

Wavelets para series de tiempo

Alondra Matos¹, Nicole Romero¹ and Sebastián Ramírez¹

¹CIMAT, Sede Monterrey, Monterrey, Nuevo León, México.

Keywords: Transformer, Wavelets, Deep learning, Time Series, Forecasting

Resumen

En la actualidad, uno de los desafíos más significativos en la predicción de series de tiempo es obtener representaciones efectivas en tiempo-frecuencia que capturen tanto las dependencias temporales como las características inherentes a los datos secuenciales. Una alternativa prometedora en este contexto es la Transformada Wavelet. Este proyecto se centra en la comprensión de la Transformada Wavelet y su aplicabilidad en la predicción de series de tiempo. Exploraremos en detalle sus ventajas y desventajas en comparación con otras representaciones utilizadas en tareas de pronóstico de series temporales. Además, se enfocará en el uso de métodos de Machine/Deep Learning basados en esta transformación para abordar estos desafíos.

Índice

1. Introducción	3
2. Objetivos	3
3. Alcance del proyecto	4
4. Marco teórico	4
4.1. Transformada de Fourier	4
4.2. Transformada de Fourier de Tiempo Reducido (Short-Time Fourier Transform, STFT)	5
4.3. Transformada Wavelet	6
4.3.0.1. Wavelet de Haar	6
4.3.0.2. Wavelet de Morlet	6
4.3.1. Transformada Wavelet Continua (Continuous Wavelet Transform, CWT)	7
4.3.2. Transformada Wavelet Discreta (Discrete Wavelet Transform, DWT)	7
4.3.2.1. Transformada de wavelet discreta de máxima superposición	8
4.4. Comparación entre transformaciones	11
4.4.1. Transformada Fourier vs Wavelets	13
4.4.2. STFT vs Wavelets	17
5. Metodología	18
5.1. Descripción de los datos	18
5.2. Pruebas de estacionariedad, no linealidad y dependencia a largo plazo	24
5.3. W-Transformer	26
5.4. Métricas de evaluación	30
5.5. Implementación	31
6. Resultados	32
7. Conclusión	41
Bibliografía	42

1. Introducción

El análisis de series de tiempo se divide en dos aspectos: el análisis y la optimización. El análisis implica caracterizar las propiedades de las series de tiempo y se utilizan dos técnicas que implican la preparación y transformación de las series de tiempo, ya sea mediante expansión o extracción de características destacadas.

En el **dominio de tiempo**, se aplican transformaciones convencionales, como la eliminación de tendencia, estacionariedad o normalización. Además, se emplean métodos más complejos, como la transformación lineal o la expansión de SVD. Sin embargo, estas técnicas pueden tener dificultades al tratar con patrones periódicos, pseudoperiódicos o caóticos en las series de tiempo.

En el **dominio de frecuencias**, se utilizan técnicas comunes como el análisis espectral y diversas transformaciones. Estas técnicas descomponen la serie de tiempo en una combinación de senos y cosenos de diferentes frecuencias y amplitudes. La mayor parte de este proceso se realiza en el espacio de frecuencias, que se obtiene mediante la transformada de Fourier.

La **transformada de Fourier** es eficaz para señales estacionarias, pero puede perder información importante en señales no estacionarias. Para abordar este problema, es necesario procesar tanto el dominio del tiempo como el de la frecuencia simultáneamente. Esto se logra mediante modificaciones a la transformada de Fourier o el uso de las wavelets.

De igual manera, desde la perspectiva de aprendizaje profundo, para lidiar con series de tiempo se tienen:

- **Modelos estadísticos:** estos tienen una interpretabilidad fácil, pero fallan cuando se tiene una situación de alta complejidad.
- **Métodos de ML:** Puede aprovechar los datos de referencia para aprender las tendencias y patrones de manera automatizada, aunque a veces falla cuando existe una dependencia a largo plazo en los datos de series temporales observados.
- **DL actualmente:** Se tienen modelos como la red temporal convolucional o las recurrentes los cuales pueden lidiar con la no linealidad y no estacionariedad de la serie de tiempo. Sin embargo, estos pueden ser difíciles de entrenar por el problema del *exploding and vanishing gradient*.

El problema persiste a pesar de numerosas versiones de redes neuronales como LSTM o GRU, que suelen ser mucho mejor que los modelos clásicos, ya que estas versiones tienen problemas para capturar dependencias a largo plazo.

Con el paso del tiempo se han introducido los transformers a diferentes aplicaciones. En el caso de las series de tiempo se ha visto que tienen mucho éxito extrapolando y modelando dependencias de largo plazo e interacciones en datos temporales.

El mecanismo de atención con múltiples cabezas de los Transformers los hace particularmente adecuados para el análisis de datos de series temporales. También, comparados con las RNN y otras estructuras neuronales tienen como ventaja el hecho de que pueden permitir el procesamiento paralelo y reducir el tiempo de cómputo.

2. Objetivos

Corto plazo:

- Comprender los fundamentos teóricos de la transformada Wavelet y su aplicabilidad en la representación de series de tiempo.
- Identificar las ventajas y desventajas de la transformada wavelet en contraste con otra representación de datos, como la transformada de Fourier de tiempo reducido.

Largo plazo:

- Aplicar el algoritmo de W-transformer sobre diversas bases de datos.
- Comparar el desempeño de los modelos de Machine/Deep Learning considerando los filtros wavelet de Haar y otras representaciones en tiempo-frecuencia.

3. Alcance del proyecto

El alcance del proyecto incluirá el uso de representaciones de wavelets, específicamente las wavelets de Haar y Morlet. Además, se emplearán dos enfoques de deep learning: RNN (Redes Neuronales Recurrentes) y el modelo W-Transformer. Estos métodos se utilizarán para abordar las tareas de pronóstico de series de tiempo. Los resultados obtenidos se compararán y evaluarán en función de métricas relevantes para determinar su eficacia en la modelización y predicción de series de tiempo.

El proyecto también incluirá una comparación con transformaciones de Fourier de tiempo reducido (STFT) para evaluar cómo se desempeñan las representaciones de wavelets (Haar y Morlet) en comparación con el enfoque de STFT en tareas de pronóstico de series de tiempo. Esta comparación ayudará a determinar cuál de estas técnicas es más efectiva para capturar las características de las series de tiempo y modelar sus patrones.

En el proyecto se evaluará el ajuste y el rendimiento de las representaciones de Haar y Morlet, así como los métodos de RNN y W-Transformer, utilizando las métricas MAPAE, MASE, RMSE, MAE y SMAPE para determinar cuál de estos enfoques es más eficaz en la tarea de pronóstico de series de tiempo. La comparación se realizará exclusivamente mediante las métricas mencionadas.

4. Marco teórico

4.1. Transformada de Fourier

Mediante la transformación de Fourier, se puede descomponer cualquier función en una suma de senos y cosenos con diferentes frecuencias, los cuales proveen una base ortogonal para un espacio de funciones.

La representación de una serie de tiempo en el dominio de frecuencias queda determinada por:

$$\text{Transformada de Fourier: } X(\nu) = \int_{-\infty}^{+\infty} x(t)e^{-j2\pi\nu t} dt \quad (4.1)$$

donde

$x(t)$ es la serie de tiempo

$X(\nu)$ es la representación de la serie de tiempo en el dominio de frecuencia

Al aplicar la transformada de Fourier, la representación de $x(t)$ en el dominio del tiempo (t) se cambia a una representación $X(\nu)$ en el dominio de frecuencias, permitiendo conocer la contribución

relativa de cada frecuencia (ν). Sin embargo, dado que la transformada de Fourier solo toma la frecuencia como argumento, no es posible determinar el inicio o el final de ciertas frecuencias, por lo que se pierde la dinámica temporal.

Por otra parte, la transformada inversa de Fourier posibilita la recuperación de la función original a partir del dominio de frecuencia.

En particular, el producto del ancho de banda por tiempo (*Time-Bandwidth product*, $T \times B$) es un aspecto cualitativo relevante en las series temporales. Según la desigualdad de Heisenber-Gabor, se cumple que:

$$T \times B \geq 1 \quad (4.2)$$

La desigualdad 4.2 implica que es fundamentalmente imposible lograr una resolución perfecta tanto en tiempo como en frecuencia de manera simultánea. Siempre hay un intercambio de información entre ambos dominios.

En particular, la transformada de Fourier no es adecuada para señales no estacionarias principalmente porque se fundamenta en la descomposición en ondas infinitas que no están localizadas en el tiempo. Para integrar la multiespectro o multirresolución y satisfacer la necesidad de localización en las frecuencias, se recurre a una descomposición atómica de naturaleza bidimensional.

La bidimensionalidad implica que, para un momento específico en el tiempo, se tendrá acceso al espectro completo o a una parte de él, lo que permite conocer la contribución de la frecuencia ν al tiempo t . El término *atómico* se refiere a que la señal o serie temporal se dividirá en intervalos de tiempo elementales para evaluar el espectro.

4.2. Transformada de Fourier de Tiempo Reducido (Short-Time Fourier Transform, STFT)

Una transformación que se fundamenta en la descomposición atómica es la Transformada de Fourier de Tiempo Reducido, también llamada transformada de Gabor o Transformada de Fourier con Ventana (Windowed Fourier Transform). Esta transformación posibilita la incorporación de la dependencia temporal al establecer una ventana previa en la señal en torno a un punto específico en el tiempo, para luego evaluar la transformada de Fourier en esa ventana.

$$F_x(t, \nu, h) = \int_{-\infty}^{+\infty} x(u)h^*(u-t)e^{-j2\pi\nu u} du \quad (4.3)$$

donde

$x(u)$ es la serie de tiempo

$h(t)$ es la ventana de análisis a corto plazo.

$F_x(t, \nu, h)$ es la transformada de Fourier de tiempo reducido

Las componentes multifrecuencia se derivan al limitar la señal de entrada, ya sea con límites precisos o difusos, en un intervalo que es constante para todas las frecuencias. Como resultado, las frecuencias vinculadas a períodos más cortos que el intervalo de restricción se estiman con mayor precisión que las relacionadas con períodos más largos.

En general, una buena resolución en frecuencia requiere una función de ventana grande, mientras que una buena resolución en tiempo se obtiene con funciones de ventana más pequeñas.

4.3. Transformada Wavelet

Una limitación del enfoque STFT es la concesión de la resolución de tiempo-frecuencia, ya que los átomos se generan únicamente por traslación de la función elemental, lo que resulta en funciones atómicas con el mismo ancho para todas las frecuencias. En cambio, en la transformada Wavelet, los átomos se obtienen mediante traslación como dilatación (escalamiento) de la función elemental.

En este caso, se emplea el término *wavelets* para referirse a los átomos (escalados y trasladados), mientras que se le llama *función madre* a la función elemental asociada a la función de ventana. A partir de la wavelet madre, se genera una familia de versiones escaladas y trasladadas de la función.

En esencia, una wavelet es una onda pequeña, la cual crece y decae en un período de tiempo limitado. En consecuencia, una wavelet pertenece a una familia de funciones real valuadas $\psi(\cdot)$ que satisfacen las siguientes condiciones:

- Media cero

$$\int_{-\infty}^{\infty} \psi(t) dt = 0$$

- El cuadrado de ψ integra la unidad

$$\int_{-\infty}^{\infty} \psi(t)^2 dt < 1$$

La segunda condición implica que $\psi(t)$ debe realizar algunas oscilaciones alejándose de cero, mientras que la primera ecuación señala que las oscilaciones positivas se compensan con las oscilaciones negativas, lo que significa que las funciones wavelet deben mostrar un comportamiento oscilatorio.

4.3.0.1. Wavelet de Haar

La función wavelet madre de las funciones de Haar $\psi(t)$ puede ser descrita como

$$\psi(t) = \begin{cases} 1 & 0 \leq t < 1/2 \\ -1 & 1/2 \leq t < 1 \\ 0 & \text{De otra forma} \end{cases}$$

La wavelet de Haar es la wavelet más simple y es adecuada para detectar cambios abruptos en datos de series temporales. Tiene una forma constante por partes y es fácil de implementar. Es el filtro de wavelet empleado con mayor frecuencia en la literatura sobre series temporales.

4.3.0.2. Wavelet de Morlet

La wavelet de Morlet es una wavelet compuesta por una exponencial compleja multiplicada por una ventana gaussiana, es decir, se define como una constante κ_σ restada de una onda plana y luego localizada por una ventana gaussiana:

$$\Psi_\sigma(t) = c_\sigma \pi^{-\frac{1}{4}} e^{-\frac{1}{2}t^2} (e^{i\sigma t} - \kappa_\sigma)$$

Donde $\kappa_\sigma = e^{-\frac{1}{2}\sigma^2}$ está definido por el criterio de admisibilidad, y la constante de normalización c_σ es la siguiente:

$$c_\sigma = \left(1 + e^{-\sigma^2} - 2e^{-\frac{3}{4}\sigma^2}\right)^{-\frac{1}{2}}$$

4.3.1. Transformada Wavelet Continua (Continuous Wavelet Transform, CWT)

Para analizar series de tiempo que se registran en un eje real continuo, se emplea la transformada wavelet en tiempo continuo, que se define de la siguiente manera:

$$W_x(t, a, \psi) = \int_{-\infty}^{+\infty} x(u) \bar{\psi}_{t,a}(u) du \quad (4.4)$$

donde

x es la serie de tiempo

$W_x(t, a, \psi)$ es la Transformada Wavelet Continua de x

$\psi_{t,a}^*$ es la función wavelet $\psi \in L^2(\mathbb{R})$

a es el factor de escala

La función wavelet escalada y desplazada se describe como:

$$\psi_{t,a}(u) = ||a||^{\frac{1}{2}} \psi\left(\frac{u-t}{a}\right) \quad (4.5)$$

Específicamente, un valor grande de a reduce el tamaño de la wavelet.

Para que la transformación sea invertible, la wavelet $\psi(\cdot)$ debe satisfacer la condición de admisibilidad:

$$C_\psi = \int_{-\infty}^{\infty} \frac{||\hat{\psi}||^2}{w} dw < \infty$$

Como consecuencia, la norma debe estar restringida por una función en la que la integral sea convergente.

4.3.2. Transformada Wavelet Discreta (Discrete Wavelet Transform, DWT)

En la transformada continua de wavelet, los parámetros de tiempo η y frecuencia s están definidos en intervalos continuos. Sin embargo, cuando se restringan a un conjunto de valores discretos, la transformada continua de wavelet se denomina Transformada de Wavelet Discreta (DWT). Por lo tanto, la transformada wavelet discreta puede considerarse como una submuestra de la transformada wavelet continua.

Un marco muy popular es la rejilla dyádica, también conocida como la cuadrícula dyádica, en el cual se escoge un factor de escala de la forma 2^{j-1} , $j = 1, 2, 3$ y luego se seleccionan los tiempos t que están separados por múltiplos de 2^j .

