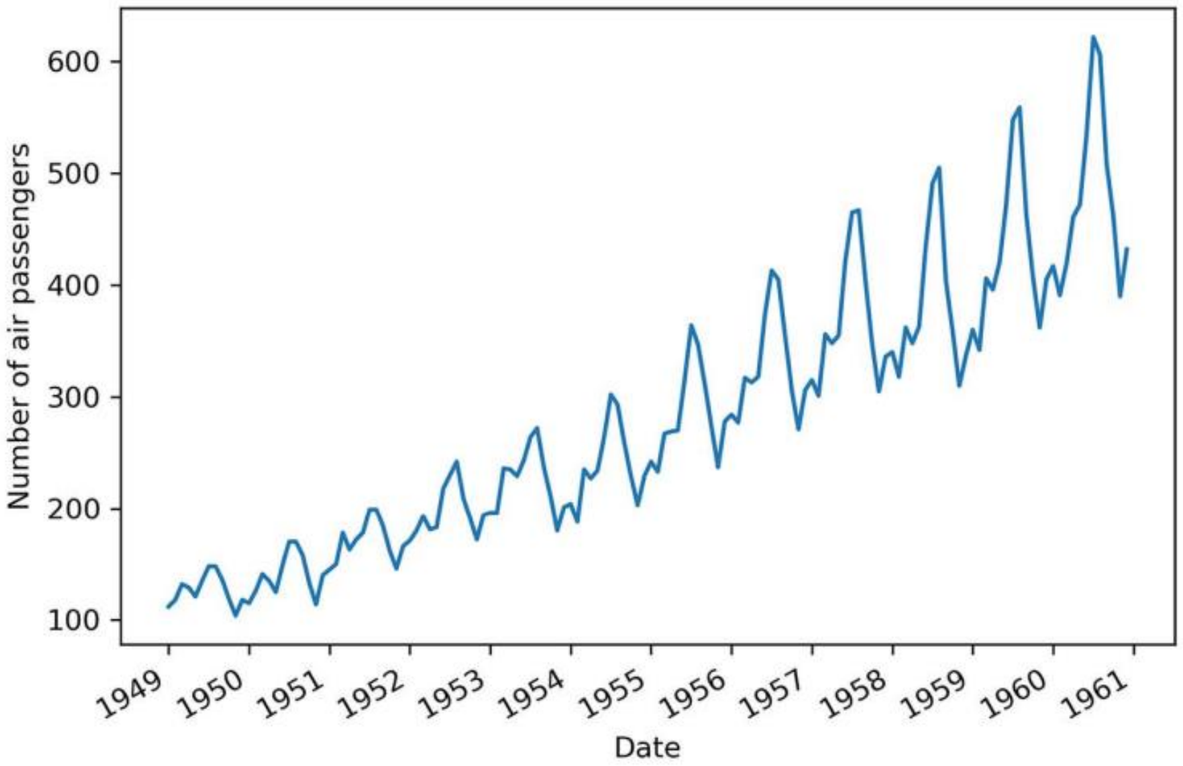
Estacionalidad: ARIMA no tiene un mecanismo específico para modelar la estacionalidad. Es más adecuado para datos que no muestran patrones estacionales fuertes o que han sido previamente desestacionalizados.

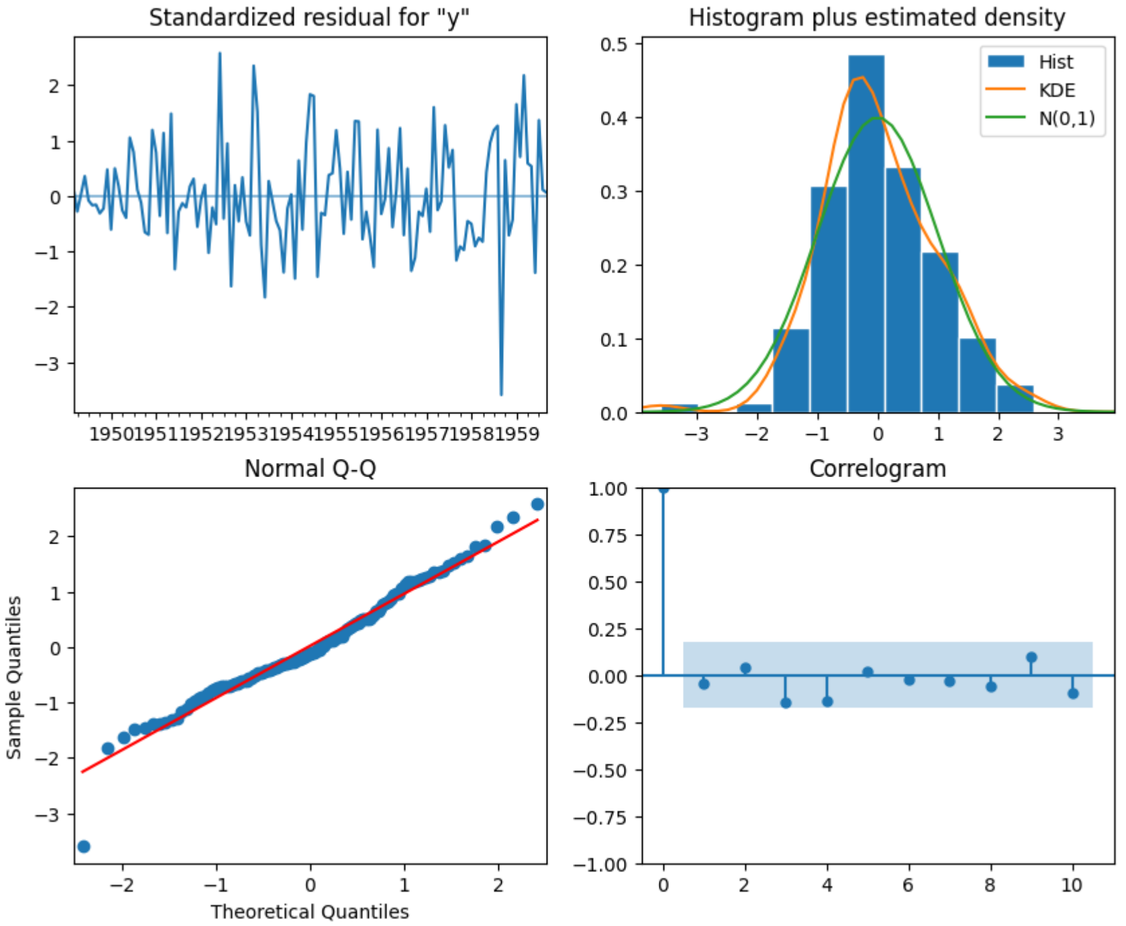


ARIMA(p,d,q), nos permite modelar series temporales no estacionarias

Ejercicio

La presencia de un patrón estacional, con picos en los meses de verano (junio, julio y agosto) y una disminución al principio y al final del año, sugiere la necesidad de un modelo que no solo capture las tendencias y autocorrelaciones, sino también la estacionalidad de los datos. ¿Qué hacemos?



