Universidad del Rosario

Analisis Avanzado de Datos

W14.Analisis de datos dependientes

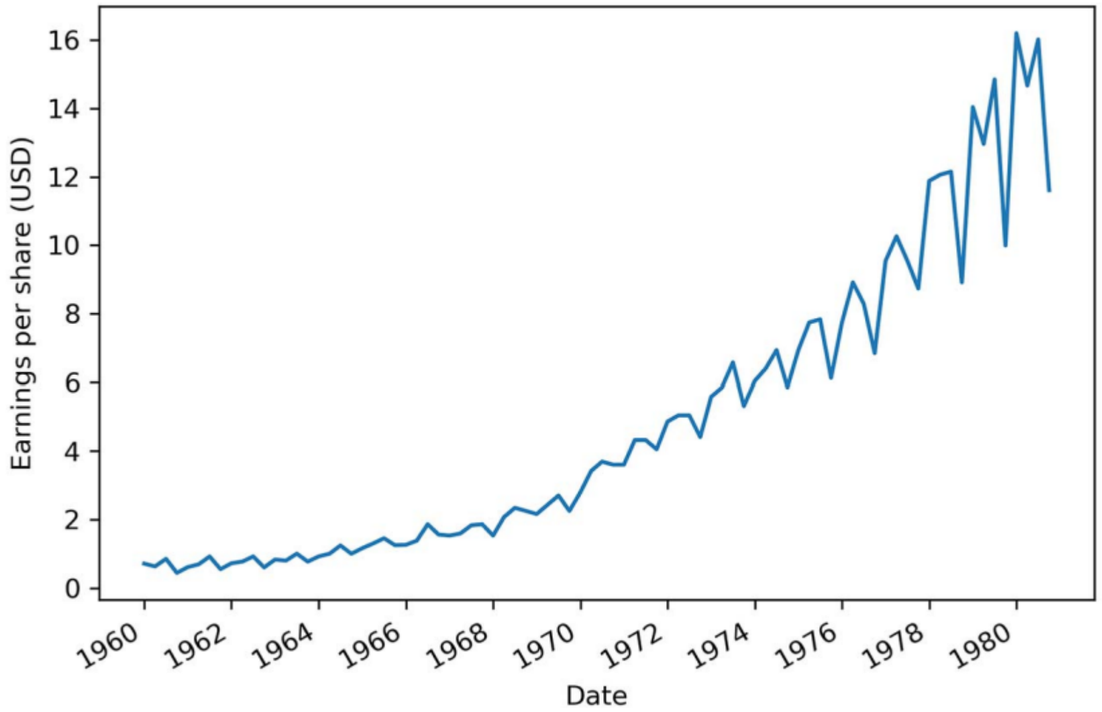
FERNEY ALBERTO BELTRAN MOLINA
Escuela de Ingeniería, Ciencia y Tecnología
Matemáticas Aplicadas y Ciencias de la Computación

Datos dependientes

- Los **datos dependientes** se refieren a datos en los que hay una relación o dependencia entre las observaciones.

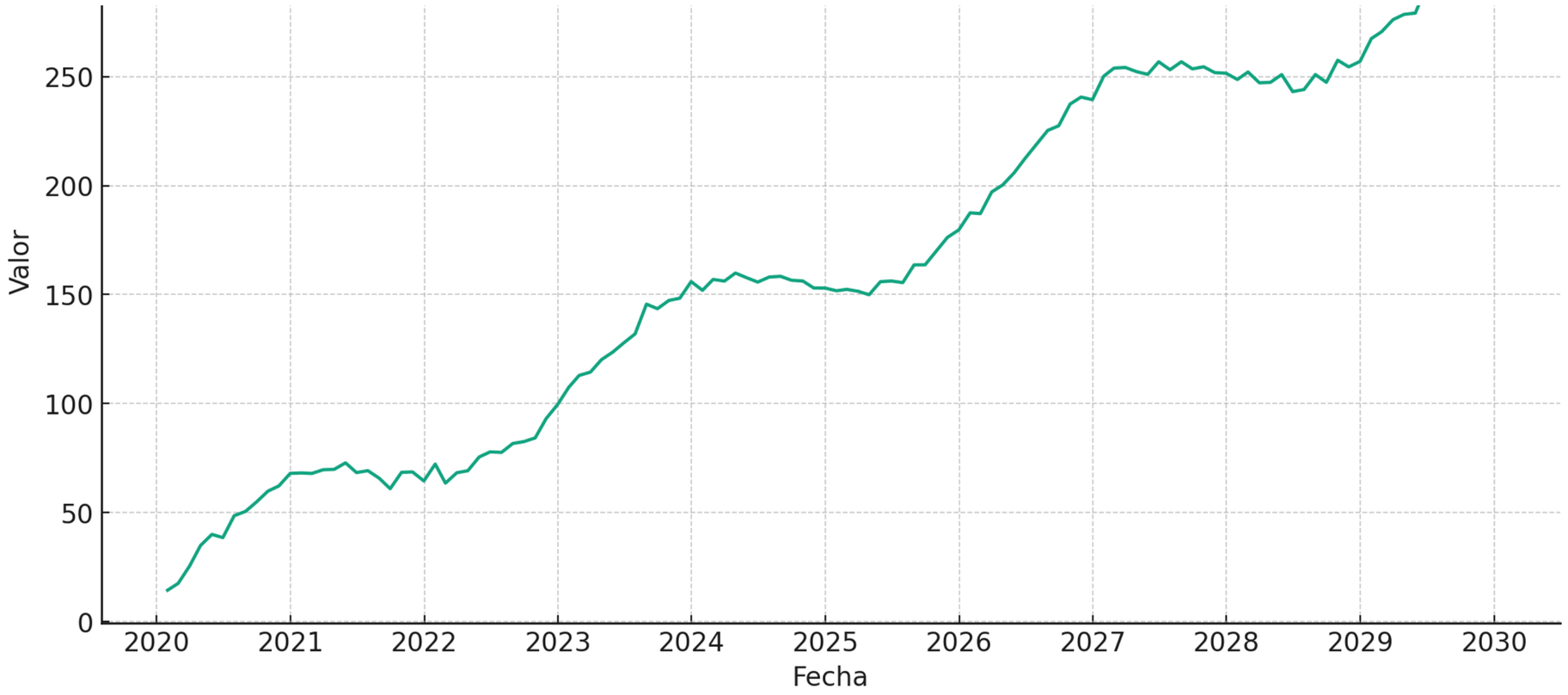
- Un ejemplo clásico son los precios de las **acciones a lo largo del tiempo**; el precio en un momento dado puede estar influenciado por los precios en momentos anteriores.

- Una serie temporal es un conjunto de puntos de datos ordenados en el tiempo. Los datos están espaciados de manera uniforme en el tiempo, lo que significa que se registraron cada hora, minuto, mes o trimestre



Descomposición series de tiempo

Los datos de series temporales pueden exhibir una variedad de patrones y, a menudo, resulta útil dividir una serie temporal en varios componentes, cada uno de los cuales representa una categoría de patrón subyacente. (ejm: **tendencia- ciclos** , **estacionalidad, residuos**).