Bajo dicha cuadrícula, la transformada wavelet discreta viene dado por:

$$W_x(n, m, \psi) = \sum_t x(t) \bar{\psi}_{n,m}(t)$$

en donde la función wavelet se define como:

$$\bar{\psi}_{m,n}(t) = 2^{m/2} \psi(2^m t - n) \quad n, m \in \mathbb{Z}$$

Es importante notar que el número entero m (parámetro de escala) determina el ancho de la wavelet, mientras que el índice n (parámetro de desplazamiento) indica la posición en el tiempo. En este contexto, la wavelet madre se expande o reescala en múltiples escalas que son potencias de 2 y se desplaza en desplazamientos enteros.

La DWT inversa se calcula como:

$$x(t) = \sum_n \sum_m W_x(n, m, \psi) \psi_{n,m}(t)$$

en donde $x(t)$ es la función y ψ es la wavelet madre.

La descomposición mediante la transformada wavelet proporcionará un conjunto de series de tiempo en el que cada serie de tiempo tiene un coeficiente que describe la evolución en el tiempo de la señal en una banda de frecuencia específica.

4.3.2.1. Transformada de wavelet discreta de máxima superposición

Una variación del DWT es el Transformada de wavelet discreta de máxima superposición (Maximal overlap discrete wavelet transform, MODWT).

El MODWT se puede conceptualizar como una submuestreo del CWT en escala diádica. Sin embargo, a diferencia del DWT, se tienen en cuenta todos los momentos temporales t en lugar de solo aquellos que son múltiplos de 2^j . La preservación de todos los momentos temporales posibles puede resultar en un resumen más adecuado del CWT, ya que esto puede eliminar ciertos artefactos de alineación que son consecuencia de cómo el DWT submuestra el CWT a lo largo del tiempo.

Los coeficientes de detalle ($\tilde{d}_{j,l}$) y los coeficientes de escala ($\tilde{s}_{j,l}$) que son generados por un algoritmo de MODWT con filtro Haar se definen como:

$$\tilde{d}_{j,l} = d_{j,l}/2^{j/2} \quad y \quad \tilde{s}_{j,l} = s_{j,l}/2^{j/2}$$

donde $d_{j,l}$ y $s_{j,l}$ son los wavelet y los filtros de escala del algoritmo MODWT. Los coeficientes del nivel j ($l = 0, 1, \dots, L - 1$) convolucionan la serie temporal original $Y_t : t = 0, 1, \dots, N - 1$ y el algoritmo piramidal MODWT genera los coeficientes de detalle MODWT ($\hat{D}_{j,t}$) y los coeficientes de escala ($\hat{S}_{j,t}$) de la siguiente manera:

$$\hat{D}_{j,t} = \sum_{l=0}^{L_j-1} \tilde{d}_{j,l} Y_{t-l} \bmod N \quad y \quad \hat{S}_{j,t} = \sum_{l=0}^{L_j-1} \tilde{s}_{j,l} Y_{t-l} \bmod N$$

donde $L_j = (2^j - 1)(L - 1) + 1$. Todos los niveles y la serie suavizada tendrán la misma longitud que la serie original.

Lo que se busca en este tipo de descomposición es una señal como en la Figura 2. Se observa que la señal (izquierda) es caótica, cuando se aplica la *descomposición MODWT* (derecha).

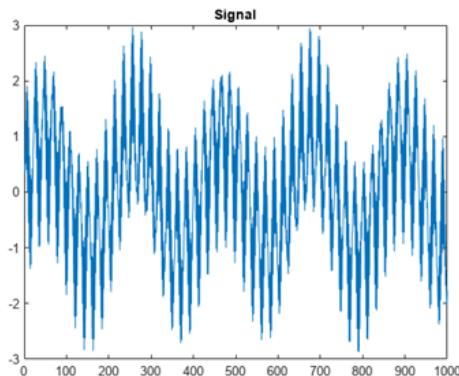


Figura 1. Serie de tiempo caótica.

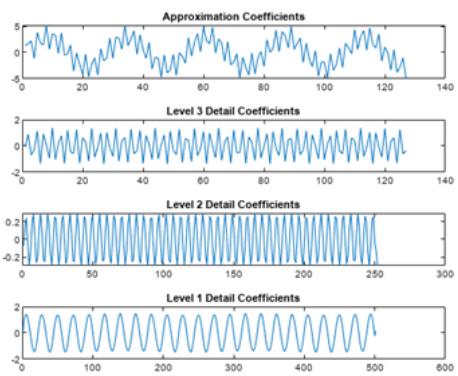


Figura 2. Descomposición MODWT.

En cada nivel de descomposición se obtienen coeficientes de aproximación y de detalle. La gráfica de aproximación muestra una versión suavizada de la señal original, resalta los componentes de baja frecuencia.

Las gráficas de detalle resaltan las componentes de alta frecuencia de la señal original. Muestra los detalles finos o cambios rápidos en la señal.

Se pueden emplear diversos estilos de wavelets según las características de los datos y los objetivos del proyecto. La Figura 10 muestra diferentes modelos de wavelets, y en este caso, se utilizará el wavelet de Haar (segunda imagen de la izquierda en la Figura 3).

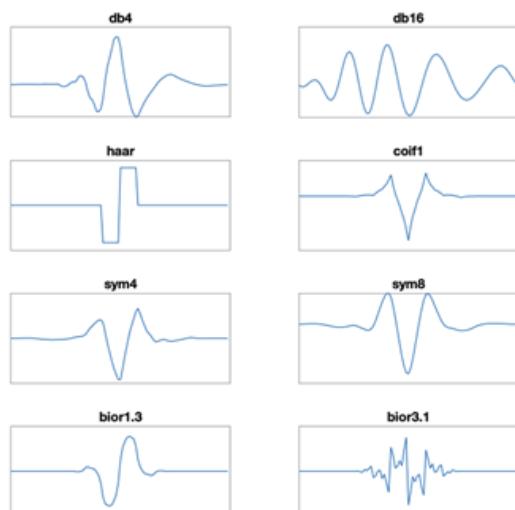


Figura 3. Distintos modelos de wavelets..

En el caso discreto se utilizan filtros para realizar la descomposición de la señal, se utilizan los filtros de *LF* que extraen componentes de baja frecuencia de una señal, selecciona las partes más suaves o lentas de la señal original.

También están los filtros *HF* los cuales seleccionan las partes más abruptas o rápidas de la señal original y produce los coeficientes de detalle. En la Figura 4 se puede ver esto con más detalle.

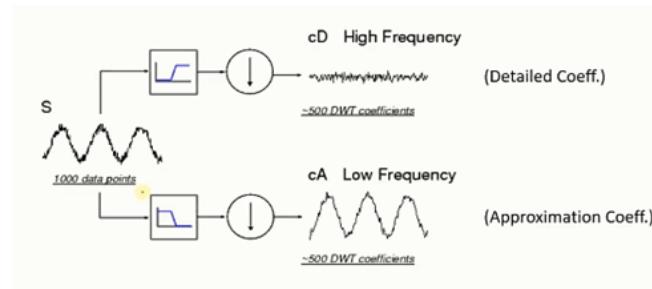


Figura 4. Filtros de descomposición de señal.

En general, la *DWT* se utiliza para dividir una señal en diferentes sub-bandas de frecuencias. Si se generan características en cada una de las sub-bandas y utilizamos la colección de características como entrada para un predictor, este debería ser capaz de distinguir entre los diferentes tipos de señales para poder generar un patrón a futuro.

Específicamente en el caso de *MODWT*, esta técnica busca proporcionar una mayor resolución en los extremos de la señal. Maximiza la sobreposición entre las ventanas de análisis y esto hace que la información de los extremos de la señal se capture de manera más eficiente. Este tipo de método es bueno para tratar cambios bruscos, o sea, series caóticas. En la Figura 5 se observa como se va descomponiendo un pulso con este método.

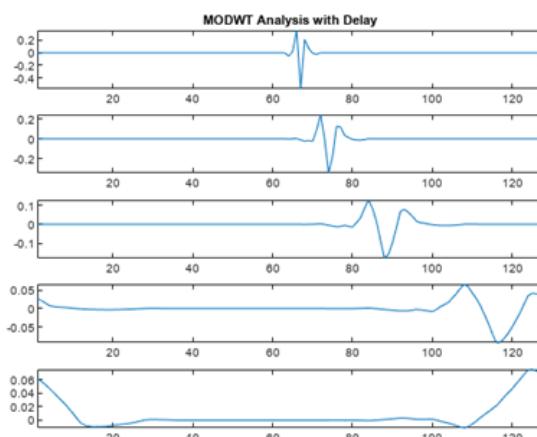


Figura 5. Descomposición MODWT de un pulso.

4.4. Comparación entre transformaciones

La Figura 6 muestra las cuatro cuadrículas básicas utilizadas en el análisis de series temporales.

- La primera ilustra la serie temporal discreta (el tiempo continuo se segmenta en intervalos), donde existe una resolución completa en el tiempo pero una incertidumbre total en la frecuencia.
- En la transformada de Fourier, se divide el plano tiempo-frecuencia en intervalos de frecuencia y se examina el espectro en toda la serie temporal. Aunque se tiene un conocimiento preciso del contenido de frecuencia de la serie, no se dispone de información completa sobre el momento en que esas frecuencias se manifestaron en el tiempo. Como resultado, se logra una alta resolución en frecuencias, pero se mantiene una incertidumbre acerca de cuándo ocurrieron esas frecuencias en el tiempo.
- En la Transformada de Fourier de Tiempo Corto, se segmenta el plano en regiones (ventanas) cuadradas con dimensiones uniformes. Esto implica una ponderación equitativa entre el tiempo y la frecuencia, resultando en una resolución temporal menor que en la serie temporal original y una resolución de frecuencia menor que en la transformada de Fourier convencional. Sin embargo, se tiene cierto conocimiento aproximado sobre cuándo se activan y desactivan frecuencias individuales en el tiempo, proporcionando así una combinación de información temporal y de frecuencia.
- En el caso de wavelets, se obtienen múltiples escalas en el tiempo y la frecuencia, ya que las ventanas presentan tamaños distintos para diferentes frecuencias, siguiendo una progresión geométrica. Esto da lugar a una estructura jerárquica de información temporal y de frecuencia. Las frecuencias más bajas, asociadas a períodos más extensos, se examinan en ventanas amplias a lo largo del eje temporal, mientras que las frecuencias más altas se analizan en ventanas más reducidas en dimensiones temporales. Este enfoque se debe a que las frecuencias bajas experimentan cambios más lentos con el tiempo, requiriendo así una menor exactitud temporal, mientras que las frecuencias más altas experimentan cambios más rápidos, justificando la necesidad de una mayor exactitud temporal en dimensiones más reducidas.

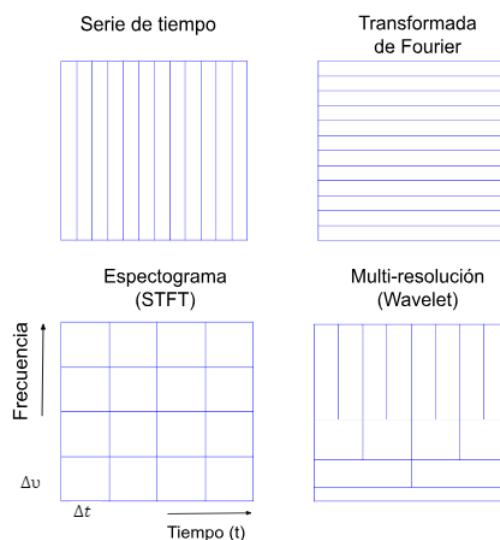


Figura 6. Muestreo en el plano tiempo-frecuencia.

En la Figura 7, se muestran tres wavelets de Haar: $\psi(1, 0)$, $\psi(1/2, 0)$ y $\psi(1/2, 1/2)$ para los dos primeros niveles de la multi-resolución. La wavelet $\psi(1, 0)$ tiene una escala $n = 1$ y un desplazamiento $m = 0$, por lo que no se modifica en tamaño ni se desplaza. Bajo dicha ventana, para obtener la frecuencia más baja, se calcula la transformada de Fourier para toda la parte inferior del plano, ya que no importa cuándo ocurra en el tiempo; se asume que está presente constantemente como la frecuencia base.

En el siguiente nivel, se divide el dominio del tiempo por la mitad utilizando las ventanas $\psi(1/2, 0)$ y $\psi(1/2, 1/2)$, para luego calcular la transformada de Fourier. Esto proporciona menos resolución en frecuencia, pero ofrece información sobre las frecuencias que se encuentran en la primera mitad y la segunda mitad de la serie temporal. En términos generales, a medida que n aumenta, las ventanas se vuelven más pequeñas, mientras que m desliza la ventana correspondiente (determinada por la wavelet) a lo largo de la señal temporal.

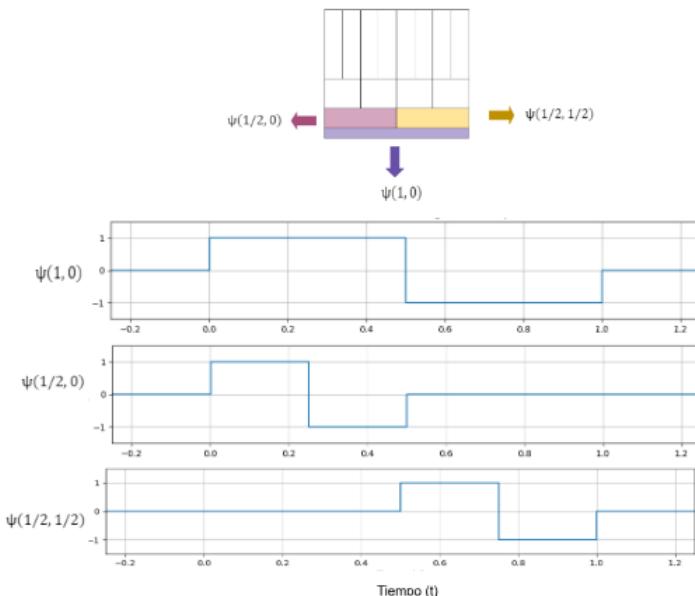


Figura 7. Tres wavelets Haar para los dos primeros niveles de la multi-resolución..

Por lo tanto, se puede decir que la descomposición de wavelet se asemeja a un espectrograma diseñado para optimizar la información en las regiones según sea necesario. Por ejemplo, las frecuencias bajas no requieren tanta información, a diferencia de las frecuencias más altas que necesitan una mayor resolución temporal, y así sucesivamente.

Si las wavelets son ortogonales, entonces la base (conformada por las frecuencias) puede ser utilizada para la proyección, de manera similar a la transformada de Fourier. En esencia, la idea detrás de una descomposición de wavelets es casi idéntica a la transformada de Gabor: al abordar una serie temporal, esta se proyecta sobre una base ortogonal. Sin embargo, en la transformada de wavelets, la base ortogonal no consistirá únicamente en senos y cosenos, sino que será una jerarquía de funciones ortogonales que se reducirán gradualmente en el tiempo o en el espacio, dentro de estas pequeñas ventanas temporales.

4.4.1. Transformada Fourier vs Wavelets

La transformación de Fourier tiene sus limitaciones, para acceder a la información acerca de frecuencia se pierde información sobre el tiempo.

