



UNIVERSIDAD
DE LA REPUBLICA
URUGUAY



Aprendizaje por transferencia en contexto de dependencia de datos

Aplicación en series de tiempo

Federica Eva Silberberg Gadea

Programa de Posgrado de Maestría en Ciencia de Datos y Aprendizaje
Automático Maestría en Ciencia de Datos y Aprendizaje Automático
Facultad de Ingeniería
Universidad de la República

Montevideo – Uruguay
Noviembre de 2025



UNIVERSIDAD
DE LA REPUBLICA
URUGUAY



Aprendizaje por transferencia en contexto de dependencia de datos

Aplicación en series de tiempo

Federica Eva Silberberg Gadea

Tesis de Maestría presentada al Programa de Posgrado de Maestría en Ciencia de Datos y Aprendizaje Automático Maestría en Ciencia de Datos y Aprendizaje Automático, Facultad de Ingeniería de la Universidad de la República, como parte de los requisitos necesarios para la obtención del título de Magíster en Maestría en Ciencia de Datos y Aprendizaje Automático.

Director:

Ph.D. Prof. Mathias Bourel

Director académico:

Ph.D. Prof. Ignacio Ramírez

Montevideo – Uruguay

Noviembre de 2025

Silberberg Gadea, Federica Eva

Aprendizaje por transferencia en contexto de dependencia de datos / Federica Eva Silberberg Gadea. - Montevideo: Universidad de la República, Facultad de Ingeniería, 2025.

VI, 126 p. 29, 7cm.

Director:

Mathias Bourel

Director académico:

Ignacio Ramírez

Tesis de Maestría – Universidad de la República, Programa de Maestría en Ciencia de Datos y Aprendizaje Automático Maestría en Ciencia de Datos y Aprendizaje Automático, 2025.

Referencias bibliográficas: p. 106 – 114.

1. 1ra palabra clave. I. Bourel, Mathias, .
II. Universidad de la República, Programa de Posgrado de Maestría en Ciencia de Datos y Aprendizaje Automático Maestría en Ciencia de Datos y Aprendizaje Automático. III. Título.

INTEGRANTES DEL TRIBUNAL DE DEFENSA DE TESIS

Ph.D. Prof. Kalemkerian

Ph.D. Prof. Marroig

Ph.D. Prof. Cugliari

Montevideo – Uruguay
Noviembre de 2025

Tabla de contenidos

Introducción	1
Contribuciones	3
Organización del documento	4
1 Aprendizaje por Transferencia (Transfer learning)	5
1.1 Fundamentos teóricos	7
1.2 Categorización de técnicas de aprendizaje por transferencia . . .	12
1.2.1 Categorización del problema	13
1.2.2 Categorización de la solución	15
1.3 Aprendizaje negativo (Negative transfer)	18
1.4 Trabajos relacionados	22
1.4.1 Series de tiempo	23
2 Aprendizaje por Transferencia en Series de tiempo	26
2.1 Series de Tiempo	26
2.1.1 Componentes de una serie temporal	27
2.1.2 Categorización de una serie temporal	30
2.1.3 Propiedades de la dependencia temporal	33
2.1.4 Evaluación de estacionariedad	38
2.1.5 Procesos autoregresivos (AR)	39
2.1.6 Procesos de media móvil y ARMA(p,q)	42
2.2 Simulaciones	44
2.2.1 Metodología	45
2.2.2 Similitud fuente-objetivo	50
2.2.3 Resultados	53
2.2.4 Conclusión	63

3	Aplicación: Índice de Precios al Consumo en Uruguay	65
3.1	Motivación	65
3.2	Presentación de los datos	67
3.3	Caracterización de la serie de tiempo	70
3.4	Evaluación de estacionariedad	74
3.5	Evaluación de modelos	79
3.6	Interior como conjunto objetivo	81
3.7	Conclusión	84
4	LLM para Series de Tiempo	86
4.1	Introducción	86
4.2	Fundamentos de los modelos de lenguaje	88
4.3	Desafíos y motivación	90
4.4	Trabajos relacionados	93
4.5	Aplicación: LLM para el Índice de Precios al Consumo en Uruguay	97
5	Conclusiones y trabajo futuro	102
	Referencias bibliográficas	106
	Apéndices	115
	Medidas de similitud	116
	Tests de estacionariedad	123
	Prompts - Aplicación: LLM para el IPC en Uruguay	124

Introducción

Como seres humanos, al enfrentarnos a una nueva tarea, rara vez partimos desde cero en términos de aprendizaje. En lugar de ello, utilizamos nuestros conocimientos y experiencias adquiridos en el pasado de forma inherente para adaptarnos a los nuevos desafíos que se nos presentan. Esta capacidad de transferir y aplicar conocimientos previos, que es intrínseca a la condición humana, es fundamental para nuestro proceso de aprendizaje. Esto se puede dar siempre y cuando haya una conexión entre las dos tareas. Por ejemplo, una persona que aprende a tocar el violín posiblemente pueda aprender a tocar el piano más rápido, dado que ambas actividades implican instrumentos musicales y es posible que haya conocimiento compartido. Por el contrario, es probable que aprender a andar en bicicleta no genere un aprendizaje más rápido de instrumentos musicales como el piano (Zhuang et al. [2020](#)).