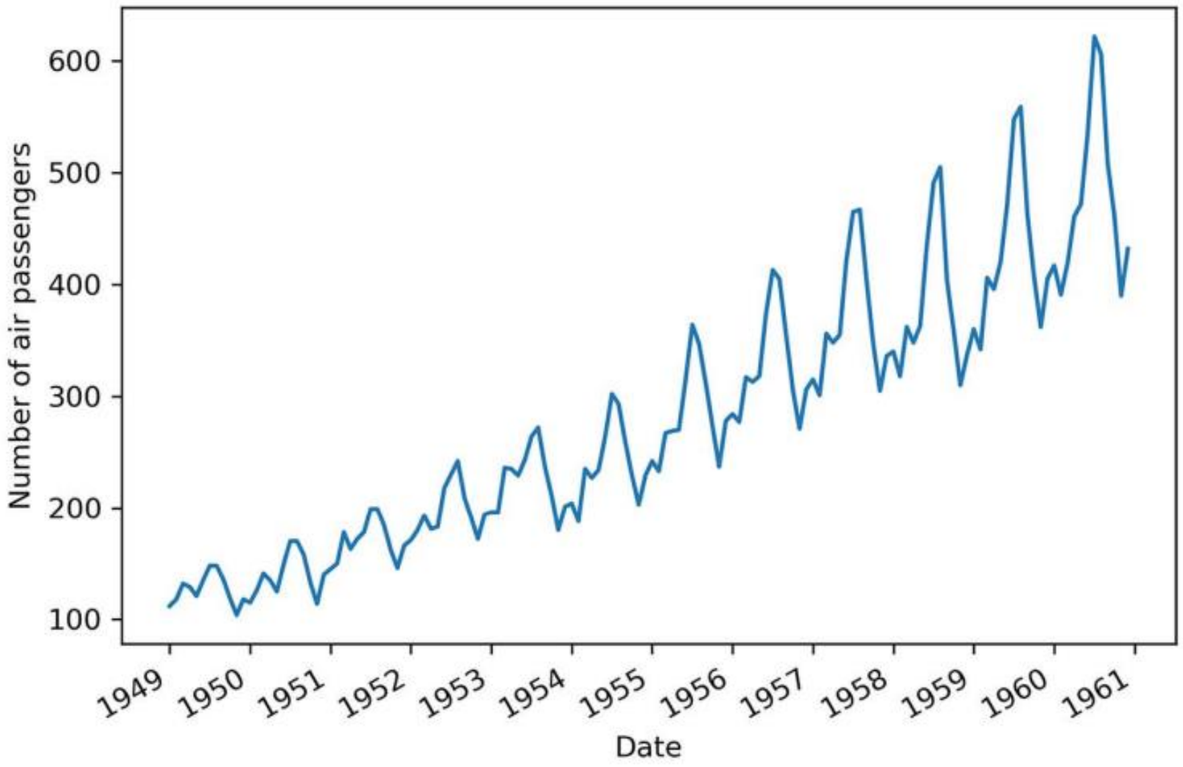


SARIMA (p,d,q)(P,D,Q)m

Enfoque: Es una extensión del modelo ARIMA que añade explícitamente la estacionalidad en el modelo.

Componentes: Además de los componentes AR, I, y MA del ARIMA, SARIMA introduce componentes estacionales adicionales: autoregresivo estacional (SAR), integrado estacional (SI), y promedio móvil estacional (SMA).

Estacionalidad: Está diseñado específicamente para capturar la estacionalidad en los datos. Puede modelar patrones que se repiten en intervalos regulares, como los datos mensuales, trimestrales o anuales típicos en economía, finanzas y otros campos.



PARAMETRO m

El parámetro **m** en SARIMA refiere a la "**frecuencia**" o el número de observaciones por ciclo en tu serie de tiempo.

Este parámetro es crucial para capturar la estacionalidad. Ayuda al modelo a entender cuán a menudo se repiten los patrones estacionales.

Para determinar m, debes mirar la frecuencia con la que tus datos muestran un patrón repetitivo.Ejemplo: Si estás analizando datos de ventas mensuales y notas que hay un patrón que se repite cada año (como un aumento en diciembre por las fiestas), tu ciclo es anual.