Por ejemplo, se utiliza como demostración de esto el ejemplo de la figura 3, que indica una señal de tráfico que cambia de frecuencia por los distintos colores de semáforo que tiene.

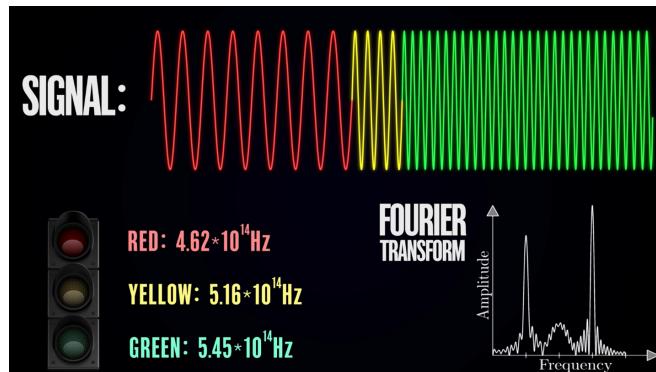


Figura 8. Transformada Fourier Señal Semáforo. Recuperado de [].

Una transformada de Fourier se observa en la esquina inferior derecha de la imagen, la cual tiene 3 picos, por la descomposición de las 3 frecuencias diferentes, pero solo indica que hay 3 frecuencias existentes, no da información de tiempo.

Ahora supóngase que el semáforo se descompone y prende de los 3 colores al mismo tiempo, dando una señal como la que se observa en la figura 4,

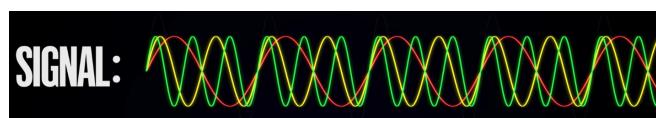


Figura 9. Señal Semáforo Descompuesta. Recuperado de [].

La transformada de Fourier sigue mostrando 3 picos, como se observa en la figura 5, ya que la señal sigue estando compuesta de 3 frecuencias,

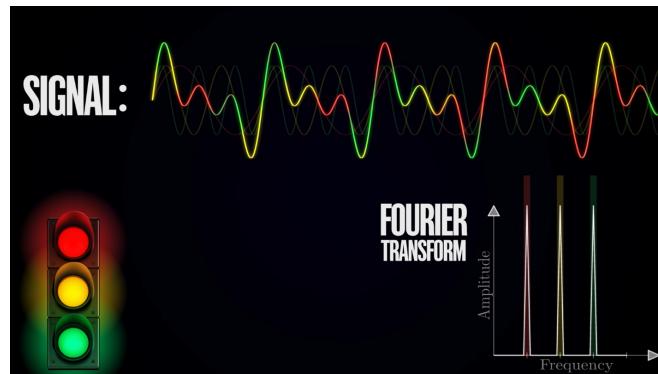


Figura 10. Transformada Fourier Señal Semáforo Descompuesta. Recuperado de [].

Es por eso que para estos casos, la transformada de Fourier no es la mejor opción, ya que es completamente ciega respecto al tiempo.

Ahora, cuando se habla de la transformada de Wavelets, si se piensa primero en el caso de una señal de pulso, al observar la figura 6, se puede observar la transformada de Wavelet de esta señal y cómo es que da información tanto de frecuencia como de tiempo en el escalograma mostrado.

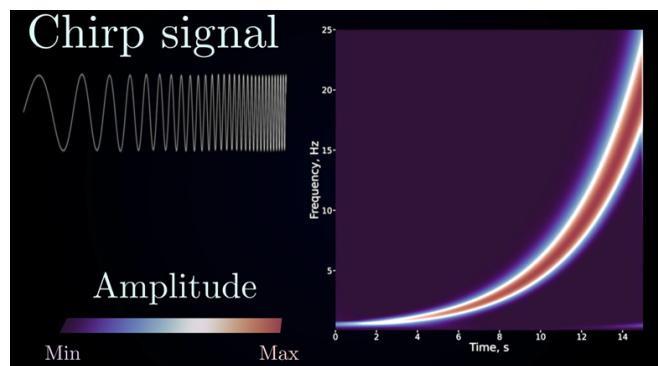


Figura 11. Wavelet de Señal de Pulso. Recuperado de [].

Se observa que instantáneamente la imagen muestra la dinámica esperada, hay un aumento gradual en la frecuencia conforme el tiempo avanza y la amplitud es prácticamente constante.

Regresando al ejemplo del problema del semáforo, se observa en la figura 7, cómo es que la transformada de Wavelet ataca a este mismo,

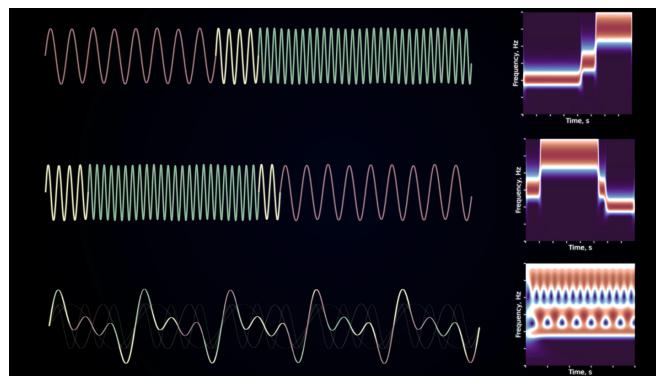


Figura 12. Transformada Wavelet Señal Semáforo. Recuperado de [].

Con esta transformada se puede detectar anomalías instantáneamente.

Considerando otro ejemplo (figura 8) en donde se observa una señal cerebral, se puede ver que hay 3 tipos de ritmos de baja frecuencia que decrecen secuencialmente en frecuencia.

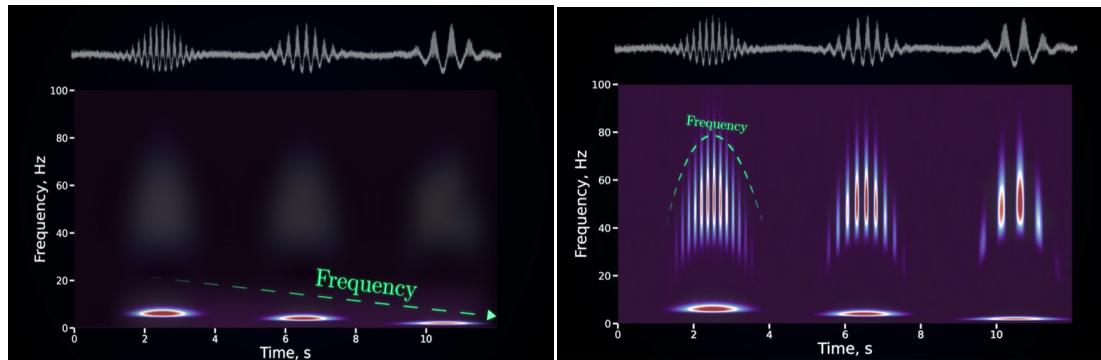


Figura 13. Transformada Wavelet Señal Cerebral. **Figura 14.** Transformada Wavelet Señal Cerebral. Recuperado de [].

También se puede cuantificar la duración de la frecuencia de esos patrones, adicionalmente cada uno de ellos se puede asociar con varios períodos de alta frecuencia. Como se observa en la figura 9.

Ahora, considerando el problema de incertidumbre, la transformada de Wavelet también lo tiene. Analizando las gráficas de las cajas de Heisenberg, como se observa en la figura 10, estos gráficos dan información en el plano tiempo-frecuencia, se dibujan cajas que sus medidas serán proporcionales a la incertidumbre acerca de la frecuencia o tiempo.

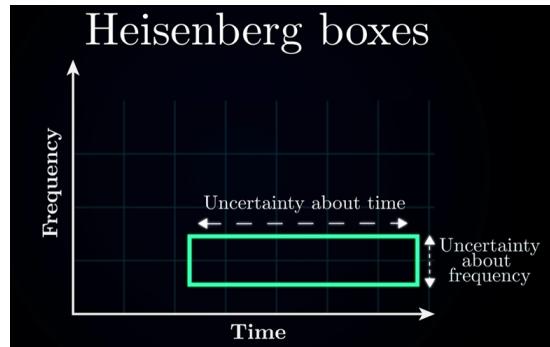


Figura 15. Gráfico de cajas de Heisenberg. Recuperado de [].

Analizando la figura 11, en donde se tienen un gráfico sobre una serie de tiempo sin trabajar se observa una resolución infinita en el dominio del tiempo, pero se tiene nula información acerca de la frecuencia, es por ello que las cajas de incertidumbre son delgadas y altas. En cambio, en la transformada de Fourier (figura 12) se tiene una resolución perfecta, pero no se tiene información sobre el tiempo.

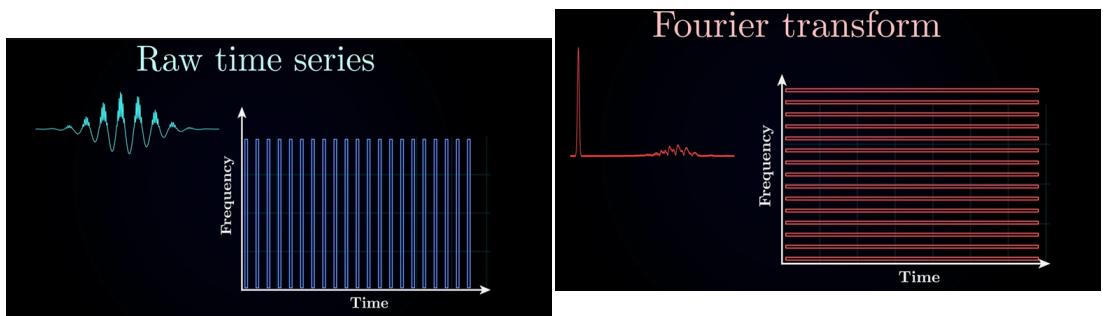


Figura 17. Transformada de Fourier. Recuperado de [].

Figura 16. Serie de Tiempo. Recuperado de [].

Un punto medio sería la transformada de Wavelet ya que está diseñada de manera que en bajas frecuencias las cajas de incertidumbre sean muy anchas para el tiempo, pero para altas frecuencias las cajas son altas y delgadas.

Esto tiene sentido, ya que frecuencias bajas duran mucho tiempo, así que necesitan la resolución de frecuencia alta, en cambio las altas frecuencias son usualmente muy cortas por lo que se necesita una resolución de tiempo alta. Esto se explica gráficamente en la figura 13.

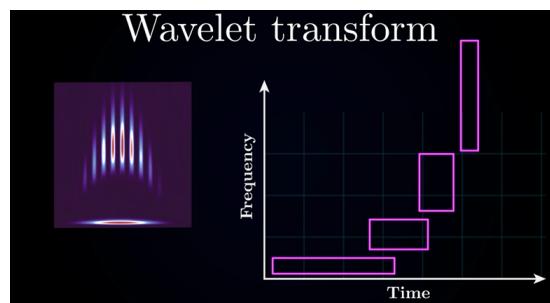


Figura 18. Gráfico de cajas de Heisenberg para Wavelets. Recuperado de [].

4.4.2. STFT vs Wavelets

Ahora, una de las soluciones de la transformada de Fourier es la Transformada de Fourier de Tiempo Corto (STFT).

La STFT proporciona información sobre el momento y las frecuencias en las que ocurre un evento de la señal. Sin embargo, la elección del tamaño de la ventana (segmento) es crucial. Para el análisis tiempo-frecuencia con la STFT, elegir un tamaño de ventana más corto ayuda a obtener una buena resolución temporal a expensas de la resolución de frecuencia. Por el contrario, elegir una ventana más grande ayuda a obtener una buena resolución de frecuencia a expensas de la resolución temporal.

Una vez que elijas un tamaño de ventana, permanece fijo durante todo el análisis. Si puedes estimar los componentes de frecuencia que esperas en tu señal, puedes utilizar esa información para elegir un tamaño de ventana para el análisis.

Se puede observar en la figura 14 que, las frecuencias instantáneas de los dos pulsos en sus puntos iniciales son aproximadamente 5 Hz y 15 Hz. Usa la función auxiliar `helperPlotSpectrogram` para trazar el espectrograma de la señal con un tamaño de ventana temporal de 200 milisegundos. La función auxiliar traza las frecuencias instantáneas sobre el espectrograma como segmentos de línea punteada negra. Las frecuencias instantáneas se resuelven temprano en la señal, pero no tan bien más tarde.

Después en la figura 15 se usa el espectrograma con un tamaño de ventana temporal de 50 milisegundos. Las frecuencias más altas, que ocurren más tarde en la señal, se resuelven ahora, pero las frecuencias más bajas al principio de la señal no lo están.

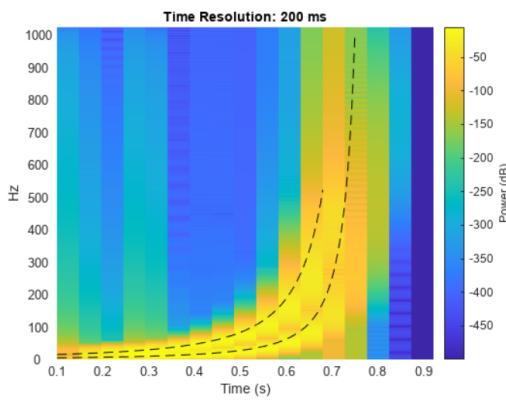


Figura 19. Espectrograma con ventana temporal de 200 milisegundos. Recuperado de [].

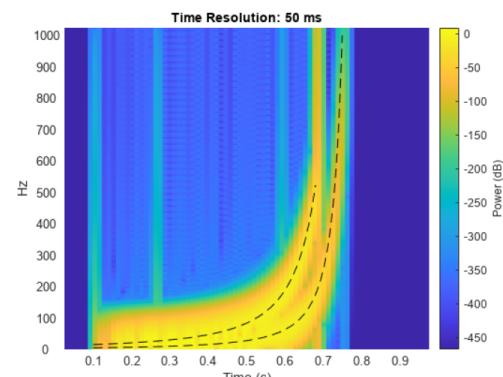


Figura 20. Espectrograma con ventana temporal de 50 milisegundos. Recuperado de [].

Ahora, considerando esta misma señal de dos pulsos, pero para la transformada de Wavelet, se observa en el escalograma en 3D y 2D de las figuras 16 y 17.