La transferencia de conocimiento fue estudiada por primera vez en el campo de la psicología educativa para mejorar la habilidad de las personas para aprender tareas nuevas y resolver problemas (Z. Chen y Daehler, [1989](#)). En este contexto, resulta interesante establecer un paralelismo entre nuestra forma de aprender nuevas tareas y el proceso de aprendizaje de los modelos de aprendizaje automático.

Cuando los datos provienen de distintas distribuciones o difieren en su espacio de características, los modelos deben entrenarse desde cero, implicando la necesidad de tener una gran cantidad de datos etiquetados (en caso de estar trabajando bajo aprendizaje supervisado), resultando costoso y hasta imposible en muchos casos. Aun cuando es posible llevar a cabo el proceso de etiquetado, debido a su alto costo, el número de instancias etiquetadas puede resultar insuficiente para asegurar que el modelo funcione de acuerdo a los estándares esperados. Incluso los datos disponibles pueden ser escasos de por sí debido a la naturaleza del dominio, generando que el modelo entrenado en los mismos arroje predicciones pobres como consecuencia de la insuficiencia

de información provista durante el proceso de entrenamiento. Esto último es común en ámbitos como la medicina, donde la baja ocurrencia de cierto tipo de fenómenos, como algunas enfermedades poco comunes, limita la disponibilidad de datos para entrenar un modelo desde cero.

Por otra parte, para aquellos problemas que son altamente dependientes de cambios contextuales, los datos pueden quedar obsoletos fácilmente. Un problema común en la minería de datos, específicamente en el análisis de tráfico de redes de alta velocidad para encontrar eventos anómalos, es la aplicación de modelos sobre datos que se extraen de entornos en constante cambio, en donde se vuelve crucial la actualización del mismo para mantener su precisión en la predicción del flujo actual de datos (Yang y Wu, 2006).

En estos escenarios, se vuelve crítico contar con técnicas de aprendizaje automático que logren retener y reutilizar el conocimiento adquirido en el pasado para resolver nuevos problemas, tema central que fue discutido en el workshop posterior a la conferencia *Neural Information Processing Systems* de 1995 (NIPS-95), denominado "*Learning to Learn*" (Pan y Yang, 2009). Esta es la principal motivación del aprendizaje por transferencia: poder hacer un pasaje de conocimiento de un dominio para el cual tenemos datos etiquetados a otro dominio, con una distribución o espacio de características diferente, para el cual no contamos con suficientes datos etiquetados.

Acompañando las limitaciones mencionadas, muchos modelos estadísticos clásicos asumen que las observaciones son independientes e idénticamente distribuidas (*i.i.d.*), una hipótesis que rara vez se sostiene en contextos del mundo real. Diversos fenómenos complejos, como la evolución de los mercados financieros, los flujos de movilidad urbana o la dinámica de indicadores macroeconómicos, se modelan mediante *series de tiempo*, donde la dependencia entre observaciones sucesivas constituye un aspecto central.

Si bien existen trabajos que han explorado el aprendizaje por transferencia en escenarios simples, como la regresión lineal (Obst et al. 2022), y en enfoques más complejos, como las redes neuronales profundas (Ye y Dai (2021), F. Yu et al. (2022), (Malhotra et al. 2017)), estos estudios presentan limitaciones importantes. En los modelos lineales, se suele dejar de lado la estructura temporal de los datos, mientras que en los modelos neuronales, aunque se han aplicado técnicas de transferencia con cierto éxito, la falta de interpretabilidad impide comprender con claridad en qué condiciones la transferencia resulta efectiva o no.

Contribuciones

Ignorar la dependencia entre los datos puede conducir a modelos mal especificados y predicciones poco confiables, lo cual subraya la necesidad de enfoques alternativos que consideren explícitamente la dinámica subyacente en las series de tiempo. A pesar de la relevancia crítica de este problema, la aplicación del aprendizaje por transferencia en datos con dependencia ha sido escasamente explorada en la literatura. Esta brecha representa una oportunidad significativa para avanzar en el campo, y el presente estudio busca ser un aporte en este sentido.