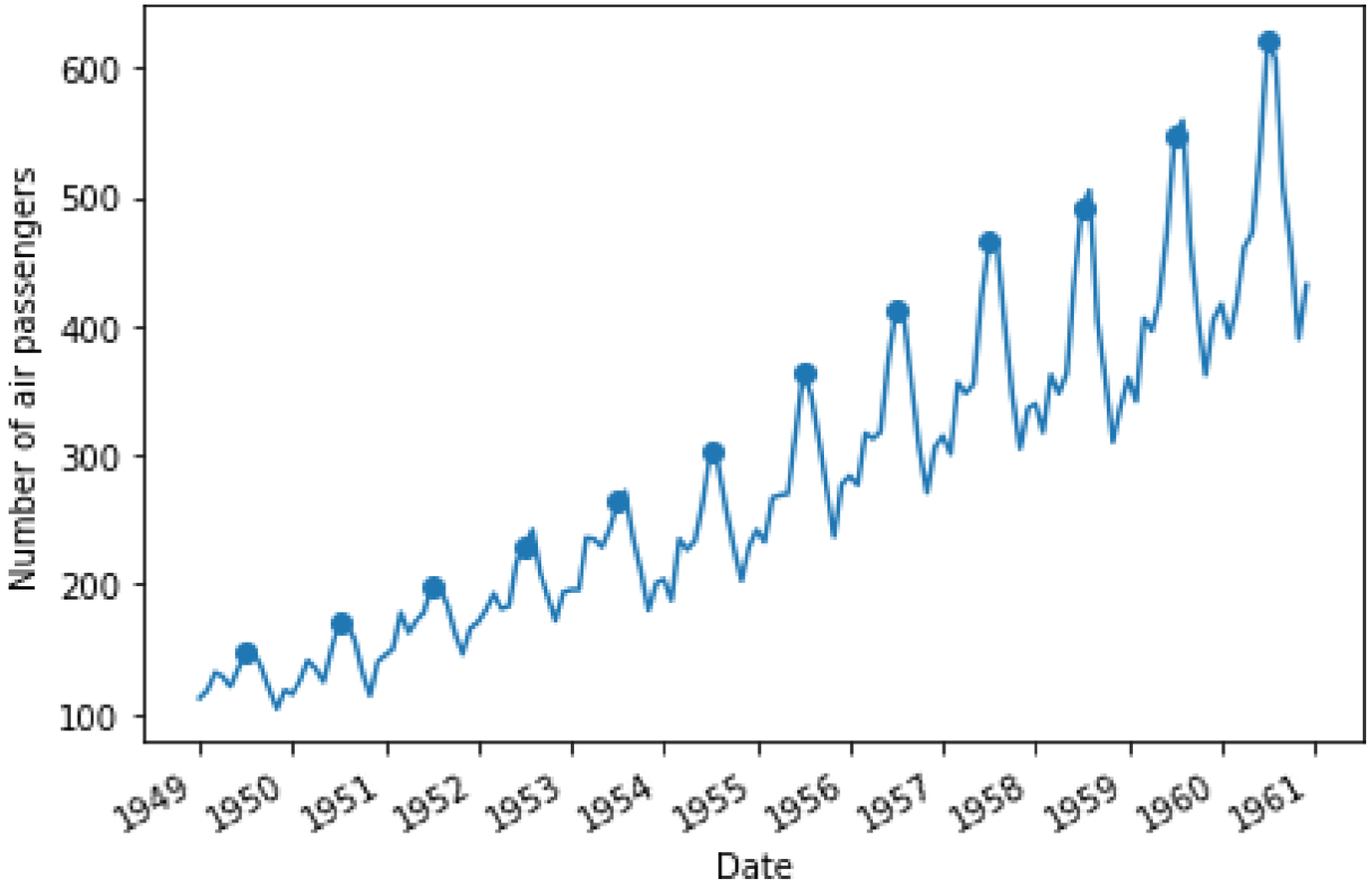
Recolección de Datos	Minuto	Hora	Día	Semana	Mes	Trimestre	Año
Cada Segundo	60	3600	86400	604800	2592000	7776000	31536000
Cada Minuto		60	1440	10080	43200	129600	525600
Horario			24	168	720	2160	8760
Diario				7	30	90	365
Mensual					1	3	12
Anial							

SARIMA (p,d,q)(P,D,Q)m

(SARIMA) añade parámetros estacionales al modelo:

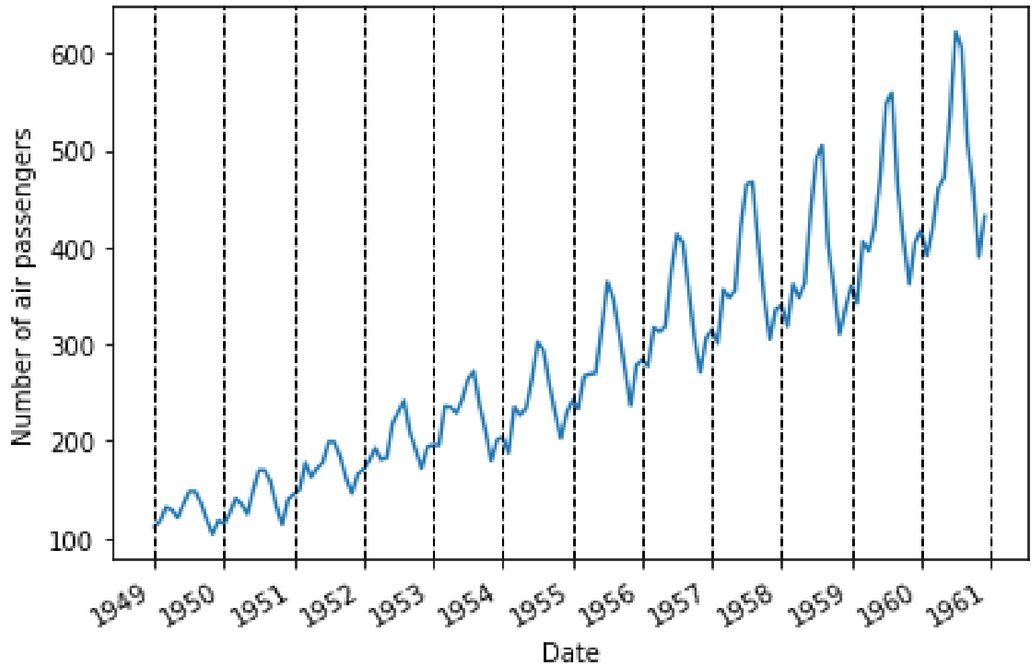
- P es el orden del proceso AR(P) estacional,
- D es el orden estacional de integración,
- Q es el orden del proceso MA(Q) estacional, y
- m es la frecuencia, o el número de observaciones por ciclo estacional.

Cabe señalar que un modelo SARIMA(p,d,q)(0,0,0)m es equivalente a un modelo ARIMA(p,d,q).