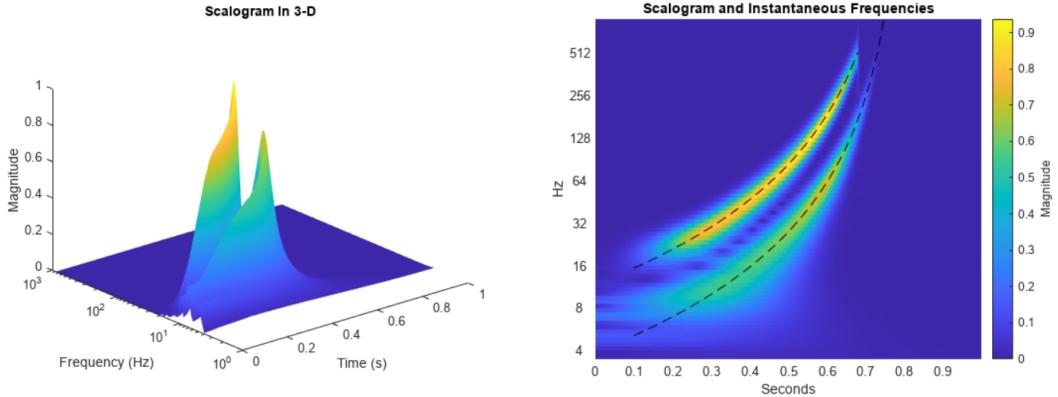


Figura 21. Espectrograma 3D de Wavelet. Recuperado de [1]. **Figura 22.** Espectrograma 2D de Wavelet. Recuperado de [1].

Se observa que la presencia de los dos pulsos en la señal es evidente en el escalograma. Con la Wavelet, se puede estimar con precisión las frecuencias instantáneas a lo largo de toda la duración de la señal, sin preocupación por elegir una longitud de segmento.

5. Metodología

En esta sección presentamos la metodología utilizada, explicamos brevemente los métodos a comparar y las métricas utilizadas para evaluar su rendimiento. Asimismo se comparten los parámetros y condiciones que permitan la reproducibilidad de los experimentos.

Para obtener pronósticos de las diversas series, se utilizaron los algoritmos el W transformer para la wavelet Haar y DB6, que son las más utilizadas por su utilidad. El W-transformer está hecho específicamente para pronóstico a largo plazo de series de tiempo, llegando a tener mejor rendimiento cuando se requiere pronóstico una serie que tiene dependencia a largo plazo. Se realizaron preprocesamientos a las series para validar si la estandarización, diferenciación y descomposición son de utilidad para mejorar la métricas del paper de W-Transformer original.

5.1. Descripción de los datos

A continuación, se describen brevemente las bases de datos que se van a utilizar y sus características importantes. Después de las descripciones individuales se encuentra un breve resumen de características importantes de todas ellas en la tabla 1. El orden en que se presentan aquí es el mismo que en la implementación y el código que acompaña este reporte.

Inflación del IPC. En el Índice de Precios al Consumidor y la tasa de interés trimestral se registra para bonos del Tesoro (línea discontinua), desde el primer trimestre de 1953 hasta el segundo trimestre de 1980, con 110 observaciones. Estos datos se obtienen de Newbold y Bos. Para la prueba de

estacionariedad tuvo un p-value de 0.938 por lo que es una serie no estacionaria. Para la prueba de no trend-stationary de KPSS se obtiene un estadístico de 1.459 por lo que describe una comportamiento no estacionario con tendencia media, de la prueba de exponente de Hurst se obtiene que la serie exhibe un comportamiento de dependencia a largo plazo con un valor de exponente de 0.518, por último para la prueba de Terasvirta se obtiene un p valor de 0.04015, que indica la presencia de no linealidad en la serie de tiempo.

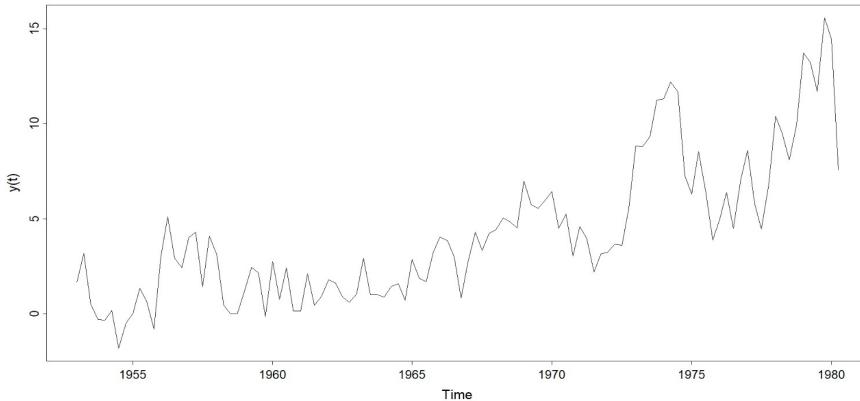


Figura 23. Inflación del IPC.

Airline Dataset. International Airline Passengers, 1949–1960; contiene el total mensual de pasajeros internacionales desde enero de 1949 hasta diciembre de 1960, en miles de pasajeros. En total son 144 datos. Es una base de datos típica para el estudio de series de tiempo y fue discutida desde Box y Jenkins en 1976. Para la prueba ADF se obtiene un p-v-value de 0.991 por lo que es una serie no estacionaria. Para la prueba KPSS (Kwiatkowski-Phillips-Schmidt-Shin) obtiene un estadístico de 1.65 con un p-value: 0.01, por lo que es una serie no trend-stationary. De la prueba del exponente Hurst se obtiene un exponente de 0.5184 por lo que exhibe una dependencia a largo plazo. Finalmente para la linealidad se aplicó la prueba de Terasvirta obteniéndose un p-valor bajo (0.0472) lo que podría indicar la presencia de no linealidad en la serie de tiempo.

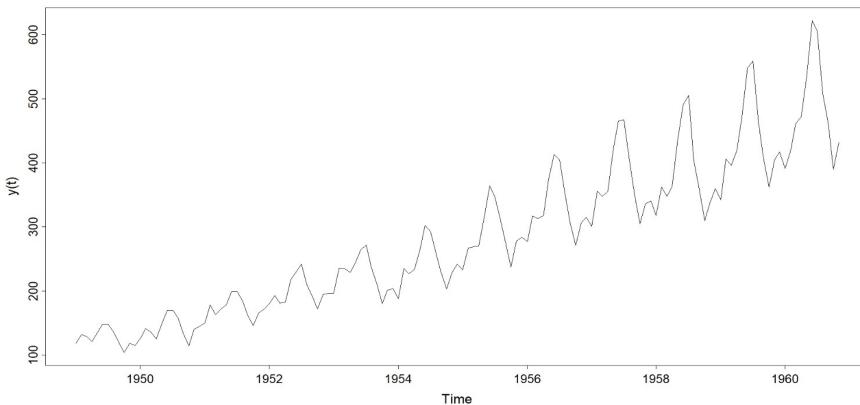


Figura 24. Airline Passengers.

Reserva Federal. Índice mensual de la Reserva Federal de los EEUU con datos de 1948 a 1978. Contienen la producción mensual de EE. UU, medida por el Índice de Producción de la Junta de la Reserva Federal. En total tiene 372 registros. La prueba de ADF arroja un p-value de 0.982 por lo que es una serie no estacionaria. De la prueba KPSS se obtiene un estadístico de 3.14 y un p-value de 0.01, por lo que es una serie no trend-stationary. De la prueba del exponente Hurst se obtiene un exponente de 0.7086 por lo que exhibe una dependencia a largo plazo. Para la linealidad se aplicó la prueba de Terasvirta obteniéndose un p-valor de 0.04848 que indica la presencia de linealidad, pasando apenas la prueba para una significancia del 0.05.

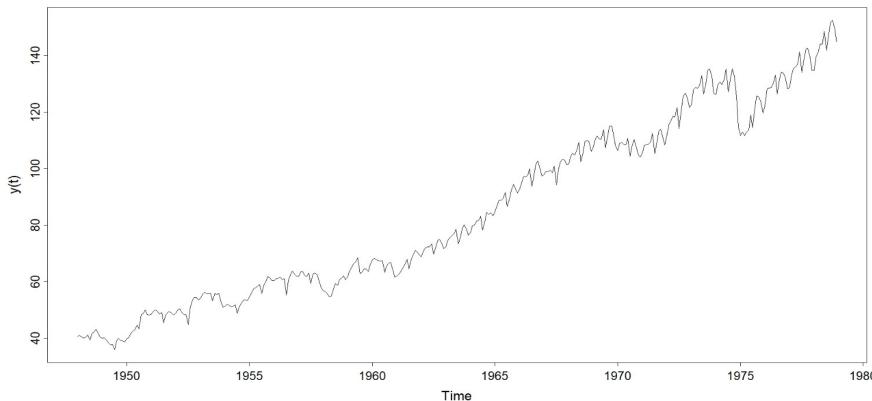


Figura 25. Reserva federal USA.

Dengue de Colombia. Contiene 626 registros semanales de número de infectados con Dengue en Colombia, de 2005 a 2016. Estos datos fueron tomados en cuenta porque son parte de un estudio biológico (epidemia) que se exhibe en el paper original del W-Transformer. En la prueba ADF tuvo un p-value de 0.157 por lo que es una serie no estacionaria. En la prueba de KPSS se obtiene un estadístico de 2.065 y un p-value de 0.01 por lo que la serie no sigue un comportamiento trend-stationary. Su índice de Hurst es de 0.4423 por lo que esta serie no tiene dependencia a largo plazo. Finalmente de la prueba de Terasvirta se obtiene un p-value muy pequeño de 0.00083 por lo que no exhibe un comportamiento lineal.

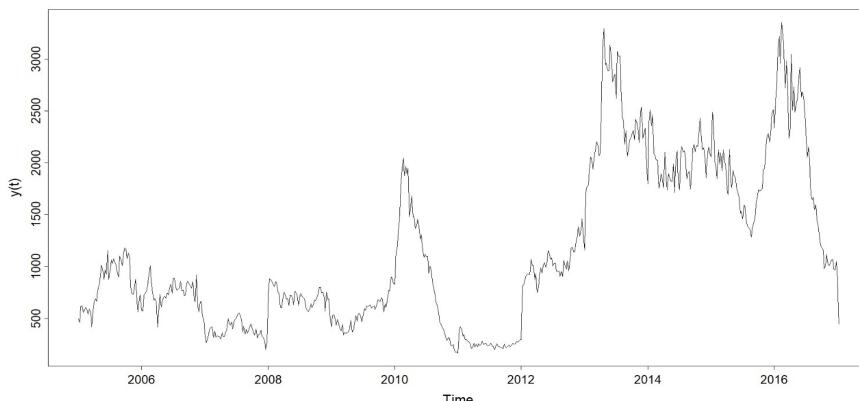


Figura 26. Dengue de Colombia.

Dengue en Bangkok. El conjunto de datos consiste en 180 medidas mensuales de número de casos de dengue, en Bangkok, de 2003 a 2017. También es una serie biológica que indica el comportamiento de una epidemia, también es una serie tomada del paper de W-Transformer, para efectos de comparación y cubrir diferentes áreas de aplicación. En esta serie se observa un comportamiento particular que es de interés en 2015, si fuera una serie de otro ámbito podría considerarse un evento estructural o aislado, en la predicción veremos que es parte del proceso que la misma serie describe. De la prueba ADF se obtiene un p-value de 4.7179×10^{-6} lo que indica que la serie es estacionaria. Para la prueba KPSS el estadístico es de 0.5301 y el p-value es de 0.0348 por lo que no describe un comportamiento trend-stationary. Se obtuvo un índice de Hurst de 0.417 por lo que la serie tampoco es dependiente a largo plazo. Finalmente de la prueba Terasvirta se obtuvo un valor 2.2×10^{-6} por lo que no describe linealidad.

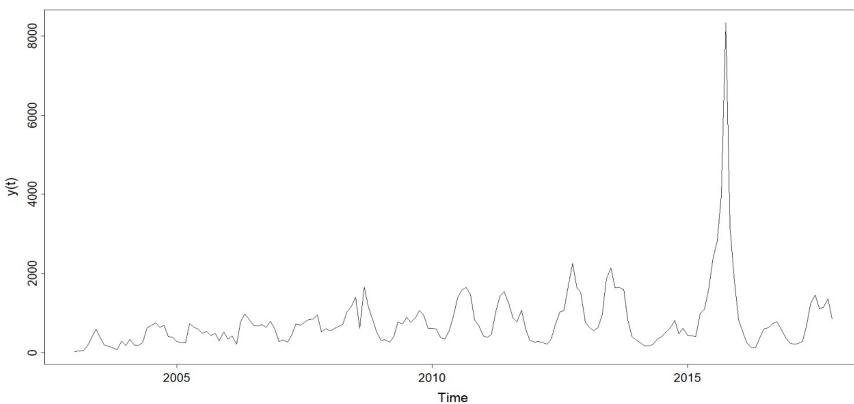


Figura 27. Dengue de Bangkok.

Japan Flu. Esta serie describe un comportamiento epidémico con mayor caos, por lo que es de interés para validar si los transformers y variantes pueden predecir valores futuros apesar de sus características. El conjunto de datos consiste en un reporte semanal de casos de influenza en Japón, de 1998 a mediados de 2016. Tiene 964 registros. Para la prueba de ADF se obtiene un valor p de 6.596×10^{-12} , por lo que la serie sigue un comportamiento estacionario. De la prueba KPSS se obtiene un valor de estadístico de 0.1626 por lo que la serie tiene un comportamiento trend-stationary, aunque como se visualiza la gráfica simplemente se asegura que no hay evidencia estadística (por la prueba) que pueda rechazar la hipótesis nula. En cuanto a la prueba de dependencia a largo plazo se obtiene un valor de Hurst de 0.2581 por lo que la serie no tiene dicho comportamiento, finalmente de la prueba de linealidad se obtiene un valor p de 0.001747 por lo que no describe un comportamiento lineal.

Netflix stock prices. Este dataset es el precio de cierre de stock price diario de cierre desde junio 2021 hasta junio 2022. Se consideran solamente días hábiles. Los datos se recopilaron de Yahoo Finance y son registros diarios. Contiene 254 registros. Este data set también es parte del paper de W-transformer y es de importancia porque es de comportamiento financiero y apoya a validar el uso del wavelet DB6 y de la diferenciación de series. Para la prueba ADF se obtuvo un valor de 0.984 lo que indica que no es una serie estacionaria, la estadística KPSS es de 1.441 que indica que no sigue un comportamiento trend-stationary. El exponente de Hust que se obtuvo fue de 0.649 por lo que se indica que la serie tiene un comportamiento con dependencia a largo plazo. Finalmente de la prueba de linealidad de Terasvirta el p-value fue de 0.4119 por lo que no describe un comportamiento lineal.

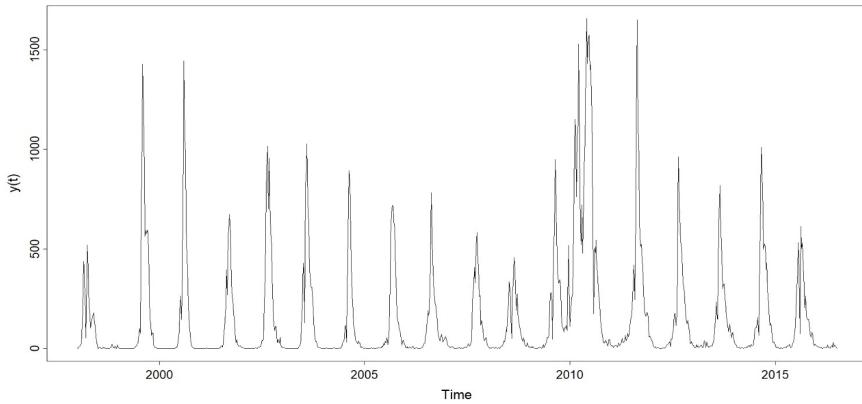


Figura 28. Influeza de Japón.