En el presente documento presentamos y analizamos aplicaciones de transfer learning en el contexto de series de tiempo, combinando enfoques clásicos con modelos de estado del arte, entre ellos los grandes modelos de lenguaje (LLM). Nuestro objetivo central es identificar bajo qué condiciones la transferencia de conocimiento entre series resulta efectiva. Para ello, proponemos una metodología propia de transferencia (descrita en la sección 2.2.1.2), inspirada en el trabajo de Obst et al. (2022) originalmente planteado para modelos lineales. A través de simulaciones con modelos clásicos, exploramos el comportamiento del enfoque propuesto y evaluamos su desempeño considerando distintas medidas de distancia entre series, características estadísticas y tamaños de muestra del conjunto objetivo, con el propósito de determinar los factores que favorecen o limitan la efectividad del transfer learning.

Posteriormente, aplicamos esta metodología a datos reales del Índice de Precios al Consumo (IPC) de Uruguay, con el objetivo de validar su utilidad en un entorno empírico y analizar su capacidad para capturar patrones de transferencia presentes en series económicas reales. Complementamos este análisis con una interpretación económica de los resultados, buscando vincular los patrones observados en la transferencia con fenómenos macroeconómicos relevantes que ayuden a contextualizar y explicar las relaciones detectadas entre las series.

Por último, estudiamos la predicción de series de tiempo utilizando LLM, realizando pruebas sobre los mismos datos del IPC. En este caso, en lugar de analizar las condiciones bajo las cuales conviene transferir, como se hace con el modelo clásico, evaluamos la eficacia de diferentes técnicas de predicción, con el fin de comparar sus capacidades y limitaciones frente a los enfoques tradicionales.

Organización del documento

El presente trabajo se estructura de la siguiente manera. En el capítulo 1 se presentan los fundamentos teóricos del aprendizaje por transferencia, acompañados de una revisión de la literatura y de estudios previos relevantes en la temática. En el capítulo 2 se desarrollan los conceptos teóricos esenciales de series temporales, que sientan las bases para las aplicaciones posteriores. Asimismo, en este capítulo se experimenta con técnicas de aprendizaje por transferencia aplicadas a datos simulados mediante modelos clásicos de series temporales. El capítulo 3 está dedicado a la aplicación de estas metodologías en datos reales, utilizando el IPC de Uruguay como caso de estudio. Posteriormente, en el capítulo 4 se examina el estado del arte en torno a las aplicaciones de los grandes modelos de lenguaje (LLM) y se exploran algunas de sus implementaciones sobre los mismos datos del IPC. Finalmente, en el capítulo 5 se presentan las conclusiones y las perspectivas de trabajo futuro, donde se sintetizan los principales hallazgos y se delinean posibles líneas de investigación a desarrollar.

Capítulo 1

Aprendizaje por Transferencia (Transfer learning)

El aprendizaje por transferencia es una técnica de aprendizaje automático que ha captado la atención de investigadores de diversas disciplinas debido a su capacidad para mejorar el entrenamiento de modelos en dominios con datos limitados, aprovechando el conocimiento adquirido en tareas previas. Su origen puede rastrearse hasta finales de los años 70 y principios de los 80, cuando los investigadores comenzaron a explorar cómo compartir conocimiento entre diferentes tareas de aprendizaje (Bozinovski, 2020).

Uno de los pioneros claves en esta área fue Mitchell (1980), cuyo trabajo innovador sobre los espacios de versión y la transferencia inductiva, que definiremos más adelante, sentó las bases del aprendizaje por transferencia. Su artículo "*The Need for Biases in Learning Generalizations*" destacó cómo los modelos de aprendizaje pueden beneficiarse de reutilizar el conocimiento adquirido en experiencias previas para enfrentar nuevos problemas. En esa misma línea, Shavlik y Towell (1989) propusieron las denominadas *reglas de transferencia*, orientadas a facilitar el traspaso de conocimiento entre tareas relacionadas. Estas iniciativas sentaron las bases para el desarrollo posterior de técnicas más sofisticadas de aprendizaje por transferencia.

Con el auge del aprendizaje profundo en los años 2000, favorecido por los avances en hardware (Oh y Jung, 2004), el aprendizaje por transferencia recobró protagonismo. Las redes neuronales profundas, que son la base de estos modelos, demostraron una notable capacidad para aprender representaciones jerárquicas, especialmente en el ámbito de la visión por computadora. Dichas

representaciones organizan la información en distintos niveles de abstracción: las capas iniciales capturan características simples, mientras que las más profundas combinan esas características para formar representaciones más complejas. Como ejemplo, en imágenes, las primeras capas pueden identificar bordes o patrones locales, y las posteriores combinarlos en formas u objetos completos. Estrategias como el *fine-tuning* (la adaptación de un modelo preentrenado a una nueva tarea) y el uso de modelos entrenados en grandes conjuntos de datos, como ImageNet (Deng et al. 2009), evidenciaron el enorme potencial del aprendizaje por transferencia para resolver problemas complejos en este campo.