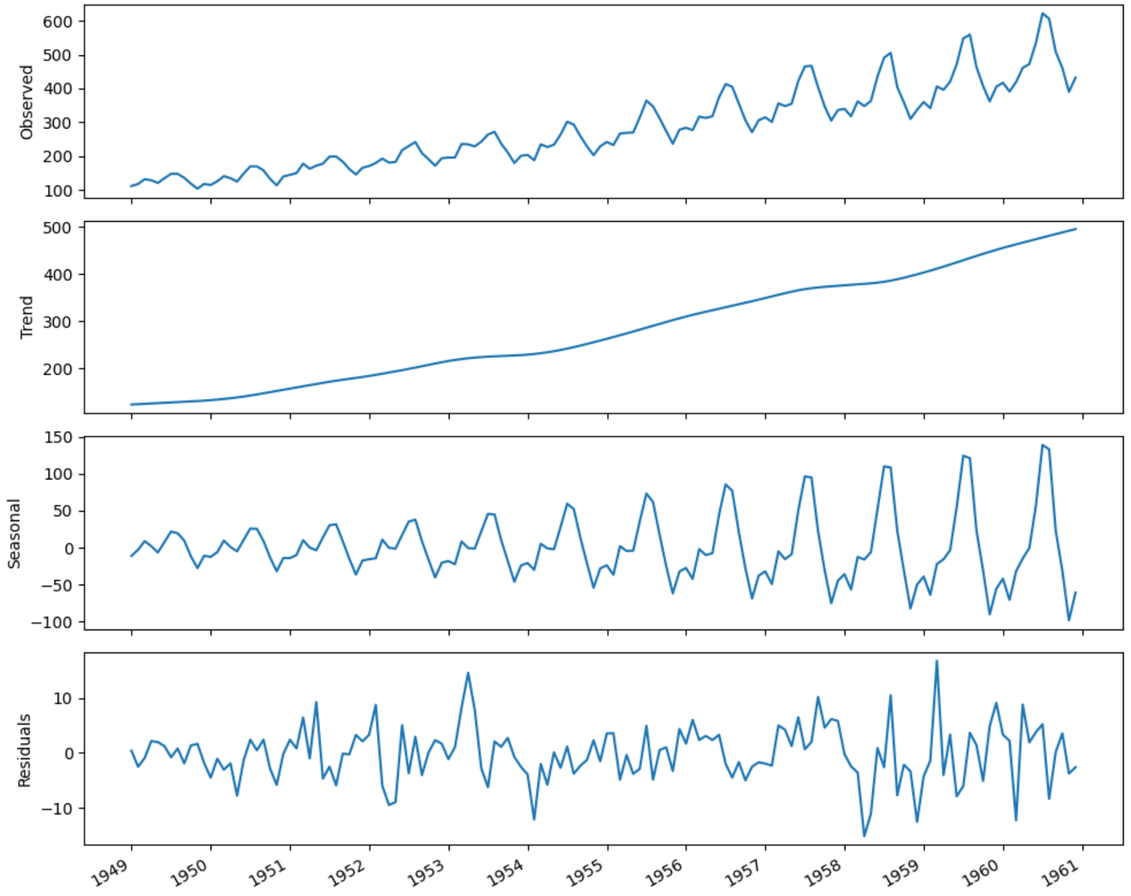


IDENTIFICACIÓN DE PATRONES ESTACIONALES

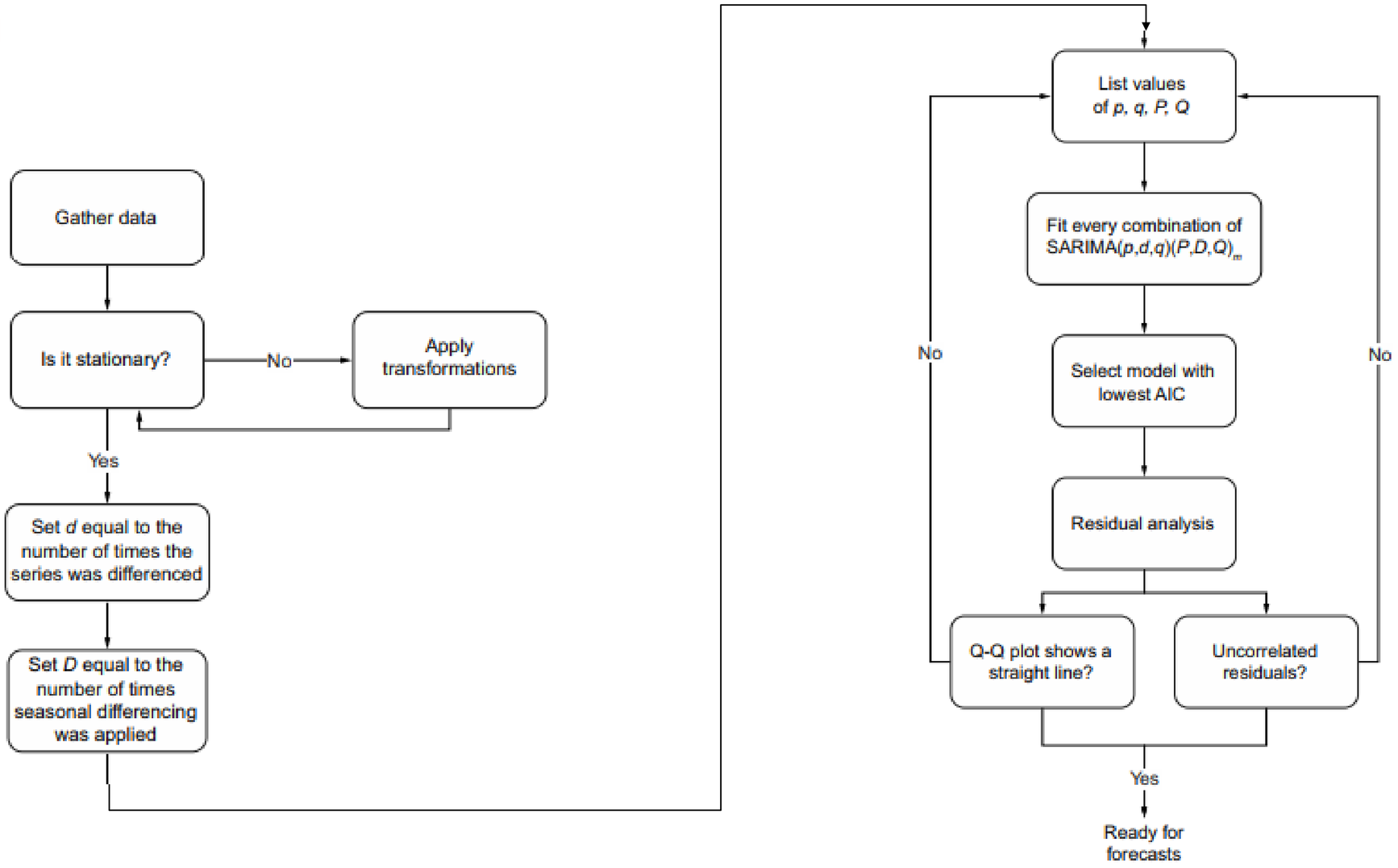
Graficar para identificar la estacionalidad