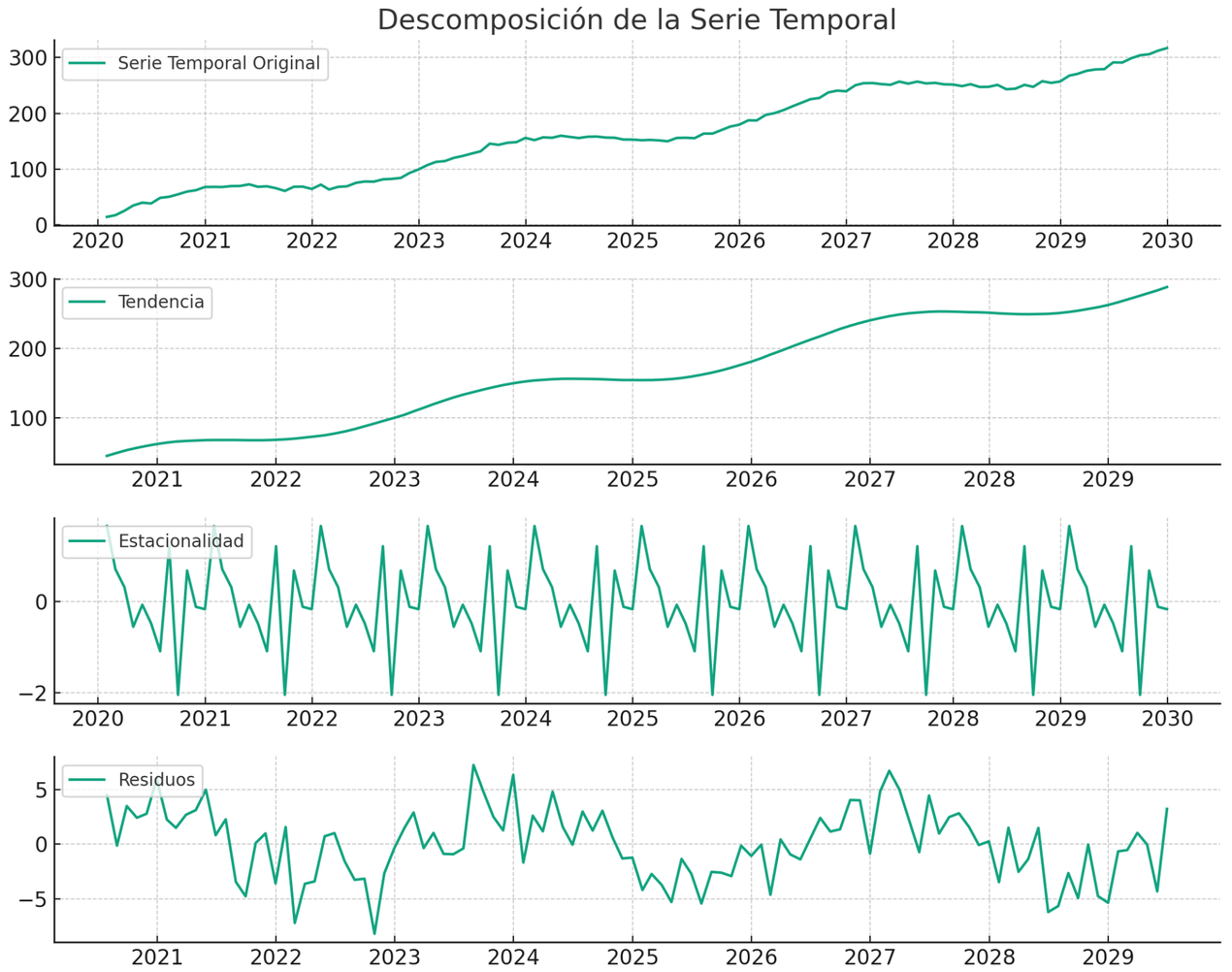


Descomposición series de tiempo

Tendencia: Cambios lentos en la serie temporal, ya sea en aumento o disminución.

Estacionalidad: Se visualizaría el patrón estacional, mostrando los ciclos que se repiten en intervalos regulares a lo largo de la serie temporal. el componente estacional muestra cómo nos desviamos de la tendencia

Residuos: comportamiento restante que no puede ser explicado por la tendencia y la estacionalidad. Este gráfico suele parecerse a ruido aleatorio.



Pronosticar series de tiempo

Buscamos predecir el futuro utilizando **datos históricos y conocimientos sobre eventos futuros** que podrían afectar nuestras previsiones.

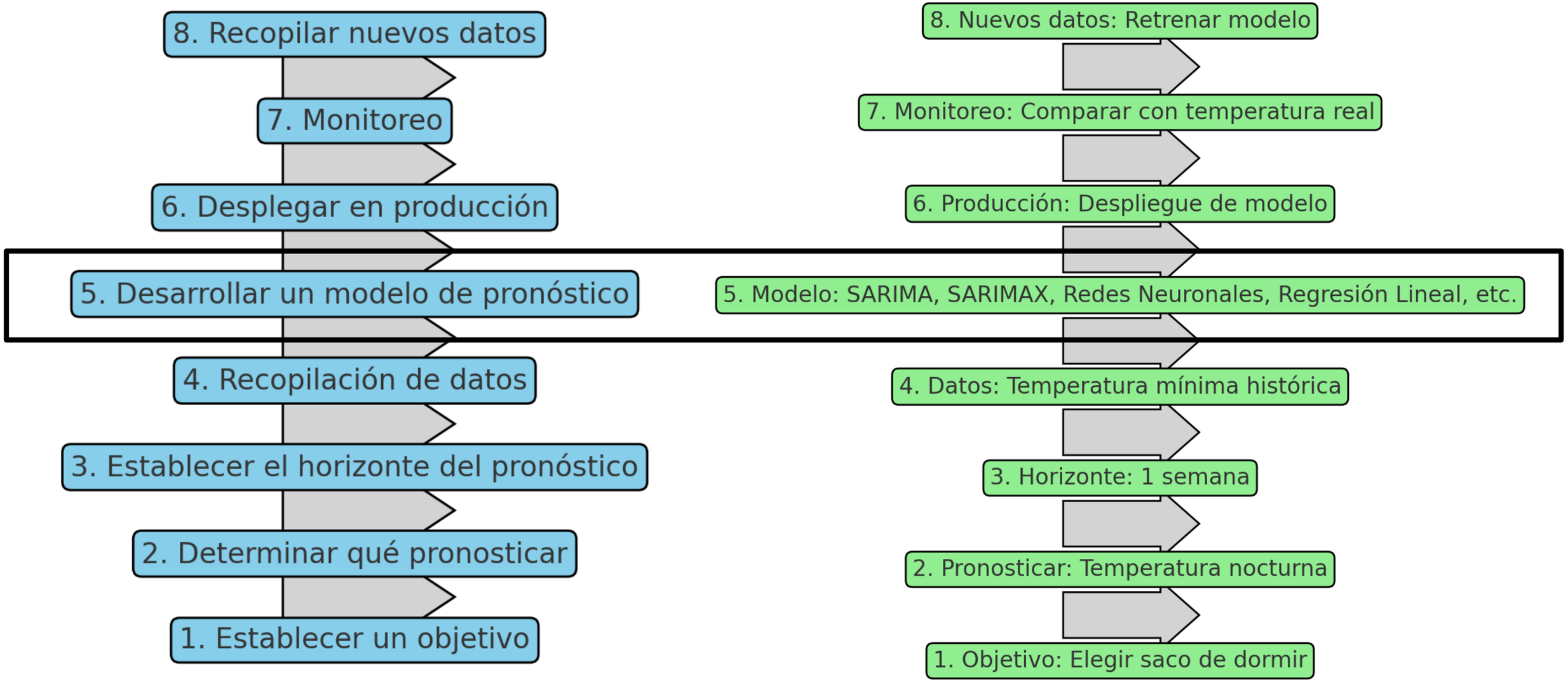
Por lo tanto el objetivo es **descubrir estructuras subyacentes** y **extraer patrones significativos de los datos históricos**. Sin embargo, el verdadero desafío radica en **separar** los patrones genuinos y relevantes para el pronóstico, de aquellos que son meramente coincidencias o ruido.