Esta revitalización alcanzó también su expansión en el procesamiento del lenguaje natural (Mou et al. 2016). Los modelos de lenguaje preentrenados, basados en arquitecturas Transformer (Vaswani et al. 2017), han mostrado gran capacidad para resolver múltiples tareas lingüísticas (Zhao et al. 2023). Entre ellos, los grandes modelos de lenguaje como GPT (Radford et al. 2018) destacan por su eficacia para transferir conocimiento con pocos datos, consolidando este enfoque como clave en contextos con información limitada. Sus aplicaciones se analizan en el capítulo 4.

El avance del aprendizaje por transferencia en visión por computadora y procesamiento de lenguaje natural, ha motivado la exploración de su aplicabilidad en otros contextos, como las series de tiempo. Pese a los desafíos particulares de este tipo de datos, principalmente la dependencia temporal, ha comenzado a emerger una línea de investigación orientada a adaptar y transferir conocimiento en tareas de esta índole. En esta tesis se investiga cómo implementar estrategias de aprendizaje por transferencia en problemas de series de tiempo, tanto mediante modelos clásicos como a través de aproximaciones más recientes basadas en LLM.

En el presente capítulo, se introducen los conceptos teóricos clave que servirán como base a lo largo del trabajo. En primer lugar, se revisarán los fundamentos teóricos del aprendizaje por transferencia. Luego, se presentará una categorización de estas técnicas, distinguiéndolas según se basen en el tipo de problema o en la solución, con el objetivo de definir y unificar los conceptos utilizados a lo largo del trabajo. Finalmente, se analiza la literatura relacionada, con el propósito de contextualizar la investigación, identificar sus aportes y señalar los vacíos existentes en el estado actual de la literatura.

1.1. Fundamentos teóricos

El objetivo de esta sección es introducir detalladamente los conceptos fundamentales del aprendizaje por transferencia, que nos van a ser útiles como base teórica para las metodologías y experimentos desarrollados a lo largo de este trabajo. Para ello, nos basamos en el artículo *A Survey on Transfer Learning* presentado por Pan y Yang (2009), que es un trabajo referente en la literatura sobre este tema. En particular, utilizamos las notaciones y definiciones establecidas allí, ya que proporcionan un marco formal que facilita la comprensión y aplicación de los principios del aprendizaje por transferencia en diferentes contextos. Se asumirá que el lector posee un conocimiento básico en teoría de conjuntos, cálculo, probabilidad y estadística.

Definimos primero el concepto de Inteligencia Artificial (Artificial Intelligence o AI, por su sigla en inglés) como el campo general de estudio que busca desarrollar sistemas capaces de realizar tareas que, si fueran llevadas a cabo por humanos, requerirían inteligencia (Kurzweil et al. 1990). Dentro de este campo se encuentra el aprendizaje automático (*machine learning*), que comprende un conjunto de procedimientos computacionales automáticos basados en operaciones lógicas o binarias, capaces de aprender una tarea a partir de ejemplos (Michie et al. 1995). A su vez, dentro del aprendizaje automático se encuentra el subcampo del aprendizaje profundo (*deep learning*), el cual se basa en redes neuronales profundas (DNN, por su sigla en inglés) y en el que es cada vez más común aplicar técnicas de transferencia de conocimiento. Esta dependencia entre los conceptos se representa en la figura 1.1.

En el paradigma tradicional del aprendizaje automático, cada tarea se aborda de manera independiente, entrenando un modelo desde cero sin reutilizar información obtenida en problemas anteriores. Este enfoque requiere, en general, grandes volúmenes de datos y tiempos de entrenamiento prolongados, ya que el modelo debe aprender desde cero cada nueva tarea. En contraste, el aprendizaje por transferencia aprovecha el conocimiento adquirido en una tarea previa para mejorar el rendimiento en una nueva tarea, especialmente cuando ambas comparten cierta estructura o características. Al reutilizar información, se reduce la cantidad de datos necesarios, se acelera el entrenamiento y, en muchos casos, se mejora la precisión del modelo. En modelos de aprendizaje profundo, cuyo entrenamiento desde cero puede requerir días o incluso semanas en hardware especializado como unidades de procesamiento