Descomposición de series temporales



resumen


Gather data
Is it stationary?
No
Apply transformations
Yes
Set d equal to the number of times the series was differenced
Set D equal to the number of times seasonal differencing was applied
List values of p, q, P, Q
Fit every combination of SARIMA(p,d,q)(P,D,Q)m
Select model with lowest AIC
Residual analysis
Q-Q plot shows a straight line?
Uncorrelated residuals?
No
No
Yes
Ready for forecasts

Ejercicio 1

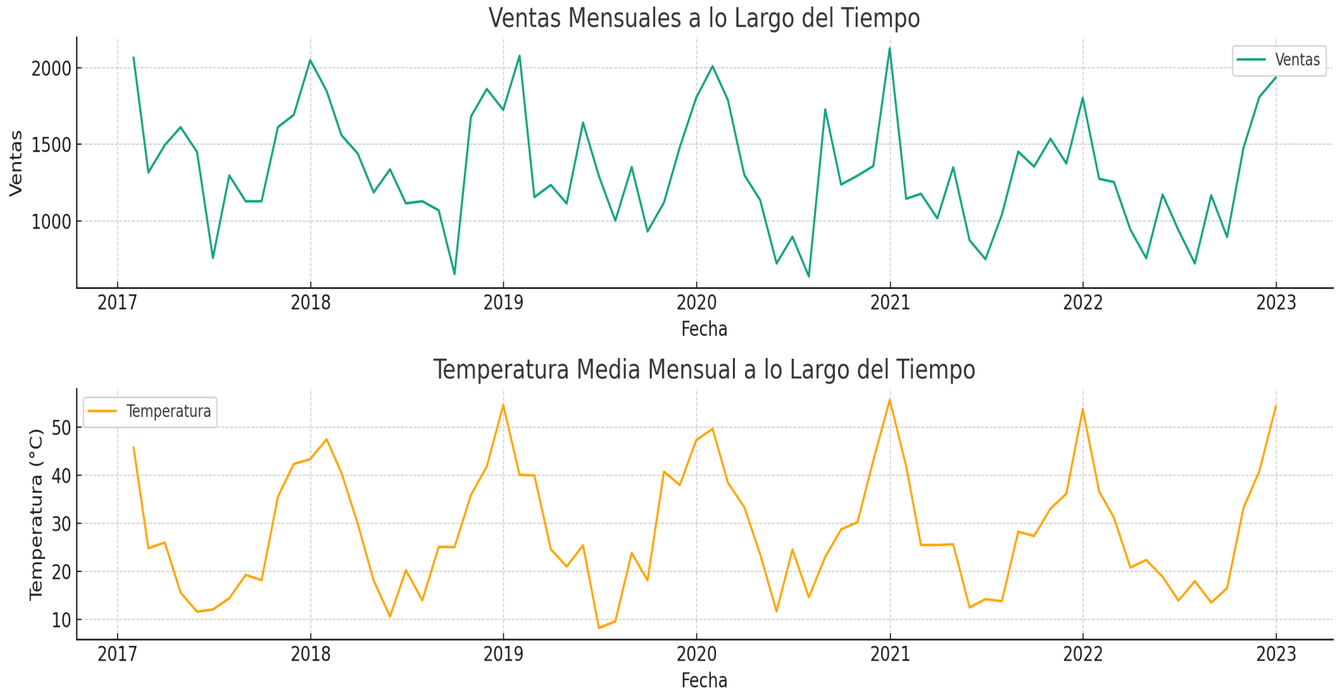
- Definir los parámetros del modelo SARIMA para le data set de pasajeros
- Haga la predicción del un año para el modelo ARIMA y SARIMA
- Compare los resultados con el modelo ARIMA
- Grafique la predicción

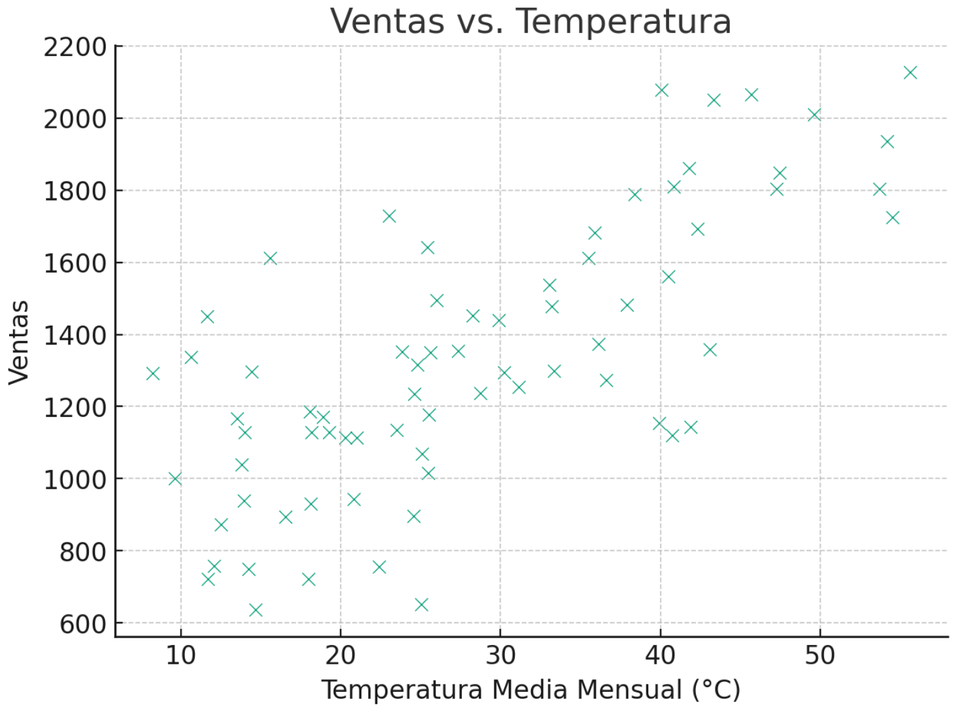
Ejercicio 2

- Aplicar el modelo SARIMA(p,d,q)(P,D,Q)m al conjunto de datos de Johnson & Johnson. para predecir el EPS trimestral durante un año.