Un ejemplo clásico de confusión entre señal y ruido es la falacia de la '**mano caliente**' en el baloncesto, donde una serie de éxitos consecutivos puede ser erróneamente interpretada como una tendencia significativa, cuando en realidad podría ser simplemente una racha de suerte.

Al pronosticar series de tiempo, es esencial aplicar métodos rigurosos y análisis cuidadosos para evitar conclusiones erróneas y asegurar pronósticos confiables.

¡No dejes que la aleatoriedad te engañe!

Pasos para pronosticas series de tiempo



Modelos de pronóstico

Complejidad de los Datos	
Linealidad	Modelos Lineales (ARIMA, AR, MA, ARMA)
Heterocedasticidad	Modelos de Volatilidad (GARCH)
Dependencias Temporales	
Autocorrelación	Modelos ARIMA y sus extensiones (SARIMA)
Relaciones de Largo Plazo	Redes Neuronales (CNN,RNN, LSTM-Long Short-Term Memory)
Estacionariedad	
Estacionarios	Modelos ARIMA, AR, MA
No Estacionarios	Modelos SARIMA, Redes Neuronales
Frecuencia y Horizonte del Pronóstico	
Frecuencia Alta	Modelos de Aprendizaje Profundo (CNN, RNN, LSTM)
Horizonte Largo	Modelos SARIMA, LSTM
Disponibilidad de Datos Exógenos	
	Modelos SARIMAX y Modelos Supervizados: SVM
Interpretabilidad	
	Modelos Simples (ARIMA, AR, MA, Regresión Lineal)
Costo Computacional	
Bajo	Modelos Tradicionales (ARIMA, AR, MA)
Alto	Modelos de Aprendizaje Profundo (RNN, CNN, LSTM)
Robustez y Estabilidad	

Estacionariedad de las Series de Tiempo

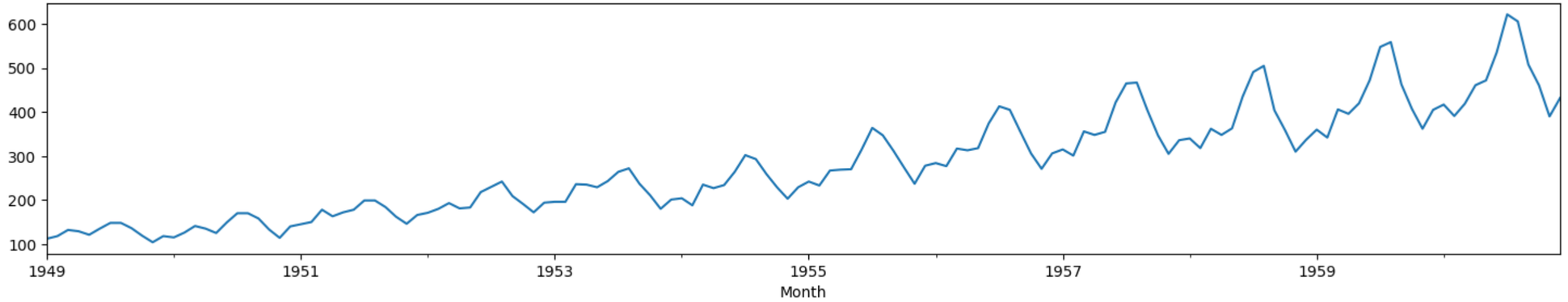
- Una serie de tiempo es estacionaria si sus propiedades estadísticas (como la media y la varianza) son constantes a lo largo del tiempo.
- La estacionariedad es importante porque muchos modelos de predicción asumen que la serie es estacionaria.
- Los modelos predictivos funcionan mejor con datos estacionarios.
- La no estacionariedad puede llevar a conclusiones engañosas y a modelos ineficaces.

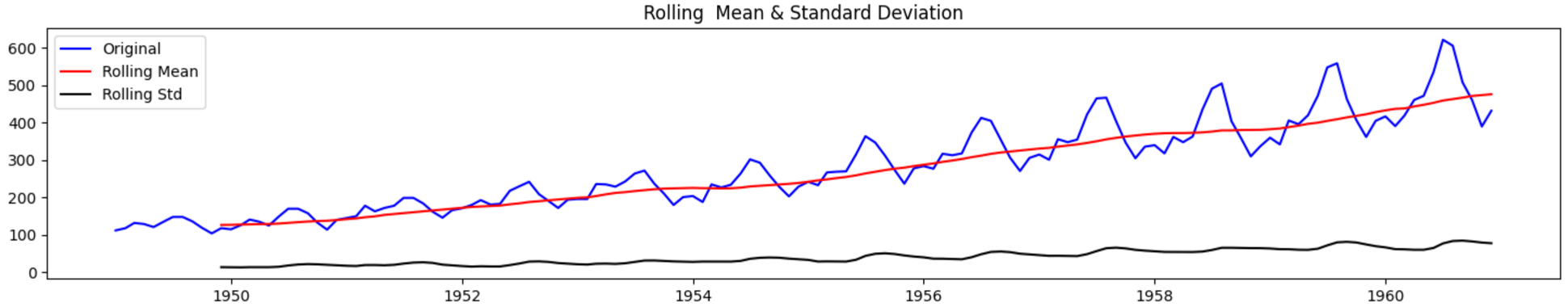
Verificar la Estacionariedad

1. Análisis Visual:
   1. Gráficos de líneas
   2. Gráficos de autocorrelación (ACF) y autocorrelación parcial (PACF)

2. Pruebas Estadísticas:
   1. Prueba de Dickey-Fuller Aumentada (ADF)
   2. Prueba KPSS (Kwiatkowski-Phillips-Schmidt-Shin)
   3. Prueba de Phillips-Perron (PP)

Estacionariedad de las Series de Tiempo





Que hacer si no hay Estacionariedad

Hay 2 razones principales detrás de la no estacionariedad de una Serie de Tiempo:

- **Tendencia**: media variable a lo largo del tiempo. Por ejemplo, en la diapositiva anterior vimos que, en promedio, el número de pasajeros estaba creciendo con el tiempo.

- **Estacionalidad**: variaciones en marcos de tiempo específicos. Por ejemplo, la gente podría tener la tendencia de comprar coches en un mes particular debido a aumentos de sueldo o festivales.