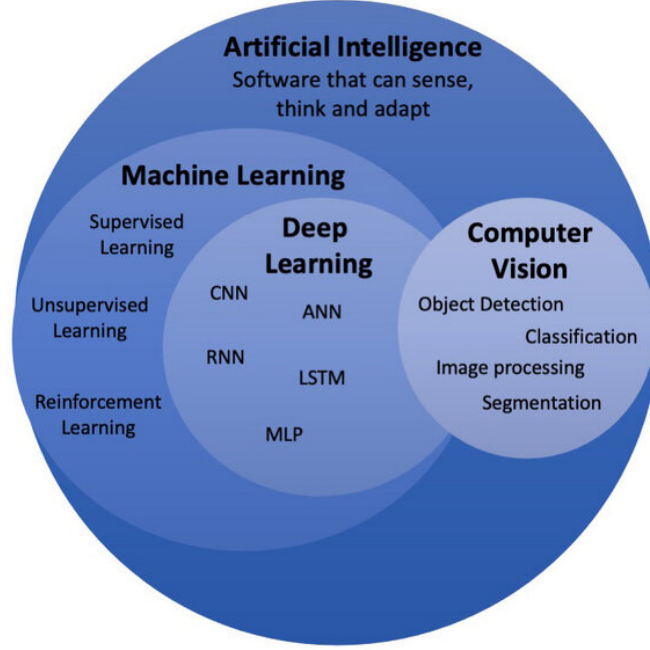


Figura 1.1: Inteligencia artificial y sus subcampos. Figura tomada del artículo *Leveraging Artificial Intelligence for Enhanced Seed Testing: Current trends and future prospects* (Peterson y Van Duijn, 2024)

gráfico (GPUs) o unidades de procesamiento tensorial (TPUs), este tipo de técnicas permite reducir drásticamente dichos tiempos al reutilizar representaciones previamente aprendidas. Esto no solo acelera el desarrollo de modelos más eficientes, sino que también hace posible su implementación en entornos con recursos limitados, donde entrenar modelos desde cero sería inviable.

La figura 1.2 muestra esta comparación entre el enfoque tradicional del aprendizaje automático y el aprendizaje por transferencia.

Comenzaremos definiendo entonces los conceptos relevantes para tratar el tema. La primera definición importante es la de *dominio*, que denotaremos como D en su forma genérica. Un dominio es un conjunto específico de datos que está compuesto por un espacio de características y una distribución marginal de probabilidad, que notamos como $\mathcal{X}$ y $P(X)$, respectivamente (Pan y Yang, 2009).

Definición 1.1.1 (Dominio). Sea $\mathcal{X}$ un espacio de probabilidad en $\mathbb{R}^n$ y $P(X)$ una distribución marginal de probabilidad. Definimos el dominio D como el

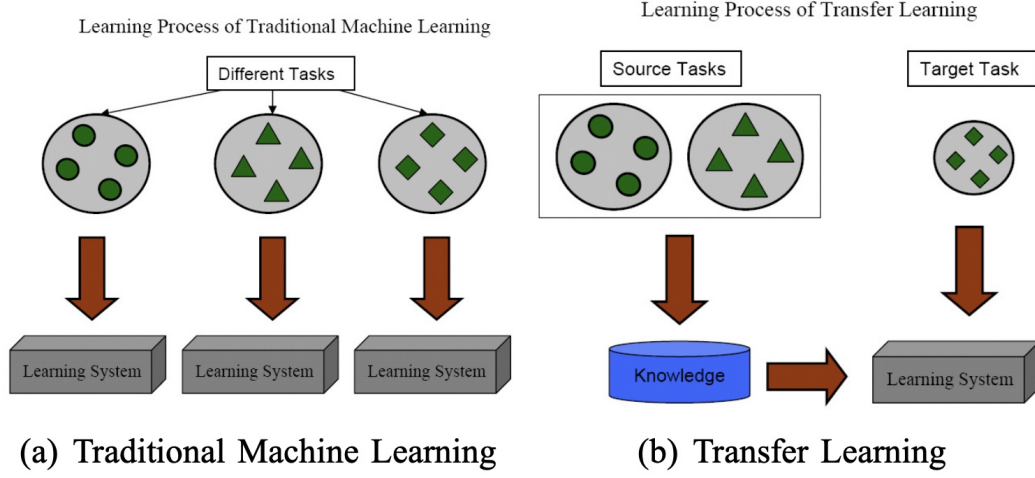


Figura 1.2: Figura tomada del artículo *A Survey on Transfer Learning* (Pan y Yang, 2009), donde se representa la comparación entre aprendizaje automático tradicional y aprendizaje por transferencia.

conjunto:

$$D = \{\mathcal{X}, P(X)\} \quad (1.1)$$

donde $X \in \mathcal{X}$, siendo $X = \{x_1, x_2, \dots, x_n\}$ una variable aleatoria y $n \in N$, donde N es el número de observaciones.

Por lo tanto, la expresión general para un dominio queda definida por el conjunto (1.1). En el marco del aprendizaje por transferencia se distinguen dos dominios: el dominio fuente y el dominio objetivo, que denotamos como D_S y D_T , respectivamente¹. Con base en esta notación, las definiciones precisas de estos conjuntos se presentan en las definiciones 1.1.2 y 1.1.3, correspondientes al dominio fuente y al dominio objetivo, respectivamente.