- Usa la descomposición de series temporales para identificar la presencia de un patrón periódico.
- Usa la función optimize_SARIMA y selecciona el modelo con el AIC más bajo.
- Pronostica el EPS para el último año y mide el rendimiento en comparación con el modelo ARIMA

Hasta ahora, cada modelo que hemos explorado y utilizado para producir pronósticos ha considerado solo la serie de tiempo en sí misma. En otras palabras, los valores pasados de la serie de tiempo se usaban como predictores de valores futuros. Sin embargo, es posible que variables externas también tengan un impacto en nuestra serie de tiempo y, por lo tanto, puedan ser buenos predictores de valores futuros





SARIMAX

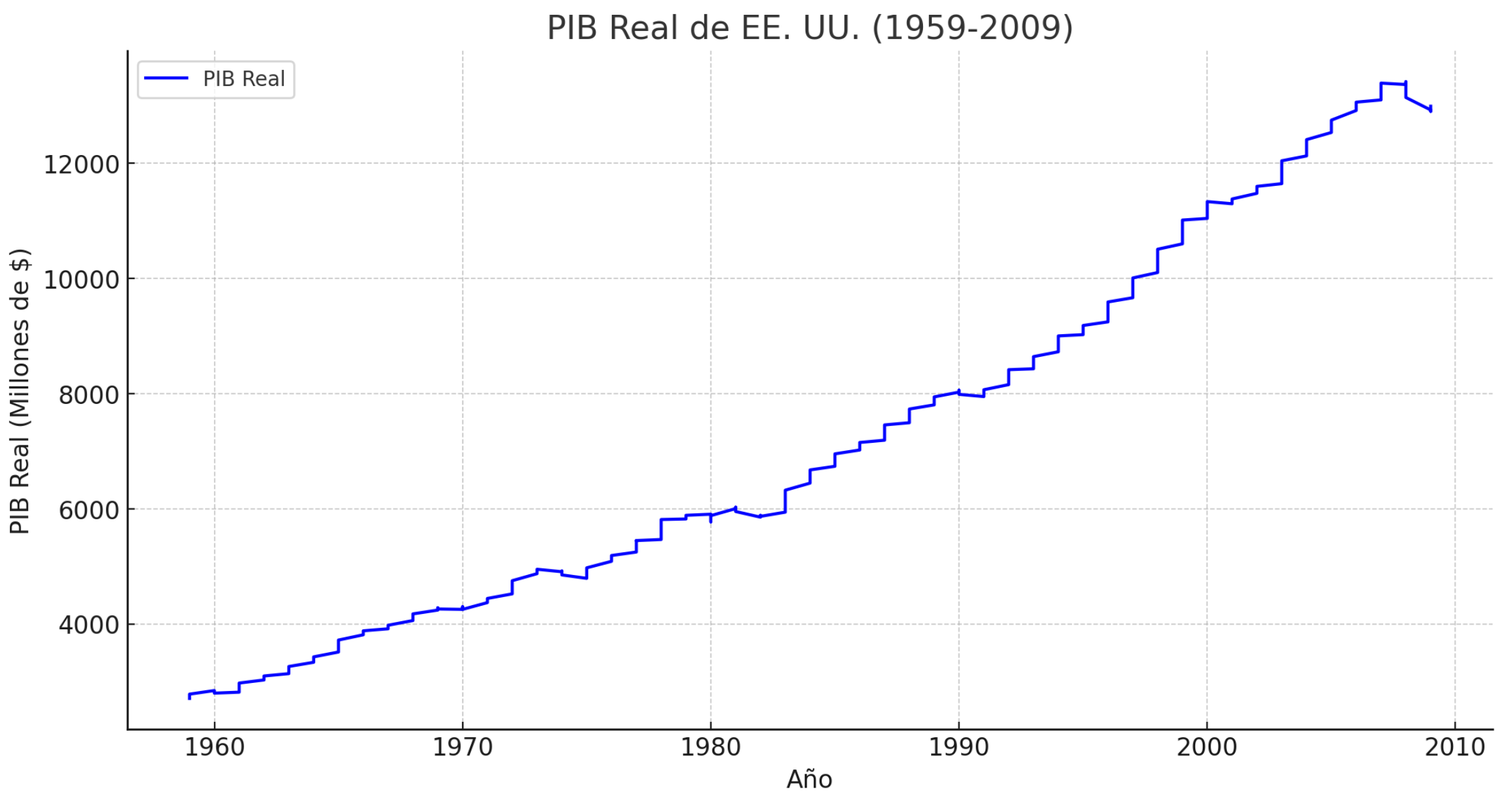
El modelo SARIMAX simplemente añade una combinación lineal de variables exógenas al modelo SARIMA. Esto nos permite modelar el impacto de variables externas sobre el valor futuro de una serie temporal

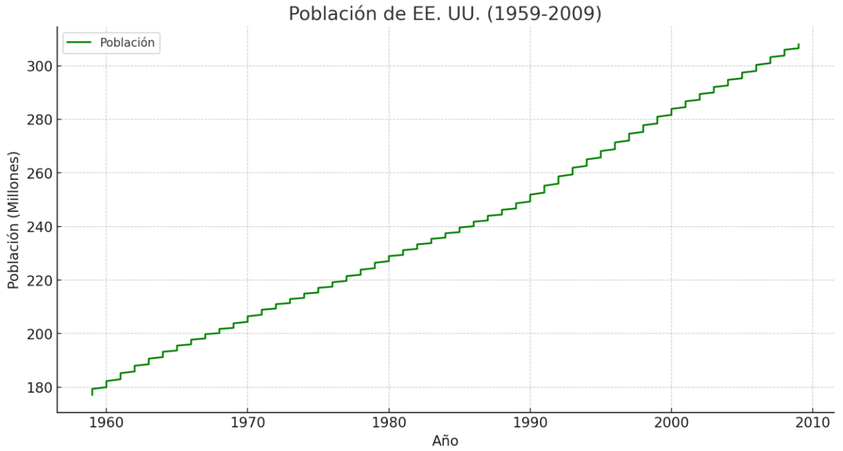
$$y_t = SARIMA(p,d,q)(P,D,Q)_m + \sum_{i=1}^{n} \beta_i X_t^i$$