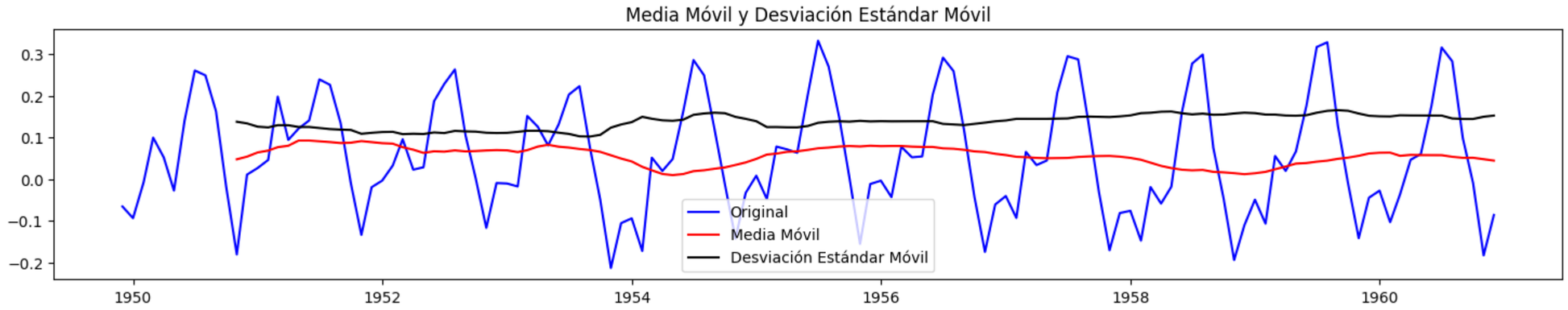
- El principio fundamental para abordar la no estacionariedad es modelar o estimar la tendencia y la estacionalidad presentes en la serie y luego eliminar estos componentes para obtener una serie estacionaria.
- Una vez hecho esto, se pueden aplicar técnicas de pronóstico estadístico a la serie.
- El paso final consistiría en convertir los valores pronosticados a la escala original, reintroduciendo las restricciones de tendencia y estacionalidad que habíamos eliminado.

Que hacer si no hay Estacionariedad

Diferenciación
Elimina tendencias y ciclos aplicando diferencias entre observaciones consecutivas.
Puede aplicarse más de una vez (diferenciación de orden superior).

Transformaciones
Estabiliza la varianza con transformaciones logarítmicas, raíz cuadrada, Box-Cox, etc.
Útil cuando los cambios en la serie son proporcionales a su nivel.

Descomposición Estacional
Separa y elimina los componentes de tendencia y estacionalidad.
Reconstruye la serie sin estas componentes para lograr estacionariedad.



Modelo de series de tiempo

Modelo AR (Auto-Regresivo)

- Se basa en los valores pasados de la propia serie.
- Parámetro ***'p'*** define el número de términos AR a considerar.
- Ejemplo: Si p=2, el modelo usa x(t−1) y x(t−2) para predecir x(t).

Modelo MA (Media Móvil)

- Se enfoca en los errores de pronóstico pasados.
- Parámetro ***'q'*** define el número de términos MA a utilizar.
- Ejemplo: Si q=2, el modelo usa e(t−1) y e(t−2) (errores pasados) para predecir x(t).

Modelo ARMA Combinación de AR y MA

- Adecuado para series de tiempo que ya son estacionarias.

Modelo ARIMA (Auto-Regresivo Integrado de Media Móvil)

- Combina AR y MA y agrega diferenciación para estacionarizar la serie.
- Parámetros: **'p'** (AR), **'d'** (diferenciación), **'q'** (MA).
- Ejemplo: Un modelo ARIMA(1,1,1) usa un término AR, una diferenciación y un término MA.
- Particularmente útil para series de tiempo no estacionarias que pueden ser estacionarizadas con diferenciación.
- **La selección de los parámetros 'p', 'd' y 'q' es crucial y se basa en el entendimiento de la autocorrelación y la autocorrelación parcial de la serie.**

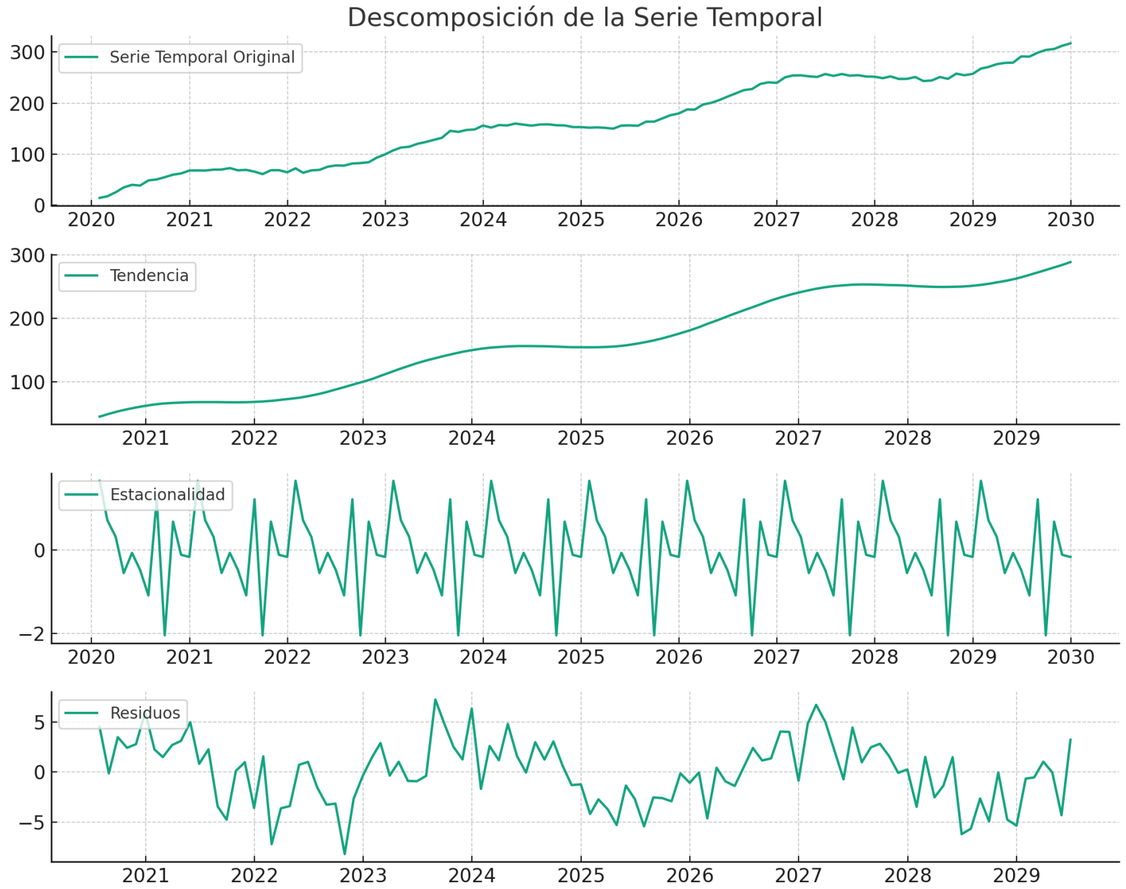
Matemática para la descomposición series de tiempo

Descomposición aditiva: Asume que la serie temporal es la suma de sus componentes.

$$Y_t = T_t + S_t + R_t$$

Descomposición multiplicativa: Asume que la serie temporal es el producto de sus componentes. Esto es útil cuando la variabilidad de la serie aumenta con el nivel de la serie (heteroscedasticidad).