Definición 1.1.2 (Conjunto fuente). Sea $D_S = \{(x_{S_1}, y_{S_1}), \dots, (x_{S_n}, y_{S_n})\}$, donde $n \in N_S$, $x_{S_i} \in X_S$ es la instancia y $y_{S_i} \in Y_S$ es su etiqueta.

Definición 1.1.3 (Conjunto objetivo). Sea $D_T = \{(x_{T_1}, y_{T_1}), \dots, (x_{T_m}, y_{T_m})\}$, donde $m \in N_T$, $x_{T_i} \in X_T$ es la instancia y $y_{T_i} \in Y_T$ es su etiqueta.

En los paradigmas tradicionales de aprendizaje automático, se asume que el espacio de características y distribución de probabilidad marginal son iguales para los datos de entrenamiento (*train*) y prueba (*test*), es decir: $\mathcal{X}_{train} = \mathcal{X}_{test}$

¹Usamos los subíndices S y T para referirnos a los términos en inglés: source y target.

y $P(X_{train}) = P(X_{test})$. Según lo establecido en la definición 1.1.1, se desprende que la discrepancia entre dos dominios puede surgir tanto por diferencias en sus espacios de características como en las distribuciones marginales de probabilidad. En la sección 1.2 haremos una distinción entre tipos de transferencia de conocimiento basada en estas diferencias.

Por otra parte, dentro de cada dominio tenemos las tareas, es decir, el problema que queremos resolver. Estas a su vez se componen de un espacio de etiquetas y una función predictiva objetivo, que notamos como $\mathcal{Y}$ y $f(\cdot) : X \rightarrow Y$, respectivamente (Pan y Yang, 2009). A continuación, se presenta la definición formal del concepto de tarea en la definición 1.1.4.

Definición 1.1.4 (Tarea). *Una tarea en su forma general se expresa como $\mathcal{T} = (\mathcal{Y}, f(\cdot))$, donde $(\mathcal{Y})$ es un espacio de etiquetas y $f(\cdot) : X \rightarrow Y$ una función predictiva objetivo.*

Respecto al espacio de etiquetas, tenemos que $\mathcal{Y} = \mathbb{R}$ para el caso de regresión y $\mathcal{Y} = [0, 1]$ para el caso de clasificación binaria.

En el aprendizaje por transferencia, generalmente se cumple que $0 \leq m \ll n$, lo que significa que la cantidad de instancias en el dominio fuente es mucho mayor que en el dominio objetivo ($|D_T| \ll |D_S|$). Esta situación se define formalmente en la definición 1.1.5 (Pan y Yang, 2009).

Definición 1.1.5 (Aprendizaje por transferencia). *Dado un dominio fuente D_S y una tarea de aprendizaje T_S , un dominio objetivo D_T y una tarea de aprendizaje T_T , el objetivo del aprendizaje por transferencia es mejorar el aprendizaje de la función predictiva objetivo $f_T(\cdot)$ en D_T utilizando el conocimiento en D_S y T_S , donde $D_S \neq D_T$ o $T_S \neq T_T$.*

Cuando los dominios objetivo y fuente son iguales ($D_S = D_T$), y las tareas también ($\mathcal{T}_S = \mathcal{T}_T$), el problema de aprendizaje por transferencia se convierte en un problema tradicional de aprendizaje automático. Cuando los dominios son diferentes, entonces puede suceder que: los espacios de características entre los dominios son distintos ($X_S \neq X_T$), o las distribuciones marginales de probabilidad entre los datos de los dominios son diferentes ($P(X_S) \neq P(X_T)$, donde $X_{S_i} \in X_S$ y $X_{T_i} \in X_T$).

Cuando realizamos transferencia de conocimiento, se aprovecha la información extraída del dominio fuente para facilitar la búsqueda de la función óptima en el dominio objetivo (f_T^*). Es decir, f_T^* es la función que minimiza el error

esperado de predicción, según una función de pérdida L_T , y que proporciona el mejor desempeño para la tarea específica en dicho dominio.

En el aprendizaje tradicional (sin transferencia), se busca directamente f_T^* resolviendo el siguiente problema de optimización:

$$\min_{f \in A_T} \mathcal{L}_T(f_T) = \min_{f \in A_T} E[L_T(Y_T, f_T(X_T))] \quad (1.2)$$

donde:

- $\mathcal{L}_T$ es la función de riesgo que mide la capacidad de predicción del modelo,
- L_T es la función de pérdida evaluada sobre los datos,
- $E[\cdot]$ denota la esperanza con respecto a la distribución de datos en el dominio objetivo,
- $A_T \subset \{f_T \mid f_T : \mathcal{X}_T \rightarrow \mathcal{Y}_T\}$ es el conjunto de modelos posibles para la tarea objetivo.