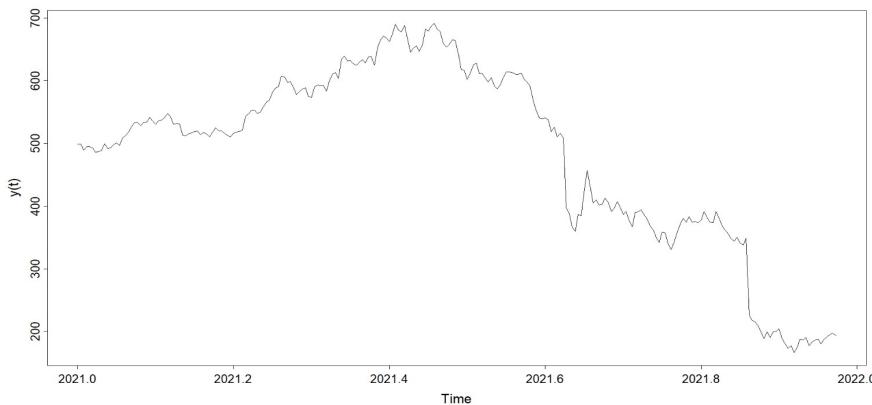


Figura 29. Netflix stock price.

Sunspots Daily. Esta serie de datos contiene el número total de sunspots diarios derivado de la fórmula $R = Ns + 10Ng$, donde Ns es el número de manchas y Ng es el número de grupos contados alrededor del disco solar. Contiene 75,179 registros desde 1818 a octubre de 2023. Esta serie es de interés en el ámbito climatológico, y se retoma también del paper de W-Transformer por su utilidad para validar el rendimiento del transformer. En la imagen de la figura 30 se observan solo 265 registros de septiembre de 2021 a junio de 2022 para efectos de visualización. En la prueba ADF se obtuvo un p-value de 1.291×10^{-5} por lo que tiene un comportamiento estacionario. De la prueba KPSS se obtuvo un p-value de 0.01 y un estadístico de 1.058 por lo que la serie no es trend-stationary, para la prueba del exponente de Hurst se obtuvo $H= 0.247$ por lo que la serie no describe un comportamiento de larga dependencia. Finalmente tuvo un p-value de 0.856 en la prueba de Terasvirta por lo que es una serie con comportamiento lineal.

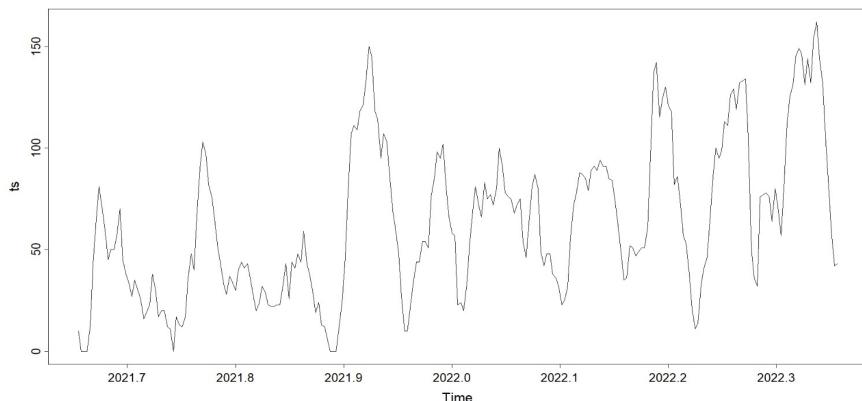


Figura 30. Sunspots Daily.

Datos	Frecuencia	Observaciones	Valor Min-Max	Características
Inflación de IPC	Trimestral	110	-1.822-15.568	No lineal, no estacionaria, no trend-stationary, dependencia a largo plazo
Airline data	Mensual	144	104-622	No lineal, no estacionaria, no trend-stationary, dependencia a largo plazo
Reserva federal	Mensual	372	36-152.6	Lineal, no estacionaria, no trend-stationary, dependencia a largo plazo
Dengue Colombia	Semanal	626	164-3354	No lineal, no estacionaria, no trend-stationary sin dependencia a largo plazo
Dengue Bangkok	Mensual	180	29-8337	No lineal, estacionaria, no trend-stationary, sin dependencia a largo plazo
Japan Flu	Semanal	964	0-1656	No lineal, estacionaria, trend-stationary, sin dependencia a largo plazo
Netflix	Diario hábil	254	166.37-691.69	No lineal, no estacionaria, no trend-stationary, dependencia a largo plazo
Sunspots	Diario	75,179	0-162	Lineal, estacionaria, no trend-stationary, sin dependencia a largo plazo

Cuadro 1. Descripción de características de bases de datos..

5.2. Pruebas de estacionariedad, no linealidad y dependencia a largo plazo

Las pruebas que se realizaron para validar el comportamiento general de las series de datos son para estacionariedad (ADF), trend-stationary (KPSS), dependencia a largo plazo (exponente Hurst) y prueba de no linealidad (Terasvirta). Esta última prueba se hizo directamente en R ya que está implementada con un solo comando y en Python requería varias librerías e implementación por tratarse de una prueba basada en redes neuronales, además el resultado no siempre era compatible con la del paquete estadístico R. A continuación se describen con detalle que comportamiento se trata de describir en las series y en qué consisten cada una de las pruebas aplicadas. Cabe destacar que las pruebas estadísticas se realizaron a las series originales y también a las series preprocesadas con estandarización, diferenciación o descomposición.

Prueba de estacionariedad

La estacionariedad en series temporales es una propiedad esencial para muchos modelos clásicos de pronóstico estadístico. Una serie temporal se considera estacionaria si sus propiedades estadísticas no cambian con el tiempo. La regularidad se conceptualiza a través de los términos de estacionariedad y estacionariedad débil, donde la constancia en la media es fundamental. Estos conceptos son cruciales para analizar datos de series temporales cuando solo se dispone de una realización.

Una serie temporal estrictamente estacionaria es aquella para la cual el comportamiento probabilístico de cualquier conjunto de valores

$$\{x_{t_1}, x_{t_2}, \dots, x_{t_k}\}$$

es idéntico al conjunto desplazado en el tiempo

$$\{x_{t_1+h}, x_{t_2+h}, \dots, x_{t_k+h}\}$$

Es decir,

$$Pr\{x_{t_1} \leq c_1, \dots, x_{t_k} \leq c_k\} = Pr\{x_{t_1+h} \leq c_1, \dots, x_{t_k+h} \leq c_k\} \quad (5.1)$$

para todos los $k = 1, 2, \dots$, todos los puntos temporales t_1, t_2, \dots, t_k , todos los números c_1, c_2, \dots, c_k , y todos los desplazamientos temporales $h = 0, \pm 1, \pm 2, \dots$

Si una serie temporal es estrictamente estacionaria, entonces todas las funciones de distribución multivariada para subconjuntos de variables deben coincidir con sus contrapartes en el conjunto desplazado, para todos los valores del parámetro de desplazamiento h . Por ejemplo, cuando $k = 1$, 5.1 implica que

$$Pr\{x_s \leq c\} = Pr\{x_t \leq c\} \quad (5.2)$$

para cualquier punto temporal s y t . Esta afirmación implica, por ejemplo, que la probabilidad de que el valor de una serie temporal muestreada por hora sea negativo a la 1 a. m. es la misma que a las 10 a. m. Además, si existe la función de media, μ_t , de la serie, la ecuación 5.2 implica que $\mu_s = \mu_t$ para todos los s y t , y por lo tanto, μ_t **debe ser constante**. [7]

La definición previa de estacionariedad es demasiado rigurosa para la mayoría de las aplicaciones y difícil de evaluar con un solo conjunto de datos. En lugar de imponer condiciones a todas las posibles distribuciones de una serie temporal, se opta por una versión más flexible que solo establece condiciones sobre los dos primeros momentos de la serie.

Una **serie temporal débilmente estacionaria**, x_t , es un proceso de varianza finita tal que:

- La función de valor medio, μ_t , es constante y no depende del tiempo t .
- La función de autocovarianza, $\gamma(s, t)$, depende de s y t solo a través de su diferencia $|s - t|$.

De ahora en adelante, se usará el término “estacionaria” para hacer referencia a estacionariedad débil.

ADF Fuller para estacionariedad

La prueba de Dickey-Fuller aumentada (ADF) se utiliza para evaluar la presencia de una raíz unitaria en una serie temporal, lo que indica la no estacionariedad. La hipótesis nula (H_0) de la prueba es que la serie temporal tiene una raíz unitaria, es decir, que no es estacionaria. La hipótesis alternativa sugiere que la serie es estacionaria después de aplicar diferenciación.

En Python, se tiene implementada la prueba con el comando `adfuller`.

KPPS para estacionariedad-tendencia (stationary-trend).

La estacionariedad y la estacionariedad en tendencia son conceptos distintos en el análisis de series temporales. La estacionariedad se refiere a propiedades estadísticas constantes en el tiempo, sin patrones sistemáticos cambiantes, ni tendencias ni variaciones cíclicas en la serie.

La estacionariedad en tendencia (Trend-Stationary) en series temporales se refiere a la capacidad de considerar una serie estacionaria a pesar de tener una tendencia determinista. Esta tendencia, ya sea lineal o no lineal, es conocida y se puede eliminar mediante diferenciación. Aunque la tendencia está presente, se puede lograr que la serie sea estacionaria restando o diferenciando dicha tendencia. Este concepto es más flexible que la estacionariedad pura, ya que permite la presencia de una componente de tendencia en la serie, pero esta tendencia es modelada o ajustada antes de realizar análisis adicionales.

La prueba Kwiatkowski-Phillips-Schmidt-Shin (KPSS) es una prueba de estacionariedad utilizada en el análisis de series temporales. A diferencia de la prueba de Dickey-Fuller aumentada (ADF), que se centra en la presencia de raíces unitarias, la prueba KPSS evalúa si una serie temporal es estacionaria alrededor de una tendencia. La prueba KPSS se formula con las siguientes hipótesis nula y alternativa:

- Hipótesis Nula (H_0): La serie es estacionaria alrededor de una tendencia determinista.
- Hipótesis Alternativa (H_1): La serie tiene una raíz unitaria, lo que implica que no es estacionaria.

La prueba KPSS se basa en la idea de que si la serie es estacionaria, entonces la varianza alrededor de una tendencia debería ser constante en todas las ventanas temporales. La prueba se realiza mediante la estimación de un estadístico que compara la varianza de subconjuntos de la serie temporal con la varianza global.

El procedimiento general de la prueba KPSS implica los siguientes pasos:

- Diferenciación: Si la serie no es estacionaria en niveles, se realiza una diferenciación para eliminar una posible tendencia.
- Estimación: Se estima un estadístico que mide la varianza en diferentes subconjuntos de la serie.
- Comparación: El estadístico estimado se compara con los valores críticos para determinar si se rechaza la hipótesis nula.

Al igual que la prueba ADF, la prueba KPSS está implementada en Python mediante el comando `kpss`.

Pruebas de no linealidad

La prueba de Terasvirta es un test que identifica no linealidades en series temporales. Evalúa si la relación subyacente en los datos puede ser eficazmente modelada por un enfoque lineal.

La prueba de Terasvirta utiliza un enfoque basado en redes neuronales para evaluar si una componente no lineal es necesaria para modelar la serie temporal. Se comparan los errores de predicción entre un modelo lineal y un modelo no lineal (basado en una red neuronal) para determinar si la no linealidad mejora significativamente la capacidad de predicción.

Como ya se comentó antes, se prefirió usar la implementación existente en el paquete estadístico R dado que se puede ejecutar con un solo comando `terasvirta.test(x)`.

Pruebas de dependencia a largo plazo

Para calcular el parámetro de dependencia a largo plazo o auto-similaridad de una serie temporal, se utiliza el exponente de Hurst (H). El valor de H se calcula mediante la función `hurstexp` del paquete R `pracma` o en Python, con el comando `compute_Hc` de la librería Hurst.

El exponente de Hurst se emplea como una medida de la memoria a largo plazo de una serie temporal. Este índice está relacionado con las autocorrelaciones de la serie temporal y la velocidad a la que disminuyen a medida que aumenta el rezago entre pares de valores.

En términos prácticos, un exponente de Hurst mayor a 0.5 sugiere una tendencia a la persistencia en la serie, mientras que un valor menor a 0.5 indica una tendencia a la antipersistencia.

5.3. W-Transformer

El marco propuesto en [1] introduce un novedoso modelo conocido como W-Transformer el cual combina el algoritmo MODWT con la arquitectura de transformer. Esta integración tiene como objetivo capturar efectivamente la dependencia de temporal a largo plazo, respecto al resto de metodologías de deep learning, e incluso es capaz de modelar los comportamientos no lineales mediante el uso de transformaciones Wavelets, al mismo tiempo que aborda la no estacionariedad y la no linealidad en los datos de series de tiempo.

Los componentes clave del W-Transformer se describen a continuación para efectos de entendimiento de la arquitectura completa.

1. Descomposición MODWT.

Primero, la serie de datos es sometida a un proceso de descomposición mediante la transformada wavelet discreta de máxima superposición (MODWT), usando el filtro de Haar. Esta descomposición ayuda a tratar la no estacionariedad y la estacionalidad de las series temporales, pues al incorporar wavelets dentro de la arquitectura de transformer, el modelo ofrece una descomposición detallada de los coeficientes de baja y de alta frecuencia, que permite una mejor separación de la señal del ruido en los datos de series temporales.

2. Redes paralelas.

Cada nivel, incluyendo el componente suave (smooth component) de la descomposición es procesada mediante un transformer codificador-decodificador independiente, en donde la parte del codificador

toma como entrada el historial del nivel de la serie temporal, mientras que la parte del decodificador predice los valores futuros del nivel de manera auto-regresiva.

Como se muestra en la Figura 31, la descomposición pasa por dos redes paralelas (el encoder y el decoder), cada una de las cuales comienza con un embedding.

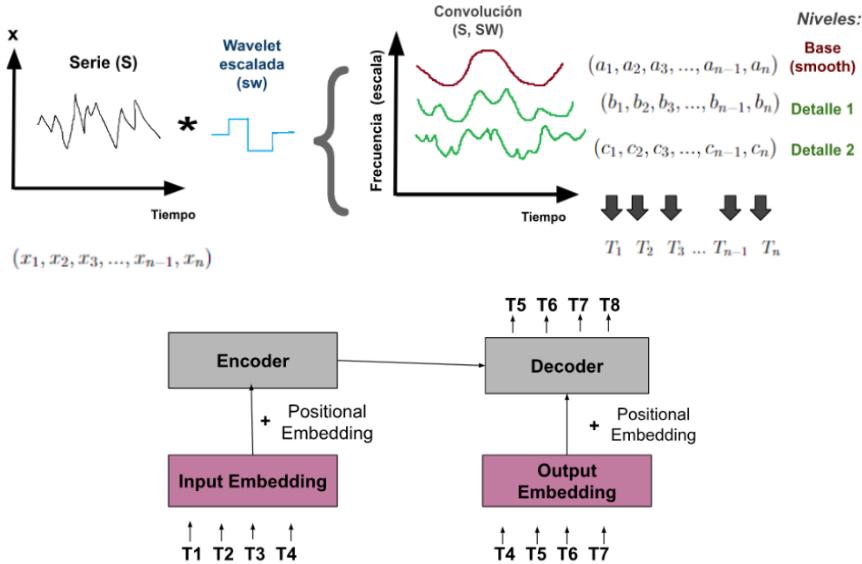


Figura 31. Ejemplo del esquema general del modelo con un input_size=4.