$$Y_t = T_t \times S_t \times R_t$$



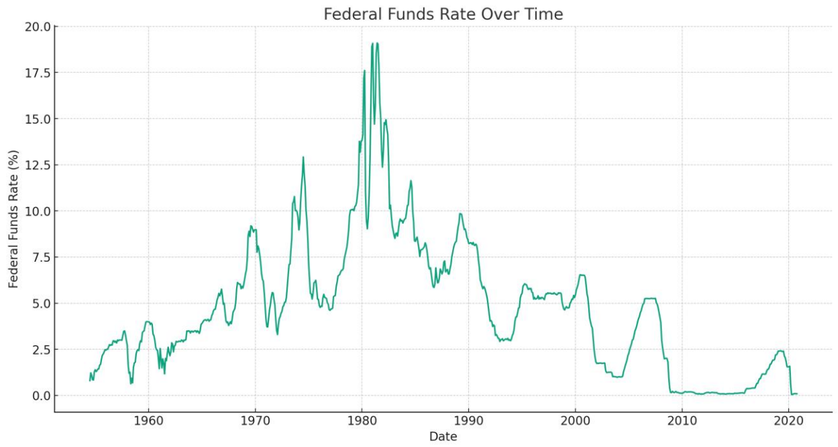
Estas descomposiciones se pueden realizar utilizando métodos estadísticos como **promedios móviles** para estimar la tendencia y la estacionalidad. Una vez estimados estos componentes, el residuo se obtiene simplemente restando (en el caso aditivo) o dividiendo (en el caso multiplicativo) estos componentes de la serie original. O usar

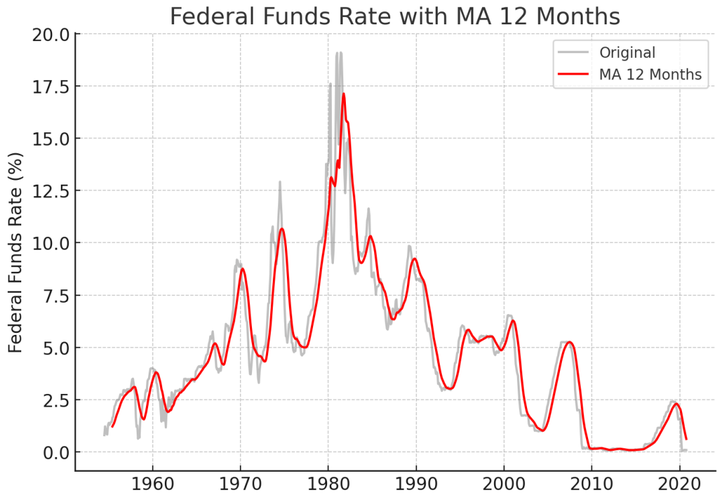
Media móvil

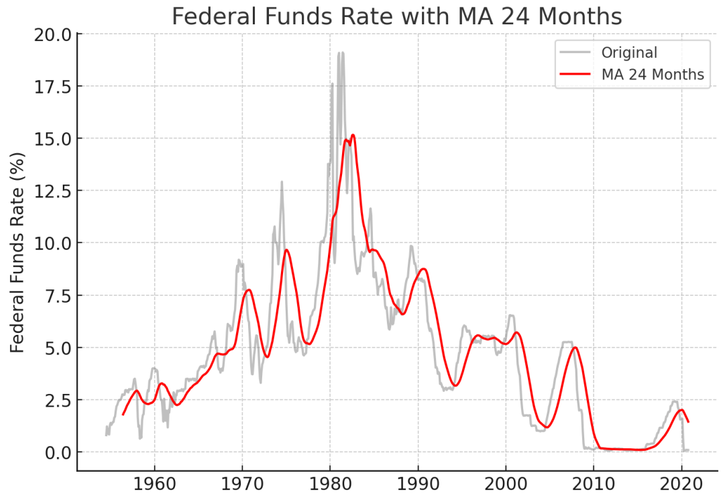
La técnica de promedio móvil es fundamental en la descomposición clásica de series temporales. Un promedio móvil de orden ***m*** se define como

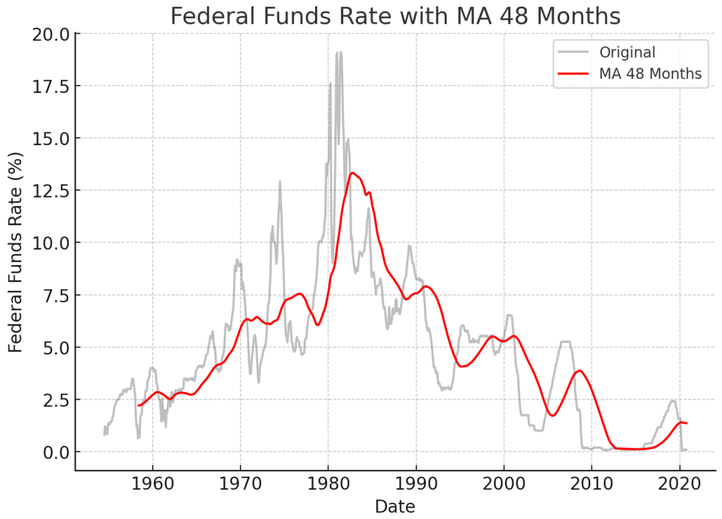
$$\hat{T}_t = \frac{1}{m} \sum_{j=-k}^{k} y_{t+j},$$

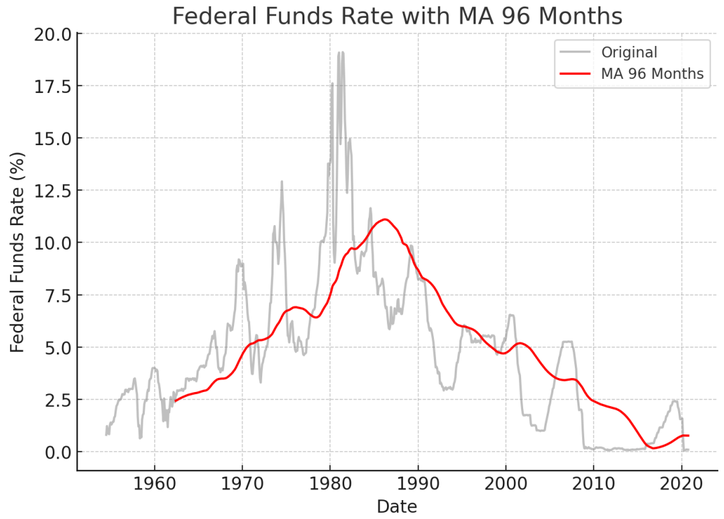
Donde **m = 2k+1**. Esta relación asegura que la ventana de promediado se centre alrededor del tiempo tt, abarcando kk periodos a ambos lados.