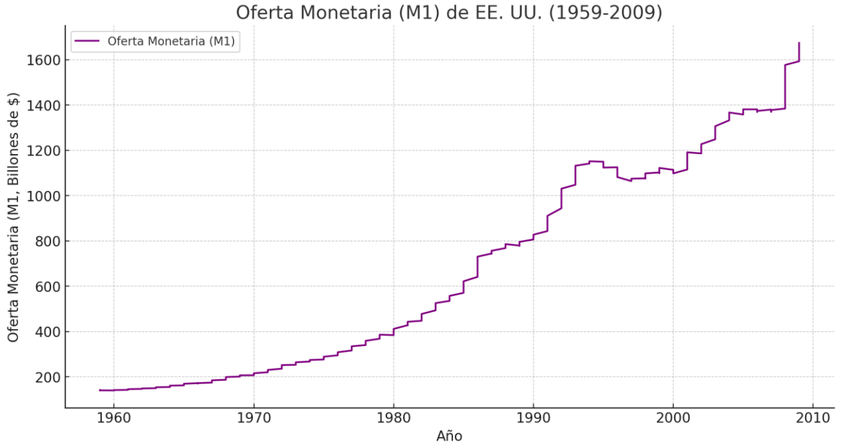
1. Sin patrones estacionales, se convierte en un modelo ARIMAX
2. Sin variables exógenas, se transforma en un modelo SARIMA
3. Sin estacionalidad ni variables exógenas, se simplifica a un modelo ARIMA.

¿qué pasa si deseas predecir dos pasos temporales en el futuro?

Qué debo hacer para pronosticar el 2012 del PIB







SARIMAX nos obliga a pronosticar también las variables exógenas.

Qué debo hacer para pronosticar el 2012 del PIB

Todas las variables exógenas tienen un valor p menor que 0.05, excepto para realdpi, que tiene un valor p de 0.896.

Esto significa que el coeficiente de **realdpi** no es significativamente diferente de 0, su coeficiente es 0.0034.

Sin embargo, el coeficiente se mantiene en el modelo, ya que el valor p no determina la relevancia de este predictor en el pronóstico de nuestro objetivo

SARIMAX Results

Dep. Variable:	realgdp	No. Observations:	200
Model:	SARIMAX(3, 1, 3)	Log Likelihood	-860.069
Date:	Thu, 23 Nov 2023	AIC	1748.138
Time:	14:20:19	BIC	1794.244
Sample:	01-01-1959	HQIC	1766.798
	- 10-01-2008		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
realcons	0.9508	0.052	18.202	0.000	0.848	1.053
realinv	1.0190	0.035	29.336	0.000	0.951	1.087
realgovt	0.7102	0.137	5.165	0.000	0.441	0.980
realdpi	0.0034	0.026	0.130	0.896	-0.048	0.054
cpi	4.9776	1.677	2.968	0.003	1.690	8.265

```python
def recursive_forecast_SARIMAX(endog: Union[pd.Series, list],
                               exog: Union[pd.Series, list],
                               train_len: int, horizon: int, window: int) -> li

    pred_SARIMAX = []
    total_len = train_len + horizon

    for i in range(train_len, total_len, window):
        model = SARIMAX(endog[:i], exog[:i], order=(3,1,3),
                        seasonal_order=(0,0,0,4), simple_differencing=False)
        res = model.fit(disp=False)
        predictions = res.get_prediction(exog=exog)
        oos_pred = predictions.predicted_mean.iloc[-window:]
        pred_SARIMAX.extend(oos_pred)

    return pred_SARIMAX
```

La **advertencia** de usar un modelo SARIMAX es que solo es razonable predecir el siguiente paso temporal, para evitar predecir también las variables exógenas, lo que nos llevaría a acumular errores de predicción en el pronóstico final.

En resumen

- Se usa la misma metodología de SARIMA
- El modelo SARIMAX te permite incluir variables externas, también denominadas variables exógenas, para pronosticar tu objetivo.
- Las transformaciones se aplican solo en la variable objetivo, no en las variables exógenas.
- Si deseas pronosticar varios pasos de tiempo en el futuro, también se deben pronosticar las variables exógenas. Esto puede magnificar los errores en el pronóstico final. Para evitarlo, debes predecir solo el siguiente paso de tiempo.

El modelo SARIMAX la variable exógena tenía un impacto en el objetivo, pero no al revés.

SARIMAX nos obliga a pronosticar también las variables exógenas.

Y Ahora …

Si necesitamos captura la relación entre múltiples series a medida que cambian con el tiempo. Es decir ,cada serie tiene un impacto en las otras, a diferencia del modelo SARIMAX

Recordemos

AR(p) expresaba el valor de una serie de tiempo como una combinación lineal de una constante C, el término de error presente ϵt, que también es ruido blanco, y los valores pasados de la serie yt –p.

$$y_t = \mathrm{C} + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

Generalizamos …..

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} \mathrm{C}_1 \\ \mathrm{C}_2 \end{bmatrix} + \begin{bmatrix} \phi_{1,1} & \phi_{1,2} \\ \phi_{2,1} & \phi_{2,2} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix}$$

El modelo de autoregresión vectorial VAR(p)

Modela la relación de dos o más series de tiempo. En este modelo, cada serie de tiempo tiene un impacto en las otras. Esto significa que los valores pasados de una serie de tiempo afectan a las otras series de tiempo, y viceversa.

El modelo VAR(p) se puede ver como una generalización del modelo AR(p) que permite múltiples series de tiempo. Al igual que en el modelo AR(p), el orden p del modelo VAR(p) determina cuántos valores rezagados impactan en el valor presente de una serie. Sin embargo, en este modelo, también incluimos valores rezagados de otras series de tiempo.