Sin embargo, en el contexto de la transferencia de conocimiento, en lugar de buscar directamente f_T^* , se aprovecha el conocimiento previo obtenido en un dominio fuente. Para ello, se usa un modelo previamente entrenado en dicho dominio (*modelo preentrenado*), obtenido al resolver el problema de optimización correspondiente, resultando en la función f_S^* . Esta función, que captura representaciones útiles o patrones aprendidos en el dominio fuente, es utilizada como base para guiar el aprendizaje en el dominio objetivo, ya sea a través de la reutilización directa del modelo, la adaptación de sus parámetros o la transferencia de representaciones intermedias, con el fin de facilitar la aproximación de f_T^* .

Estos modelos preentrenados pueden clasificarse en dos grandes categorías:

- *Modelos fundacionales*, que son entrenados con grandes volúmenes de datos y diseñados para adaptarse a una variedad de tareas posteriores (Bommasani et al. 2021). Su gran capacidad de generalización se debe a las enormes y heterogéneas cantidades de datos utilizadas en su fase inicial de entrenamiento, conocida como preentrenamiento. Esto les da la habilidad de ajustarse con facilidad a un amplio rango de tareas, ofreciendo una base sólida y versátil para diversos contextos (Germán-Morales et al. 2024).
- Modelos ad-hoc, que son específicamente desarrollados para la transferencia de conocimiento, los cuales suelen estar diseñados para un dominio

particular.

La transferencia de conocimiento puede lograrse a través de diversas estrategias, por lo que nos adentraremos en las diferentes categorizaciones en la sección que sigue.

1.2. Categorización de técnicas de aprendizaje por transferencia

Dadas las definiciones anteriores de dominios y tareas, en la presente sección es pertinente realizar una clasificación de las técnicas de transferencia de conocimiento basado en las diferentes relaciones que pueden establecerse entre ambos, basándonos en el trabajo de Zhuang et al. (2020). En la figura 1.3 tomamos de dicho trabajo un esquema de las diferentes categorizaciones de aprendizaje por transferencia que veremos a continuación, según si se trata de categorizar el problema o la solución.

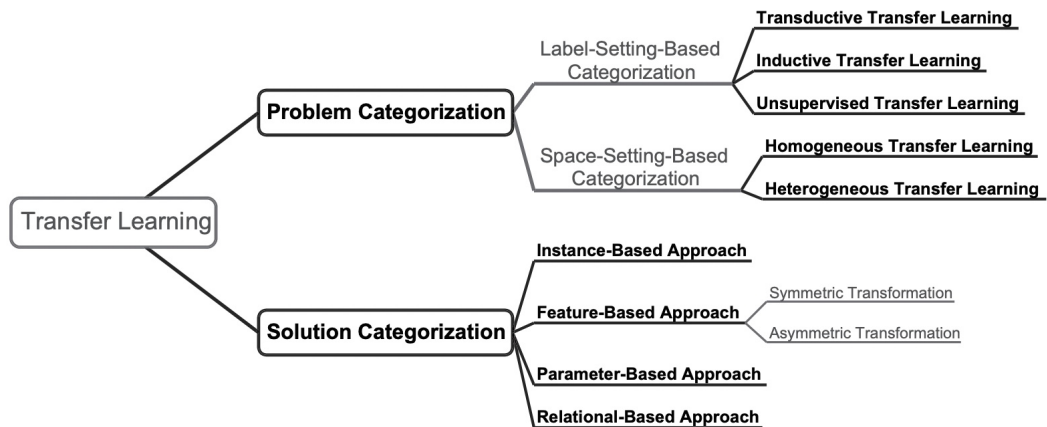


Figura 1.3: Categorización de técnicas de transferencia de conocimiento (Zhuang et al. 2020)

En la sección 1.2.1 se presentan las distintas formas de clasificar los tipos de transferencia basadas en el problema a resolver. En la sección 1.2.2 se describen las principales estrategias empleadas, abordadas desde el punto de vista de la solución.