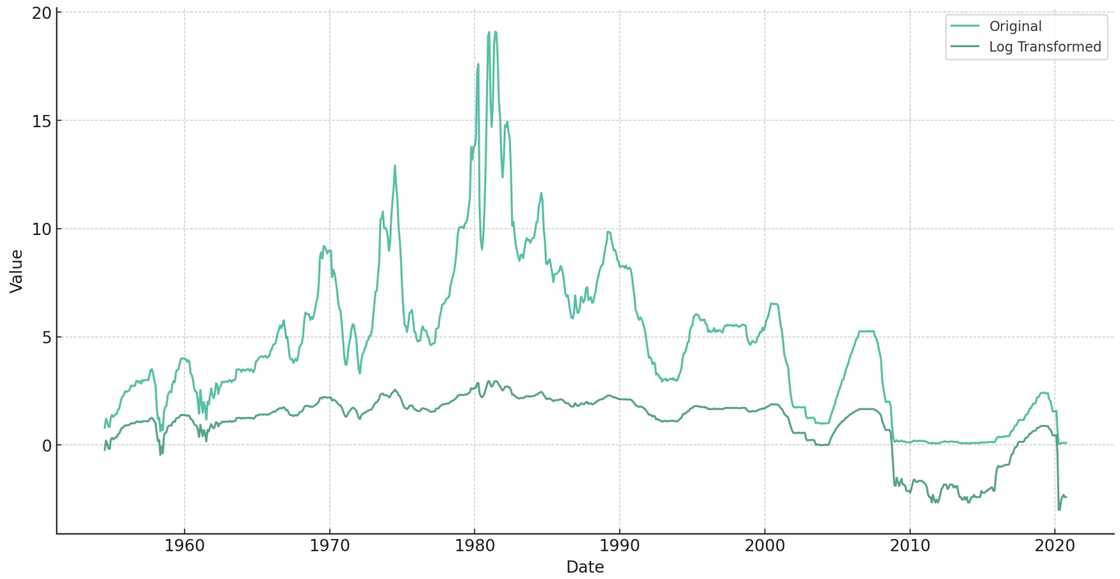


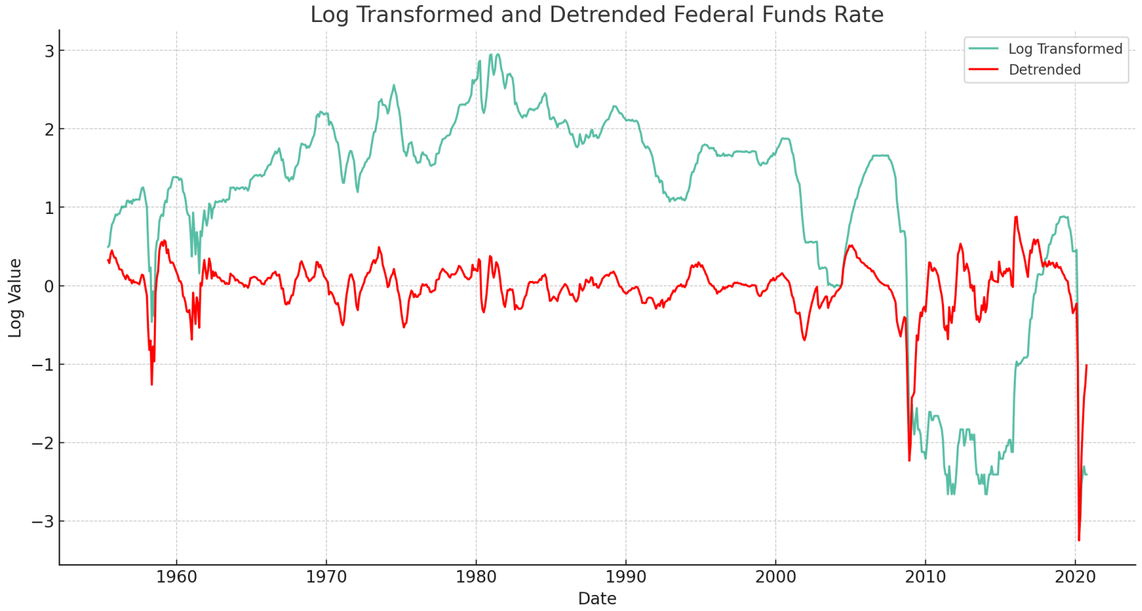






Transformación





Para los datos originales (sin transformar):
El valor-p fue mayor que 0.05, lo que indicó que **no podíamos rechazar** la hipótesis nula de que la serie tiene una raíz unitaria y, por lo tanto, es **no estacionaria**.

Para los datos después de aplicar la transformación logarítmica y detrended (restar la media móvil):
El valor-p fue extremadamente bajo (mucho menor que 0.05), lo que indicó que **podemos rechazar** la hipótesis nula de que la serie tiene una raíz unitaria y, por lo tanto, es **estacionaria**

Pasos descomposición series de tiempo

- **Calcula la tendencia:** a partir de la media móvil puede ser ejemplo 2MA del grado 12

- **Calcular la nueve serie detrended** $y_t - \hat{T}_t.$

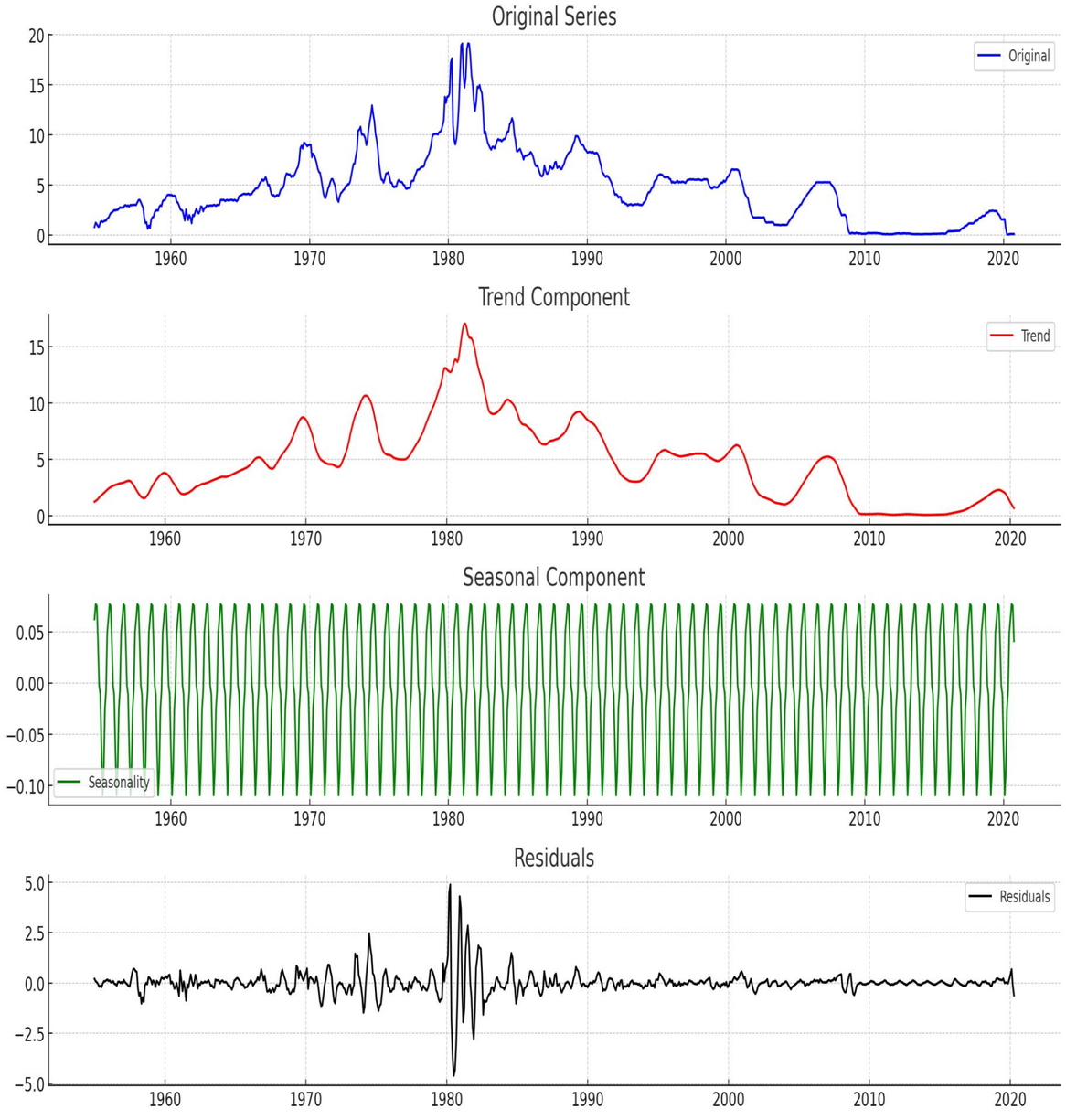
- **Estimar el componente estacional.**
  Calcular los promedios estacionales para cada período (por ejemplo, cada mes si tienes datos mensuales),
  Ajusta los promedios estacionales para que sumen cero.