En particular, para mantener seguimiento del orden de las entradas, a cada salida del embedding se le suma un vector que representa una secuencia de senos y cosenos. Por ejemplo, como se observa en la Figura 32, los valores de posición para el primer input proviene del eje y correspondiente, y así sucesivamente. A medida que el índice posicional (o posición en la secuencia) aumenta, la perturbación introducida por las funciones seno y coseno hace que el embedding de la señal codificada por posición trace un círculo unitario en el espacio vectorial. Esta codificación dinámica ayuda al modelo a capturar información posicional sin depender únicamente del contenido del input embedding.

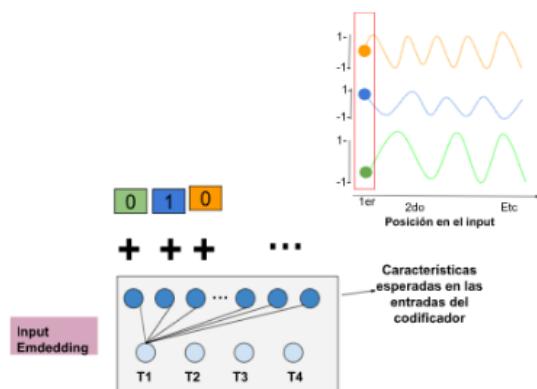


Figura 32. Ejemplificación del embedding posicional para el primer input.

El encoder toma la entrada de un nivel de la descomposición de los datos de entrenamiento por batches y el decoder toma la salida del mismo nivel de descomposición de los datos de entrenamiento también por lotes.

El decodificador está vinculado al codificador mediante un mecanismo de atención. De esta manera, el decodificador puede aprender a atender la parte más útil de los valores históricos del nivel de la serie antes de realizar una predicción.

La salida tanto del encoder como del decoder es procesada adicionalmente por un segundo decoder.

3. Arquitectura de encoder-decoder

El encoder se compone de una capa de atención de multi-cabezales, una capas directas feedforward y capas de normalización.

El primer decoder incluye una capa de atención multi-cabeza oculta y otra de normalización, mientras que el segundo decoder replica la estructura del encoder.

El mecanismo de auto-atención para cualquier transformer, dado en el paper original [5], está dado como sigue:

$$Attn(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d}} \right) V,$$

donde Q es la matriz query, K es la matriz key, V es la matriz value, y la matriz d es la dimensión de Q , K y V . Posteriormente la función softmax se usa para obtener los pesos de los valores que se introducirán en la matriz Q . A continuación, las keys, las queries y los values se proyectan linealmente, y estas proyecciones realizan la función de atención en paralelo dando lugar a los valores de salida.

Para las capas de atención multi-cabezales se definen también estas cabezales como sigue:

$$head_p = Attn \left(QW_p^Q, KW_p^W, VW_p^V \right),$$

donde las proyecciones son matrices de parámetros denotadas por W_p^Q , W_p^W , W_p^V . Por último, la capa de atención de multi-cabezales es simplemente una concatenación de la cabeza para m -veces, donde m es un hiperparámetro que suele fijarse en un valor entero preestablecido 8, según la misma referencia [5] y d la dimensión de Q y V definida por $d = d_{model}/m$, donde d_{model} es la dimensión de la salida de la capa del embedding. Para atender conjuntamente a la información de varios subespacios, utilizamos la capa de atención multicabezal definida como sigue

$$\text{MultiHead}(Q, K, V) = \text{Concat}(head_1, \dots, head_m)W^O,$$

donde W^O es la matriz de proyección de la capa de salida.

El decodificador utiliza una atención autoenmascarada para que evitar que la red haga trampa al predecir valores pasados mediante la observación de valores futuros. La atención autoenmascarada se implementa mediante la introducción de una máscara que se aplica a la matriz de atención antes de calcular los pesos de atención. Esta máscara tiene valores cero en las posiciones que corresponden a elementos futuros y valores no nulos en las posiciones actuales y pasadas. De esta manera, al calcular la atención, los elementos en una posición dada solo pueden atender a los elementos anteriores o en la misma posición, evitando la dependencia de información futura.

Se aplica una capa lineal y una capa softmax a la salida del segundo decoder para obtener los parámetros de peso del modelo, P_0, P_1, \dots, P_j .

3. Pronóstico final

Una vez completadas todas las épocas de entrenamiento, el modelo final está listo para realizar pronósticos un paso a hacia adelante.

Para obtener las predicciones finales, cada nivel de descomposición y el componente suavizado son pronosticados utilizando sus respectivos modelos transformer. Posteriormente, se aplica la inversa de la transformación MODWT para revertir el proceso de descomposición. Esto permite integrar los pronósticos individuales.

4. Pronóstico multi-step

Para realizar las predicciones se generan repite el proceso de manera iterativa aplicándolo al modelo W-Transformers ajustado.

Para generar la predicción del lag h (\hat{Y}_{N+h}), seguimos la siguiente fórmula:

$$\hat{Y}_{N+h} = \sum_{j=1}^J \hat{D}_{j,N+h} + \hat{S}_{J,N+h}$$

donde

$$\begin{aligned}\hat{D}_{j,N+h} &= f(D_{j,1}, D_{j,2}, \dots, D_{j,N}); j = 1, 2, \dots, J, \\ \hat{S}_{j,N+h} &= f(S_{J,1}, S_{J,2}, \dots, S_{J,N}),\end{aligned}$$

y f es el modelo W-transformer.

La Figura 33 muestra la arquitectura general del W-transformer.

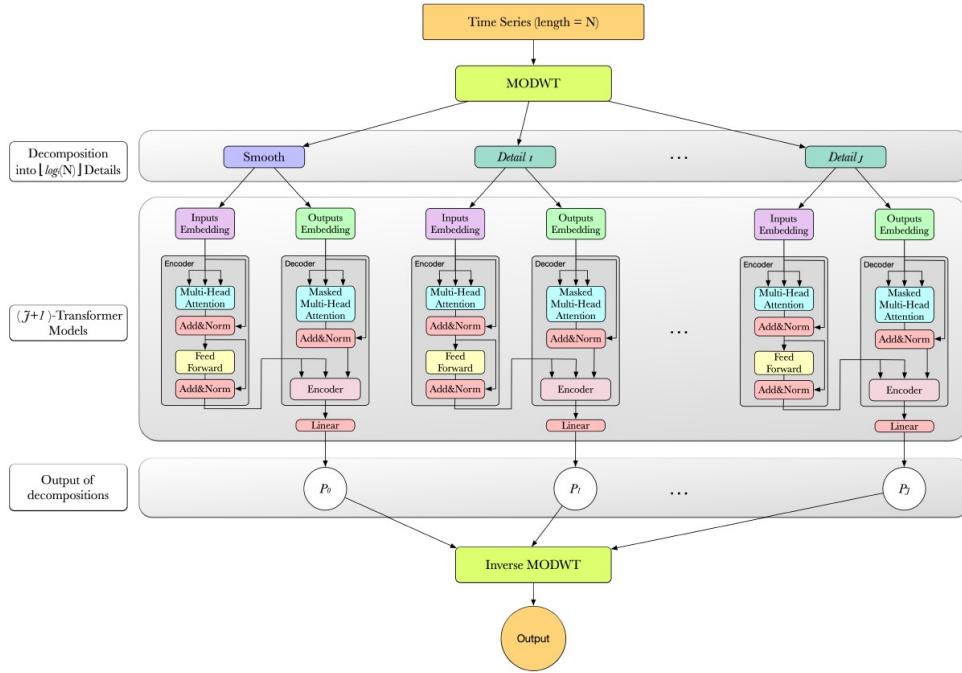


Figura 33. Diagrama de arquitectura W-Transformer tomado de [5].

5.4. Métricas de evaluación

Para permitir la evaluación del rendimiento en la tarea de pronóstico de series de tiempo se consideran los siguientes métricas.

MAPE (Mean Absolute Percentage Error): El MAPE calcula el error porcentual promedio entre las predicciones y los valores reales de la serie de tiempo. Es una métrica que mide la precisión relativa de las predicciones y se expresa como un porcentaje. Se calcula como la media del valor absoluto de los errores porcentuales para cada punto de tiempo.

MASE (Mean Absolute Scaled Error): El MASE es una métrica que compara el error absoluto promedio de las predicciones con el error absoluto promedio de un modelo ingenuo o de referencia. Se utiliza para evaluar si el modelo de pronóstico es mejor que un enfoque simple de referencia. Un valor de MASE menor a 1 indica que el modelo es mejor que el enfoque de referencia.

RMSE (Root Mean Square Error): El RMSE calcula la raíz cuadrada del error cuadrado promedio entre las predicciones y los valores reales de la serie de tiempo. Es una métrica que penaliza de manera más significativa los errores grandes. Cuanto menor sea el RMSE, mejor será la precisión del modelo.

MAE (Mean Absolute Error): El MAE calcula el error absoluto promedio entre las predicciones y los valores reales de la serie de tiempo. Mide la magnitud promedio de los errores sin considerar su dirección. El MAE es útil para evaluar la magnitud promedio de los errores en las predicciones.

SMAPE (Symmetric Mean Absolute Percentage Error): El SMAPE es una métrica que combina elementos del MAPE y el MAE. Calcula el error porcentual promedio simétrico entre las predicciones

y los valores reales de la serie de tiempo. Es útil para evaluar la precisión de las predicciones y se expresa como un porcentaje.

Estas métricas ya están implementadas en Python dentro del módulo `metrics` de la librería `darts`.

5.5. Implementación

Configuración del Entorno de Implementación

La implementación de la red profunda W-Transformer se llevó a cabo en un entorno computacional equipado con una unidad de procesamiento gráfico (GPU) NVIDIA Tesla T4. La configuración del sistema incluye el controlador de GPU versión 525.105.17 y la versión 12.0 de CUDA. Durante el proceso de implementación, se monitorizó el rendimiento de la GPU, que operó a una temperatura de 35°C con un consumo de energía de 9W, dentro de una capacidad máxima de 70W. La memoria de la GPU tiene una capacidad de 15,360 MiB. Esta configuración proporciona un entorno propicio para el desarrollo y entrenamiento eficiente de la red profunda, aprovechando la potencia computacional de la GPU Tesla T4 con el soporte de las tecnologías CUDA, además permitió determinar las métricas que se presentan en la sección de resultados.

Implementación del W-transformer y Tratamientos de preprocesamiento

Se llevó a cabo la implementación del W-transformer en Python con la configuración mencionada, centrando la atención en 8 bases de datos específicas que ya se revisaron. El objetivo principal era evaluar y comparar el rendimiento del W-transformer bajo diferentes condiciones y preprocesamientos. A continuación, se describen detalladamente los cinco tratamientos aplicados a cada serie temporal.

- **1. W-transformer Directo:** En este enfoque, se aplicó directamente el W-transformer a la serie temporal sin intervenciones adicionales. Esto sirve como referencia para entender el comportamiento intrínseco de la serie sin ajustes.
- **2. Estandarización de Datos:** Para normalizar la escala y facilitar la comparación entre diferentes series, se optó por la estandarización de los datos. Este tratamiento ayuda a mitigar posibles problemas relacionados con las unidades y magnitudes de las diferentes series temporales.
- **3. Diferenciación:** Se recurrió a la diferenciación utilizando la función `ndiffs` de `mdarima.arima.utils`. Este enfoque es relevante para abordar problemas de estacionariedad, ya que la diferenciación puede ayudar a hacer que una serie no estacionaria sea estacionaria. Si Y_t representa la serie de tiempo entonces la diferencia de primer orden esta definida como $Z_t = Y_t - Y_{t-1}$.
- **4. Descomposición aditiva:** Se implementó una descomposición clásica en componentes de tendencia, estacional y residual, siguiendo la metodología introducida por Box-Jenkins. Este método proporciona una visión más detallada de la estructura subyacente de la serie temporal, se asume que la serie puede ser descompuesta de esta manera $Y_t = S_t + R_t + E_t$, donde Y_t es la serie de tiempo, S_t es la componente estacional, R_t es la componente de tendencia y E_t es la componente residual.
- **5. Descomposición multiplicativa:** Se exploró una variante en la descomposición, utilizando parámetros o metodologías distintas. Esto permitió evaluar la sensibilidad del W-transformer a diferentes enfoques de descomposición. En este caso se asumió que la serie puede ser descompuesta

de esta manera $Y_t = S_t * R_t * E_t$, donde Y_t es la serie de tiempo, S_t es la componente estacional, R_t es la componente de tendencia y E_t es la componente residual.

Repetición y Evaluación

Cada uno de los cinco tratamientos fue repetido en cinco ocasiones, cada vez con una semilla diferente. Esto se hizo con el propósito de obtener resultados más robustos y evitar la influencia de la aleatoriedad en las métricas de evaluación. Se calculó el promedio de las métricas para cada tratamiento y repetición.

Comparación entre Wavelets

El estudio se replicó utilizando dos wavelets diferentes: la clásica wavelet de Haar y la wavelet DB6. La segunda fue implementada desde el repositorio de GitHub [modwtpy]. Este enfoque permitió comparar y contrastar el desempeño del W-transformer utilizando diferentes bases wavelet.

Consideraciones y Soluciones

Durante la implementación, surgieron desafíos relacionados con las métricas y el preprocesamiento. Por ejemplo:

- **MAPE y Valores Negativos.** La métrica MAPE es intrínsecamente válida para valores no negativos. Ante la presencia de valores negativos, se implementó una solución que notificaba la imposibilidad de calcular la métrica, garantizando transparencia y precisión en los resultados.
- **Problemas con MASE en Series Estacionales.** Se observó que la métrica MASE en ocasiones no se podía calcular debido a problemas específicos con series temporales estacionales. Este aspecto fue documentado para su consideración en la interpretación de los resultados.

Comparación de Wavelets.

Se identificó que la wavelet DB6 ofrecía una mayor ondulación en las predicciones en comparación con la wavelet de Haar. Este fenómeno será explorado y discutido con mayor profundidad en la sección de resultados obtenidos.

Esta implementación exhaustiva y detallada busca proporcionar una base sólida para la evaluación y comparación del W-transformer en diversos contextos de series temporales para bases de datos de diferentes ámbitos, incluyendo series con comportamientos financieros, epidemiológicos, meteorológicos y demográficos. En la sección se muestran los resultados más destacados.