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} \mathrm{C}_1 \\ \mathrm{C}_2 \end{bmatrix} + \begin{bmatrix} \phi^1_{1,1} & \phi^1_{1,2} \\ \phi^1_{2,1} & \phi^1_{2,2} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \phi^2_{1,1} & \phi^2_{1,2} \\ \phi^2_{2,1} & \phi^2_{2,2} \end{bmatrix} \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + \cdots$$

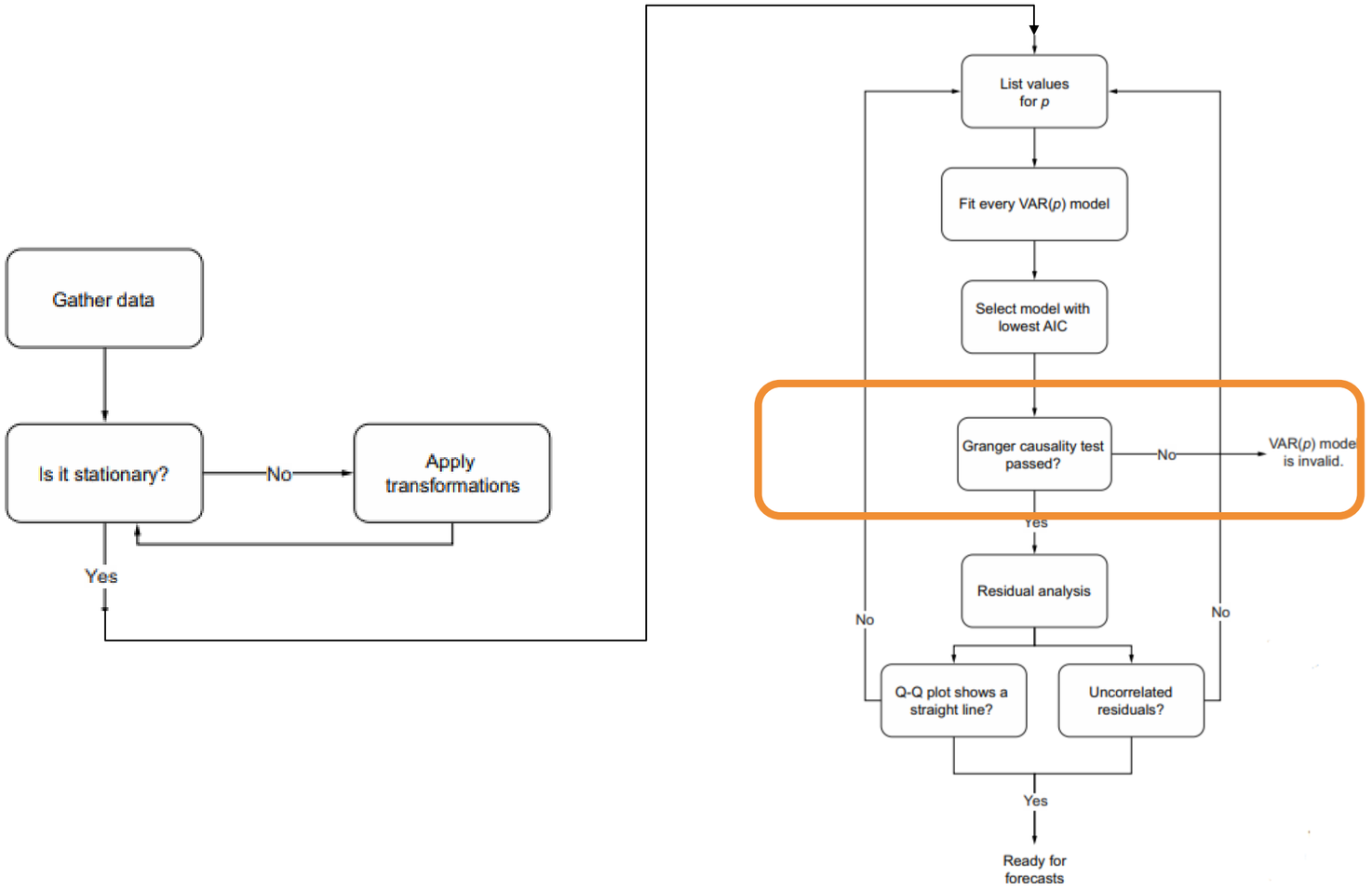
$$+ \begin{bmatrix} \phi^p_{1,1} & \phi^p_{1,2} \\ \phi^p_{2,1} & \phi^p_{2,2} \end{bmatrix} \begin{bmatrix} y_{1,t-p} \\ y_{2,t-p} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix}$$

El modelo de autoregresión vectorial VAR(p)

Una vez que se ha seleccionado el modelo con el AIC más bajo, realizamos la **prueba de causalidad de Granger**.

Esta prueba determina si los valores pasados de una serie de tiempo son estadísticamente significativos para pronosticar otra serie de tiempo.
Si la prueba falla, lo que significa que los valores p devueltos son mayores que 0.05, el modelo VAR(p) es inválido y no puede ser utilizado.

Es importante probar esta relación porque el modelo VAR(p) utiliza valores pasados de una serie de tiempo para pronosticar otra



Propuesta de Proyecto Final: Análisis Avanzado de Datos

Se espera que cada grupo elija una problemática específica, identifique y recolecte un conjunto de datos apropiado, y luego aplique una combinación de técnicas y modelos analíticos para abordar la cuestión planteada|

Objetivos del Proyecto:

- Identificar una problemática real y relevante que pueda ser analizada mediante técnicas de análisis de datos.
- Seleccionar y preparar un conjunto de datos adecuado para el análisis.
- Elegir y justificar la combinación de al menos dos metodologías o modelos analíticos distintos aprendidos durante el curso.
- Realizar un análisis exhaustivo de los datos y validar los modelos seleccionados.
- Presentar los hallazgos y conclusiones de manera clara y bien fundamentada.

Propuesta de Proyecto Final: Análisis Avanzado de Datos

Etapas del Proyecto:

1.Selección del Tema y del Dataset:

1. Definir una problemática de interés.
2. Buscar y preparar un conjunto de datos que se ajuste a la problemática elegida.

2.Metodología:

1. Seleccionar y justificar las metodologías y modelos a utilizar (por ejemplo, regresión, métodos de suavización, modelos generalizados lineales, análisis de datos dependientes, etc.).
2. Desarrollar un plan de análisis coherente con los objetivos del proyecto.

3.Análisis y Validación:

1. Aplicar las metodologías seleccionadas al conjunto de datos.
2. Realizar la validación y evaluación de los modelos empleados.

4.Presentación de Resultados:

1. Preparar un informe detallado que incluya la metodología, el análisis realizado, los resultados obtenidos y las conclusiones.
2. Realizar una presentación que resuma los aspectos más importantes del proyecto.