- **Calcular el componente residual** restando los componentes estacional y de tendencia-ciclo estimados

$$\hat{R}_t = y_t - \hat{T}_t - \hat{S}_t.$$

Tambien puedes usa las siguientes funciones

1. **Seasonal_decompose** de Statsmodels
2. **Prophet** de Prophet desarrollado por facebook
3. **Deseasonalizer** de Sktime



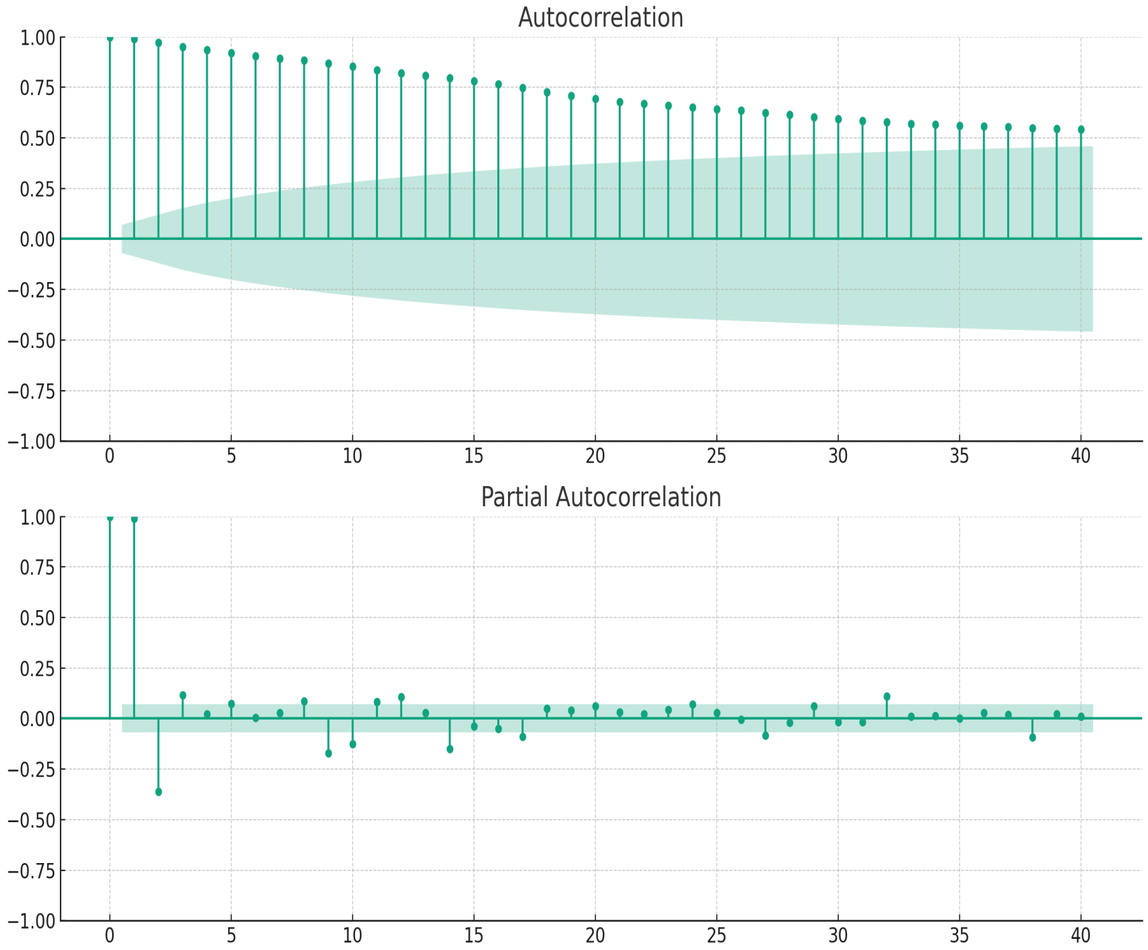
autocorrelación (ACF) y autocorrelación parcial (PACF)

Función de Autocorrelación (ACF): mide la correlación lineal entre los valores de la serie temporal separados por diferentes intervalos de tiempo (lags).

$$ACF(k) = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

Función de Autocorrelación Parcial (PACF): mide la correlación lineal entre los valores de la serie temporal con un retraso específico, k, eliminando la influencia de las correlaciones en retrasos más cortos.

$$PACF(k) = \frac{\mathrm{Cov}(y_t, y_{t-k}|y_{t-1}, \ldots, y_{t-k+1})}{\sqrt{\mathrm{Var}(y_t|y_{t-1}, \ldots, y_{t-k+1}) \times \mathrm{Var}(y_{t-k}|y_{t-1}, \ldots, y_{t-k+1})}}$$



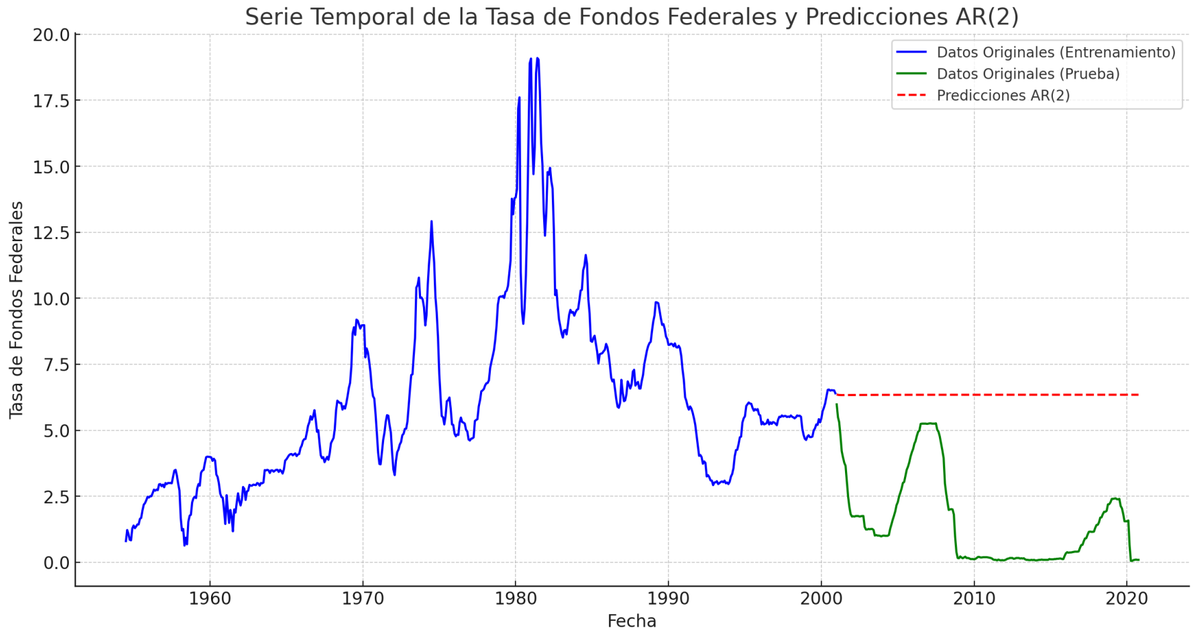
Recuerda: si los picos en la gráfica de ACF o PACF se encuentran dentro de llos intervalos de confianza para la correlación cero (sombras verdes), podríamos considerar que no hay una correlación significativa a ese retraso específico.