6. Resultados

A continuación, se proporciona un cuadro informativo para cada serie de datos con énfasis en los resultados destacados. Los tratamientos enumerados en las tablas incluyen: 1. W-Transformer directo, 2. Estandarización de datos y W-Transformer, 3. Diferenciación y W-Transformer, 4. Descomposición aditiva y W-Transformer, y 5. Descomposición multiplicativa y W-Transformer.

En el Cuadro 2, se destaca que el tratamiento 5, que implica la descomposición multiplicativa antes de la transformación con wavelets, presenta el MAE más bajo (1.369), indicando una mayor precisión en la predicción según todas las métricas evaluadas. Además, el tiempo de ejecución no difiere significativamente del mejor resultado obtenido con W-Transformer directo. Para esta serie temporal, la

6 RESULTADOS

wavelet DB6 muestra un rendimiento generalmente superior.

Es notable que al realizar la transformada wavelet DB6, las métricas no experimentan cambios sustanciales entre la aplicación directa y la escala de los datos. Esto sugiere que la descomposición multiplicativa podría ser un tratamiento preferido para aplicar en primer lugar en una serie caracterizada por no linealidad, no estacionariedad y dependencia a largo plazo.

Es relevante señalar que, a pesar de que los datos de la serie contienen valores negativos, estos no influyeron significativamente en los resultados de las métricas. La Figura 34 presenta el mejor pronóstico para la serie de Inflación del IPC, utilizando la wavelet DB6 y el tratamiento de descomposición.

Cuadro 2. Comparación de Implementaciones de Tratamientos para la Serie Inflación del IPC.

Wavelets	Tratamiento	Ejecución (s)	MAE	RMSE	MASE	sMAPE	MAPE
Haar	1	63.174	2.225	2.788	1.880	26.702	23.825
Haar	2	62.567	31.176	39.060	1.880	27.174	24.231
Haar	3	62.791	1.694	2.029	1.431	19.954	22.219
Haar	4	63.044	1.536	1.744	1.297	20.159	21.150
Haar	5	62.123	2.386	2.882	2.016	25.159	21.459
DB6	1	61.911	2.234	2.793	1.880	20.374	17.914
DB6	2	62.577	31.293	39.125	1.887	20.233	17.804
DB6	3	61.927	2.667	3.010	2.253	23.447	28.018
DB6	4	62.438	1.541	1.769	1.302	14.639	15.047
DB6	5	62.099	1.369	1.627	1.157	12.501	13.096

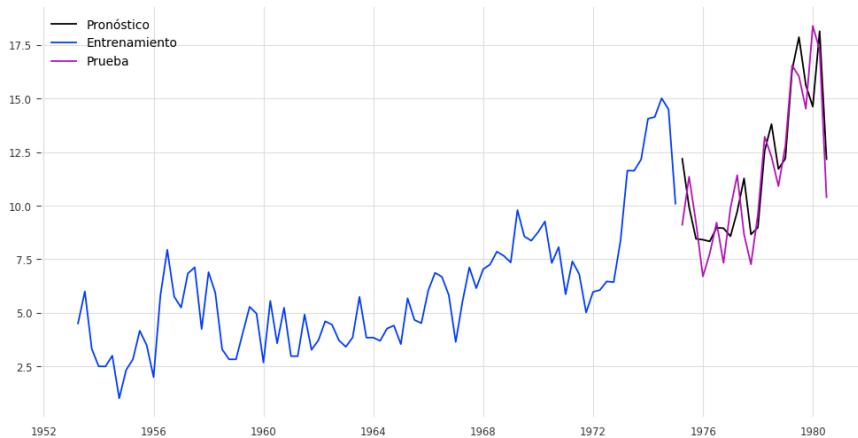


Figura 34. Mejor pronóstico para la serie Inflación del IPC..

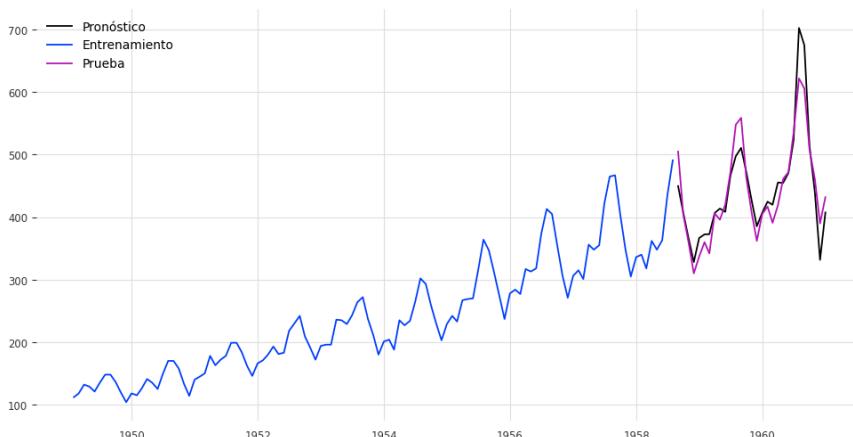
El Cuadro 3 compara diversas implementaciones de tratamientos para la serie temporal Airline Passengers, utilizando las wavelets Haar y DB6. Se destaca que el tratamiento 4 con wavelets Haar muestra una ejecución más rápida (63.087 segundos), pero presenta un MAE más elevado (30.873) y RMSE (49.267). En contraste, el tratamiento 4 con wavelets DB6 exhibe el menor MAE (24.029) y RMSE (32.147), indicando una mejor precisión en la predicción.

Cuadro 3. Comparación de Implementaciones de Tratamientos para la Serie Airline Passengers.

Wavelets	Trat.	Ejecución (s)	MAE	RMSE	MASE	sMAPE	MAPE
Haar	1	80.709	270.95	279.07	12.865	88.221	60.982
Haar	2	80.655	104859.33	108002.62	12.865	88.142	60.944
Haar	3	80.753	56.500	67.299	2.682	12.607	13.311
Haar	4	63.612	33.296	39.421	1.581	7.439	7.466
Haar	5	63.087	30.873	49.267	1.466	6.362	6.615
DB6	1	80.775	267.078	281.124	12.681	85.666	59.189
DB6	2	81.088	103359.047	108795.011	12.681	85.591	59.148
DB6	3	81.454	124.742	143.025	5.923	26.008	31.702
DB6	4	63.669	24.029	32.147	1.141	5.275	5.291
DB6	5	63.718	30.442	48.863	1.445	6.270	6.513

Es importante señalar que en esta serie temporal se observan valores notablemente altos para las métricas MAE y RMSE en el tratamiento de estandarización utilizando DB6. Sin embargo, para ambas wavelets, se observa un mejor rendimiento al aplicar las descomposiciones (tratamientos 4 y 5), lo cual es consistente con el comportamiento observado en la serie de Inflación. Se destaca que ambas series comparten características de no linealidad, no estacionariedad y dependencias a largo plazo.

En la Figura 35 se presenta el mejor pronóstico para la serie de Airline Passengers. Se observa que el pronóstico tiene una varianza mayor que la serie de prueba, lo que sugiere que se está propagando cierta incertidumbre en las predicciones.

**Figura 35.** Mejor pronóstico para la serie Airline passangers..

El Cuadro 4 de comparación de implementaciones de tratamientos para la serie temporal de la Reserva Federal de los Estados Unidos proporciona información valiosa sobre el rendimiento de distintas estrategias. El tratamiento 4 con wavelets DB6 destaca al exhibir el menor tiempo de ejecución (211.535 segundos), aunque no es muy diferente al resto, y métricas de precisión notoriamente bajas, como MAE (2.510), RMSE (3.436), MASE (1.503), sMAPE (1.961) y MAPE (1.968). Este tratamiento demuestra una predicción precisa y eficiente en comparación con otras opciones. Sin embargo, se observa un valor NaN en la métrica MASE para el tratamiento 5 con wavelets Haar, lo que sugiere la necesidad de una exploración adicional para comprender la causa de esta excepción.

6 RESULTADOS

Al igual que otras series de tiempo se tienen valores altos para las métricas de los tratamientos con estandarización. Además se observan métricas similares para los tratamientos con ambas descomposiciones, aditiva o multiplicativa.

En la Figura 36 se observa el mejor pronóstico para esta serie de tiempo sobre la reserva federal.

Cuadro 4. Comparación de Implementaciones de Tratamientos para la Serie Reserva federal USA.

Wavelets	Trat.	Ejecución (s)	MAE	RMSE	MASE	sMAPE	MAPE
Haar	1	213.982	57.381	57.956	34.373	55.813	43.553
Haar	2	213.584	5210.206	5262.423	34.373	55.598	43.422
Haar	3	213.46	4.213	6.151	2.524	3.275	3.403
Haar	4	211.597	2.654	3.494	1.590	2.072	2.076
Haar	5	211.535	3.554	4.150	NaN	2.738	2.792
DB6	1	215.510	35.553	36.662	21.297	31.087	26.756
DB6	2	215.438	3228.220	3328.945	21.297	30.980	26.676
DB6	3	215.873	5.520	7.317	3.306	4.299	4.404
DB6	4	213.057	2.510	3.436	1.503	1.961	1.968
DB6	5	216.391	3.669	4.283	2.198	2.791	2.846



Figura 36. Mejor pronóstico para la serie Reserva Federal..

El Cuadro 5, que compara distintas implementaciones de tratamientos para la serie temporal de Dengue en Colombia, revela aspectos de relevancia. El tratamiento 3, empleando wavelets Haar, destaca al lograr las métricas más bajas, como MAE (237.271), RMSE (314.607), MASE (3.155), sMAPE (13.402), y MAPE (15.811), lo que indica una predicción precisa y efectiva. En contraste, el tratamiento 2 con la misma wavelet exhibe resultados extremadamente altos, sugiriendo una influencia negativa específica de este enfoque. Además, el tratamiento 5 con wavelets DB6 presenta métricas notoriamente bajas, resaltando la eficacia de la descomposición multiplicativa en esta serie temporal.

Es relevante señalar que esta serie de tiempo no presenta dependencia a largo plazo, a diferencia de las anteriores, lo que hace que el uso de la wavelet Haar sea más apropiado para este tipo de series. En la gráfica 37, se observa la mejor predicción utilizando diferenciación de primer orden con la wavelet Haar.

Cuadro 5. Comparación de Implementaciones de Tratamientos para la Serie Dengue de Colombia.

Wavelets	Trat.	Ejecución (s)	MAE	RMSE	MASE	sMAPE	MAPE
Haar	1	433.023	1059.146	1106.451	14.082	75.999	54.483
Haar	2	435.407	3315127.089	3463190.798	14.082	75.997	54.482
Haar	3	437.812	237.271	314.607	3.155	13.402	15.811
Haar	4	385.254	334.863	413.812	4.452	16.647	16.492
Haar	5	383.458	416.865	517.568	5.542	20.561	20.395
DB6	1	435.449	982.201	1108.210	13.058	64.932	47.956
DB6	2	438.117	3074290.512	3468698.431	13.059	64.930	47.955
DB6	3	441.893	508.702	682.382	6.763	28.880	24.805
DB6	4	386.499	320.971	386.559	4.267	16.882	16.893
DB6	5	383.38	413.596	495.739	5.499	23.164	22.207

**Figura 37.** Mejor pronóstico para la serie Dengue de Colombia..

En el Cuadro 6 se presentan los resultados de diferentes tratamientos aplicados a la serie temporal de Dengue en Bangkok. El tratamiento 3, utilizando wavelets Haar, destaca por sus métricas significativamente bajas, con MAE (730.943), RMSE (1536.798), MASE (3.749), y sMAPE (56.602), indicando una predicción precisa y eficaz. Asimismo, el tratamiento 4 con la misma wavelet exhibe el RMSE más bajo (1027.258) y un destacado MASE (2.613), señalando una mayor precisión en la predicción y menor sesgo en comparación con otros tratamientos. Sin embargo, se observa que la wavelet DB6 muestra mejores métricas al combinarse con una descomposición aditiva; el tratamiento correspondiente logra la mejor MAE (495.229) y MASE (2.540). Es importante destacar que los tratamientos 1 y 2 con wavelets DB6 se consideran no válidos debido a métricas extremadamente altas.

Al examinar los pronósticos generados por estos tratamientos, se observa que la propagación de oscilaciones prominentes afecta el cálculo de las métricas al considerar valores fuera del rango de comportamiento esperado. La figura 38 ilustra el mejor pronóstico para la serie de Dengue en Bangkok (DB6 con tratamiento 4), destacando la eficacia del tratamiento 4 con wavelets Haar en esta configuración. Se destaca que, aunque el modelo pudo anticipar el punto máximo de contagio, no alcanzó los valores reales, sugiriendo una intrínseca complejidad en el proceso subyacente.

6 RESULTADOS

Cuadro 6. Comparación de Implementaciones de Tratamientos para la Serie Dengue de Bangkok.

Wavelets	Trat.	Ejecución (s)	MAE	RMSE	MASE	sMAPE	MAPE
Haar	1	119.858	904.009	1652.432	4.636	96.954	63.337
Haar	2	120.099	2015939.529	3684923.782	4.636	96.952	63.336
Haar	3	119.477	730.943	1536.798	3.749	56.602	48.057
Haar	4	99.715	509.447	1027.258	2.613	40.682	81.789
Haar	5	99.867	547.223	1024.288	2.807	47.316	93.927
DB6	1	121.591	915.597	1690.879	4.696	93.790	No válido
DB6	2	122.630	2041782.159	3770646.571	4.696	93.788	No Válido
DB6	3	122.800	858.040	1625.624	4.401	92.433	60.141
DB6	4	99.515	495.229	1027.489	2.540	40.756	77.387
DB6	5	99.278	537.530	955.113	2.757	45.353	97.081

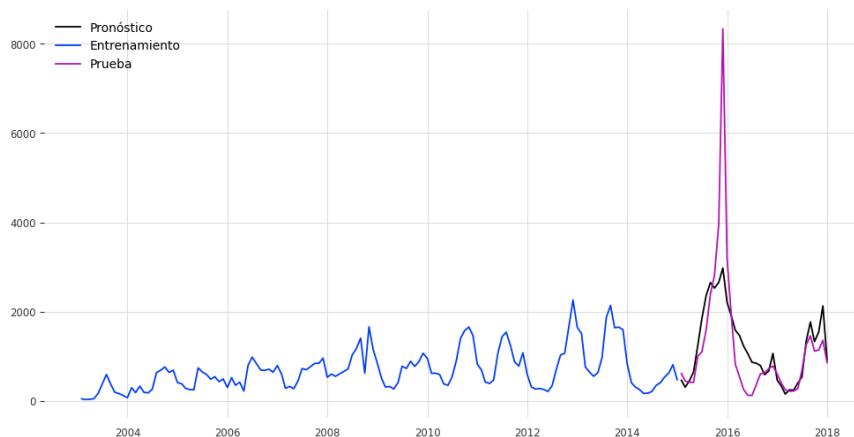
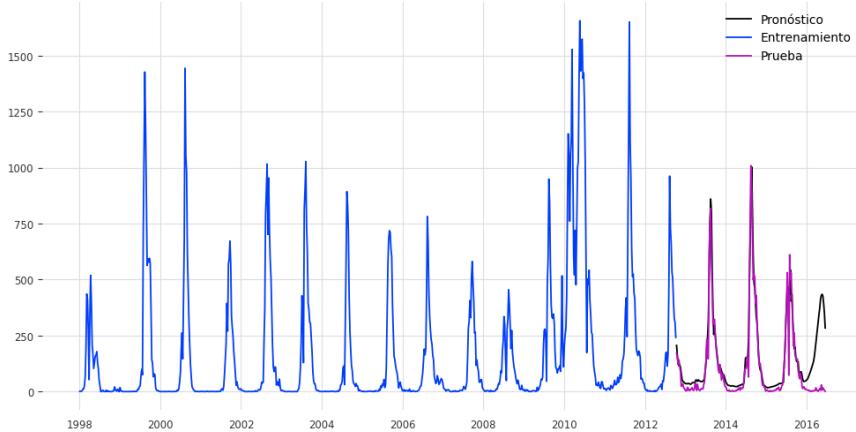


Figura 38. Mejor pronóstico para la serie Dengue de Bangkok..