Modelo AR

La autorregresión es un modelo de series de tiempo que utiliza observaciones de pasos de tiempo anteriores como entrada a una ecuación de regresión para predecir el valor en el siguiente paso de tiempo. Es una idea muy simple que puede dar como resultado pronósticos precisos sobre una variedad de problemas de series temporales.

Un modelo de autorregresión es un modelo de regresión lineal que utiliza variables rezagadas como variables de entrada

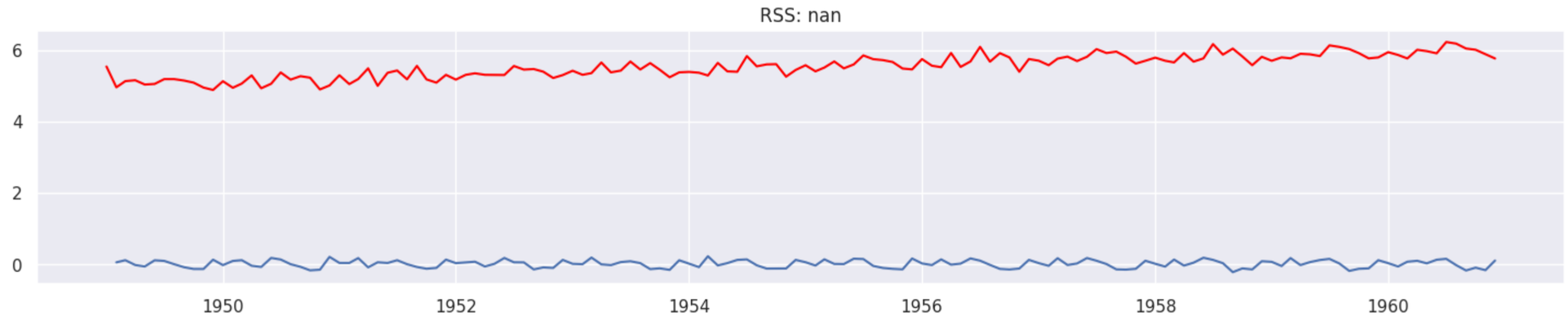
```
from statsmodels.tsa.ar_model import AutoReg
model = AutoReg(ts_log, lags=2)

results_AR = model.fit()
plt.plot(train)
plt.plot(results_AR.fittedvalues, color = 'red')
plt.title('RSS: %.4f'%
sum((results_AR.fittedvalues - ts_log_diff)**2))
#Residual sum of squares
```



Modelo MA

El modelo MA se utiliza para modelar la dependencia entre un error de pronóstico y los errores de pronóstico anteriores. En el modelo MA(q), el término q representa el orden de la media móvil, es decir, el número de términos del error de pronóstico pasado que se van a utilizar

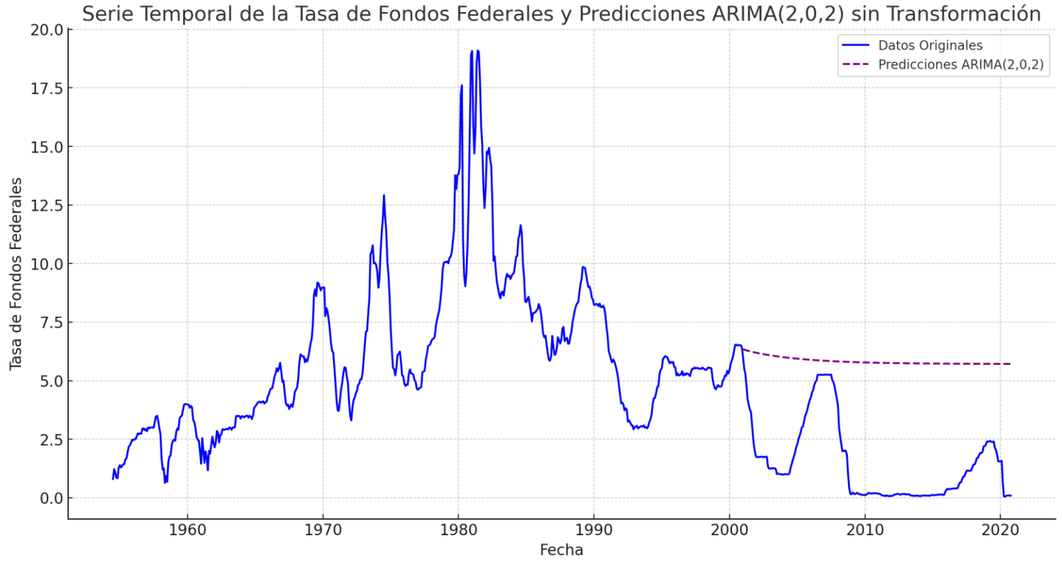
Para determinar el valor de q, puedes examinar el gráfico de autocorrelación (ACF) de la serie temporal. En un modelo MA(q), teóricamente esperarías que las autocorrelaciones sean significativamente diferentes de cero para los primeros q retardos y cerca de cero para los retardos posteriores. Esto puede darte una buena indicación inicial del orden del modelo MA a ajustar.



Modelo ARIMA
Autoregressive Integrated Moving Average

Es una extensión de los modelos AR y MA que incluye la **diferenciación** para hacer la serie temporal estacionaria. Se denota como ARIMA(p, d, q), donde:

p: Es el número de términos autoregresivos (AR).
d: Es el número de diferenciaciones necesarias para hacer la serie estacionaria. Si la serie ya es estacionaria, entonces d=0
q: Es el número de términos de media móvil (MA). Es similar al "q" en los modelos MA(q).



Diferenciación Simple: Consiste en restar el valor actual de la serie temporal del valor anterior. Si **y(t)** representa el valor de la serie temporal en el tiempo **t,** entonces la primera diferencia de la serie temporal se calcula como $\Delta y(t) = y(t) - y(t-1)$

Diferenciación de Orden Superior: A veces, una sola diferenciación no es suficiente para hacer la serie estacionaria, por lo que puedes aplicar diferenciaciones de orden superior

$$\Delta^2 y_t = \Delta y_t - \Delta y_{t-1}$$

Validación del modelo

La validación cruzada para series temporales es un poco diferente de la validación cruzada que se utiliza en otros tipos de modelos estadísticos o de aprendizaje automático. En las series temporales, no podemos dividir los datos de forma aleatoria porque las observaciones están ordenadas en el tiempo y la estructura temporal puede ser importante.