En el cuadro 7, se presenta la comparación de implementaciones de tratamientos para la serie temporal de Influenza en Japón. Los resultados revelan que el tratamiento 3, utilizando wavelets Haar, logra métricas notoriamente bajas, destacando con MAE (58.654), RMSE (106.680), MASE (1.112), y sMAPE (83.176), indicando una predicción precisa y eficaz. Por otro lado, el tratamiento 5 con la misma wavelet exhibe un elevado sMAPE (180.235), sugiriendo un rendimiento menos óptimo en términos de precisión. Los tratamientos 1 y 2 con wavelets Haar, así como los tratamientos 1, 2 y 3 con wavelets DB6, se consideran no válidos debido a métricas extremadamente altas, y la figura 39 muestra el mejor pronóstico para la serie de Influenza en Japón. A pesar de las características de la serie, como su naturaleza no lineal, estacionaria, y sin dependencia a largo plazo, el mejor modelo logra estimar con precisión los picos oscilatorios, aunque la imposibilidad de calcular la métrica MAPE sugiere que los pronósticos negativos no capturan adecuadamente la naturaleza infecciosa de la serie (hubo valores negativos que imposibilitaron calcular la métrica).

Cuadro 7. Comparación de Implementaciones de Tratamientos para la Serie Influeza de Japón.

Wavelets	Trat.	Ejecución (s)	MAE	RMSE	MASE	sMAPE	MAPE
Haar	1	649.803	58.654	106.681	1.112	83.176	No Válido
Haar	2	652.113	97131.711	176663.185	1.112	83.176	No Válido
Haar	3	655.122	58.654	106.680	1.112	83.176	No Válido
Haar	4	622.63	116.147	143.596	2.202	136.428	665.088
Haar	5	620.570	78.660	153.475	1.491	99.546	180.235
DB6	1	662.891	236.194	299.911	4.478	151.817	No válido
DB6	2	660.223	391136.705	496652.888	4.478	151.817	No Válido
DB6	3	652.579	236.194	299.911	4.478	151.817	No Válido
DB6	4	618.556	128.524	161.210	2.437	133.003	No válido
DB6	5	626.014	113.668	200.070	2.155	130.84	No Válido

**Figura 39.** Mejor pronóstico para la serie Influenza Japón..

En el cuadro 8, se presenta la comparación de implementaciones de tratamientos para la serie temporal de los precios de las acciones de Netflix. El tratamiento 4 con la misma wavelet DB6 presenta el MAE más bajo (8.660) y el MASE más destacado (1.015), sugiriendo una mayor precisión en la predicción y menor sesgo en comparación con otros tratamientos. Sin embargo, el tratamiento 5 con wavelets DB6 exhibe el RMSE más bajo (14.958), indicando una menor dispersión de los errores. Aunque las características de la serie, como su naturaleza no lineal, no estacionaria, sin tendencia estacionaria y con dependencia a largo plazo, presentan desafíos, las implementaciones, como se muestra en la figura 40, logran resultados notables al aplicar descomposiciones, se muestra el tratamiento DB6 con descomposición aditiva. Este enfoque demuestra su eficacia incluso en series con comportamientos complejos y caóticos, subrayando la utilidad de las técnicas de procesamiento previo para mejorar la precisión de los pronósticos.

Nuevamente como antes la métrica MAPE no pudo ser calculada correctamente porque algunas predicciones salieron negativas y para calcularla se requiere tomar valores estrictamente positivos del pronóstico vs el test.

6 RESULTADOS

Cuadro 8. Comparación de Implementaciones de Tratamientos para la Serie Netflix stock price.

Wavelets	Trat.	Ejecución (s)	MAE	RMSE	MASE	sMAPE	MAPE
Haar	1	147.410	102.320	119.038	11.999	36.806	50.451
Haar	2	147.610	36904.903	42934.652	11.998	36.684	50.216
Haar	3	146.990	46.140	55.372	3.475	94.538	No Válido
Haar	4	146.157	9.824	15.195	1.152	4.023	4.066
Haar	5	145.527	11.779	15.348	1.381	4.472	4.611
DB6	1	145.456	102.804	125.589	12.055	36.817	51.408
DB6	2	144.962	37079.237	45297.529	12.055	36.695	51.168
DB6	3	144.366	112.676	143.260	8.486	109.199	No válido
DB6	4	146.248	8.660	14.972	1.015	3.362	3.424
DB6	5	146.240	11.314	14.958	1.327	4.297	4.427

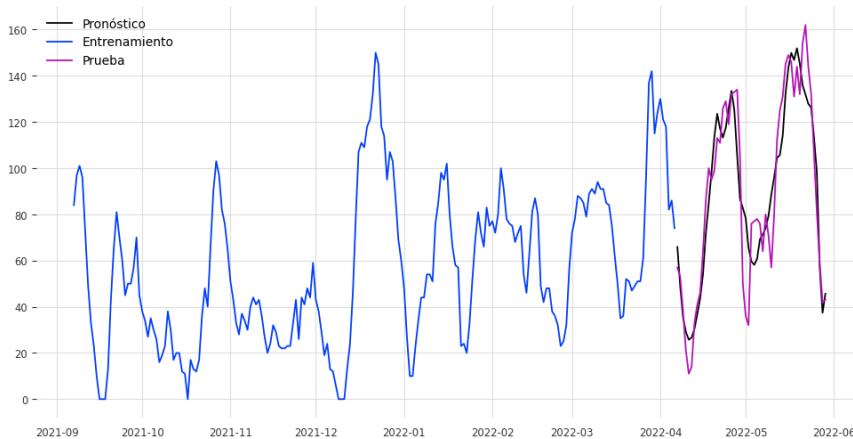


Figura 40. Mejor pronóstico para la serie Netflix Stock Prices..

En el Cuadro 9 presenta la comparación de implementaciones de tratamientos para la serie temporal diaria de manchas solares (Sunspots Daily). Se destaca que el tratamiento 4 con wavelets DB6 muestra un rendimiento destacado, logrando el MAE más bajo (12.005) y el MASE más bajo (1.244), indicando una predicción precisa con menor dispersión de errores. Además, la figura 41 ilustra el mejor pronóstico para esta serie, resaltando la eficacia de este tratamiento en particular. Otra buena predicción está dada por el tratamiento 4 (descomposición aditiva) con la otra wavelet de Haar. Tiene 3 mejores valores de métricas RMASE, sMAPE y MAPE.

Cuadro 9. Comparación de Implementaciones de Tratamientos para la Serie Sunspots Daily.

Wavelets	Trat.	Ejecución (s)	MAE	RMSE	MASE	sMAPE	MAPE
Haar	1	166.879	36.7310	41.905	3.807	54.447	No Válido
Haar	2	167.051	5509.653	6285.732	3.807	54.447	No Válido
Haar	3	166.127	18.481	21.246	1.915	25.822	32.799
Haar	4	165.289	12.560	15.499	1.302	17.609	20.971
Haar	5	165.207	16.251	19.264	1.684	21.486	25.077
DB6	1	165.605	51.818	59.283	5.370	69.889	69.307
DB6	2	164.379	7772.758	8892.515	5.370	69.882	69.294
DB6	3	166.768	39.809	45.791	4.126	50.337	63.716
DB6	4	166.420	12.005	15.517	1.244	17.690	22.708
DB6	5	167.540	14.004	18.462	1.451	19.162	24.875

**Figura 41.** Mejor pronóstico para la serie Sunspots Daily..

Es interesante observar que, a diferencia de otras series, en este caso, el tratamiento 4 con wavelets Haar supera a las implementaciones con wavelets DB6. La serie Sunspots Daily muestra un comportamiento que, a pesar de tener características lineales, estacionarias y sin tendencia, sigue siendo desafiante y caótica en su apariencia, lo que resalta la importancia de la selección adecuada de tratamientos y wavelets para mejorar la precisión de los pronósticos. Además, se evidencia que la calidad de la predicción mejora significativamente con un mayor volumen de datos de entrenamiento, como se aprecia en la figura 41, donde se muestra la predicción utilizando solo los últimos datos para una visualización más clara.

Los resultados de la implementación revelan que la combinación más efectiva para mejorar la precisión en diversas series temporales, como la inflación del IPC, Airline Passengers, Reserva Federal USA, Dengue en Colombia y Bangkok, Influenza en Japón, Netflix Stock Prices, y Sunspots Daily, es la aplicación de descomposiciones aditivas o multiplicativas junto con la transformación W-Transformer, e incluso en algunos casos es suficiente con realizar diferenciación de un orden. Para la inflación del IPC, Airline Passengers, Reserva Federal USA, Netflix Stock Prices, y Sunspots Daily, la descomposición aditiva con W-Transformer utilizando wavelet DB6 (Tratamiento 4) ha demostrado ser la opción más exitosa. Por otro lado, para Dengue en Colombia, Dengue en Bangkok, e Influenza en Japón, la diferenciación de primer orden con W-Transformer y wavelet Haar (Tratamiento 3) ha proporcionado los mejores resultados. Estos hallazgos subrayan la importancia de seleccionar cuidadosamente la combinación de wavelets y tratamientos específicos para optimizar las métricas de evaluación del

pronóstico en cada conjunto de datos.

También es importante validar la estructura caótica que presentan alguna de las series presentadas, el pronóstico puede mejorarse adaptándose a las características específicas de cada conjunto de datos. Para contextualizar, la inflación del IPC trimestral muestra ser no lineal, no estacionaria, sin tendencia estacionaria y con dependencia a largo plazo. En contraste, la serie temporal de Airline, Reserva Federal USA, Netflix Stock Prices y Sunspots Daily presentan patrones similares de no linealidad, no estacionariedad y dependencia a largo plazo, pero con variaciones en la presencia de tendencia estacionaria. Por otro lado, la serie de Dengue en Colombia exhibe características no lineales y no estacionarias sin tendencia estacionaria y sin dependencia a largo plazo. Dengue en Bangkok y la Influenza en Japón muestran patrones diversos, con Dengue en Bangkok siendo no lineal pero estacionaria, sin tendencia estacionaria y sin dependencia a largo plazo, mientras que la Influenza en Japón se presenta como no lineal, estacionaria, con tendencia estacionaria y sin dependencia a largo plazo. Estos resultados subrayan la importancia de adaptar enfoques específicos de tratamiento a las particularidades de cada serie temporal.

7. Conclusión

Los datos de series temporales constituyen un tipo especial de información secuencial en la que cada ejemplo está vinculado a una dimensión temporal. Aunque la adaptación de Transformers a series temporales es teóricamente directa, en la práctica se enfrenta a desafíos debido a la longitud extendida de estas secuencias. Este inconveniente surge porque la complejidad de la autoatención aumenta cuadráticamente en relación con la longitud de la secuencia de entrada.

En este contexto, el W-transformer se destaca como una alternativa eficaz para abordar secuencias temporales extensas, logrando resultados de predicción satisfactorios.

Referencias

- [1] Sasal L., Chakraborty T. y Hadid A. (2022) W-Transformers: A Wavelet-based Transformer Framework for Univariate Time Series Forecasting, *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, 671–676.
- [2] Aminghafari M. y Poggi J. M. (2007) Forecasting Time Series Using Wavelets, *International Journal of Wavelets, Multiresolution and Information Processing*, 5(5) 709–724.
- [3] Hewamalage, H., Bergmeir, C. y Bandara, K. (2021) Recurrent Neural Networks for Time Series Forecasting: Current status and future directions, *International Institute of Forecasters. Elsevier*, 37(1) 388–427.
- [4] Percival, D. B. y Walden, A. T. (2021) Wavelet Methods for Time Series Analysis, *Cambridge University Press*.
- [5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017), Attention is all you need, *CoRR*, vol.abs/1706.03762.
- [6] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun y Rong Jin (2022), FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting, *In International Conference on Machine Learning* (pp. 27268-27286). PMLR.
- [7] Shumway, Robert H. & Stoffer David S. (2017). *Time Series Analysis and Its Applications with R examples*. [4th. ed, Springer]
- [8] B. L. Bowerman, & D. S. Stoffer, &R. T. O'Connell, & A. B. Koehler (2005). *Forecasting, time series and regression. An applied aproach*. [4th. ed, Thomson. Brooks Cole].
- [9] Dickey, D.A. (1976). *Estimation and hypothesis testing in nonstationary time series*, Ph.D. dissertation (Iowa State University, Ames, IA).
- [10] Newbold P, Bos T (1985). *Stochastic parameter regression models [Datos de inflacion rate]*. Sage, Beverly Hills.
- [11] Qingsong Wen and Tian Zhou and Chaoli Zhang and Weiqi Chen and Ziqing Ma and Junchi Yan and Liang Sun (2023). *Transformers in Time Series: A Survey*, eprint 2202.07125.
- [12] Panja, M., Chakraborty, T., Kumar, U., & Liu, N. (2022). *Epicasting: An ensemble wavelet neural network (ewnet) for forecasting epidemics*. arXiv preprint arXiv:2206.10696.
- [13] Polwiang S. (2020) *The time series seasonal patterns of dengue fever and associated weather variables in Bangkok (2003-2017)*. BMC Infect Dis. 2020 Mar 12;20(1):208. doi: 10.1186/s12879-020-4902-6. PMID: 32164548; PMCID: PMC7068876.
- [14] Zhang, W., Song, X., Wu, H. et al. (2020). *Epidemiology, species distribution, and predictive factors for mortality of candidemia in adult surgical patients*. BMC Infect Dis 20, 506. <https://doi.org/10.1186/s12879-020-05238-6>
- [15] Denis Kwiatkowski, Peter C.B. Phillips, Peter Schmidt, Yongcheol Shin. (1992). *Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?*, Journal of Econometrics, Volume 54, Issues 1–3, 1992, Pages 159-178, ISSN 0304-4076.
- [16] Simonsen, I., Hansen, A., & Nes, O. M. (1998). *Determination of the Hurst exponent by use of wavelet transforms*. Physical Review E, 58(3), 2779.
- [17] Godahewa, R., Bergmeir, C., Webb, G. I., Hyndman, R. J., & Montero-Manso, P. (2021). *Monash time series forecasting archive*. arXiv preprint arXiv:2105.06643.