1.2.1. Categorización del problema

En cuanto a la categorización de problemas, tenemos por un lado la categorización basada en la asignación y disponibilidad de etiquetas (*label-setting-based*) y, por otro, la basada en el espacio de características (*space-setting-based*). La primera se refiere a la disponibilidad de etiquetas en los dominios origen y destino, y distingue entre transferencia *inductiva*, *transductiva* y no supervisada según si ambos, uno o ninguno de los dominios disponen de datos etiquetados, respectivamente. La segunda categoriza los problemas según qué tan parecidos son los datos y las etiquetas en el dominio fuente y en el dominio objetivo. Divide entre transferencia homogénea y heterogénea. La transferencia homogénea se da cuando ambos dominios comparten los mismos espacios y solo difiere la distribución ($\mathcal{X}_S = \mathcal{X}_T$). En cambio, la heterogénea se da cuando difieren en tipo o estructura de los datos y etiquetas, requiriendo técnicas adicionales de transformación del espacio de representación, es decir, si $\mathcal{X}_S \neq \mathcal{X}_T$ y/o $\mathcal{Y}_S \neq \mathcal{Y}_T$. En el trabajo de Weiss et al. (2016) se profundiza sobre estas dos últimas técnicas, que no serán el foco de esta sección.

A continuación enfocaremos exclusivamente en la clasificación basada en la asignación y disponibilidad de etiquetas, excluyendo del análisis el aprendizaje por transferencia no supervisado. Este es similar al paradigma de aprendizaje de transferencia inductiva, en donde la tarea objetivo es diferente a la tarea fuente $\mathcal{T}_T \neq \mathcal{T}_S$, pero están relacionadas. Las dos grandes diferencias son que este tipo de transferencia se centra en resolver tareas de aprendizaje no supervisado (clusterización, reducción de la dimensionalidad, etc.) y en ninguno de los dominios se cuenta con datos etiquetados.

Antes de avanzar en las secciones siguientes, es útil definir la transferencia inductiva y transductiva de forma general. El concepto de transferencia transductiva fue introducido por Vladimir Vapnik en su trabajo *Learning by Transduction* (Gammerman et al. 2013), donde se describe como inductivo a un algoritmo que busca extraer una regla general para clasificar instancias no observadas. En contraste, el enfoque transductivo se enfoca en clasificar correctamente ciertas instancias específicas, sin aspirar a generar un modelo que generalice a todos los ejemplos futuros. Esto implica que, durante el entrenamiento, los datos del conjunto de prueba están disponibles para el modelo, aunque no sus etiquetas, lo cual orienta al algoritmo para que aproveche las características de las instancias en las que se centra.

Es importante señalar que el aprendizaje transductivo es ortogonal al aprendizaje de transferencia, es decir, se puede tener un algoritmo transductivo que haga o no el supuesto de aprendizaje por transferencia, y viceversa (Arnold et al. 2007). A continuación, introducimos brevemente el aprendizaje por transferencia en ambos contextos.

Definición 1.2.1 (Aprendizaje por transferencia inductivo). *Dado un dominio D_S y una tarea $\mathcal{T}_S$, y otro dominio D_T y tarea $\mathcal{T}_T$, el aprendizaje inductivo busca mejorar la función predictiva objetivo $f_T(\cdot)$ en el dominio D_T , utilizando el conocimiento del dominio D_S y la tarea $\mathcal{T}_S$, donde $\mathcal{T}_S \neq \mathcal{T}_T$.*

Nos referimos entonces al aprendizaje por transferencia inductivo cuando la tarea objetivo difiere de la tarea fuente, independientemente de si los dominios son idénticos o no. Para aplicar este enfoque, es necesario contar con datos etiquetados en el dominio objetivo, lo cual permite inducir un modelo predictivo que generalice adecuadamente, logrando clasificar correctamente instancias que no fueron vistas durante el entrenamiento (Pan, 2020). No nos estaremos enfocando en este tipo de transferencia en el presente trabajo.

A continuación citamos un ejemplo de Arnold et al. (2007) para ilustrar un caso de aprendizaje por transferencia transductivo. Supongamos que tenemos un problema de reconocimiento de entidades (entity recognition), donde se cuenta con una serie de artículos de enciclopedia en donde tenemos etiquetados todos los nombres personales. Consideremos ahora el mismo problema (detectar nombres personales) pero aplicado a mensajes de texto. Es evidente que los problemas de identificar nombres en artículos de enciclopedia y mensajes de texto tienen una relación estrecha. Sin embargo, hay factores que pueden ser relevantes en uno de los dominios pero no en el otro, por ejemplo, la capitalización de los nombres es relevante en los artículos de enciclopedia que traen aparejado cierto formalismo, mientras que esto no sería tan relevante en el contexto de mensajes instantáneos donde prima la informalidad. Entonces, surge la necesidad de alterar el clasificador aprendido en el primer contexto para que pueda ser adaptado al segundo, en donde no contamos con etiquetas para los datos.

Definición 1.2.2 (Aprendizaje por transferencia transductivo). *Dado un dominio fuente D_S y una tarea $\mathcal{T}_S$, y por otro lado un dominio objetivo D_T y su correspondiente tarea $\mathcal{T}_T$, el enfoque de aprendizaje por transferencia trans-*

ductivo intenta mejorar el aprendizaje de la función predictiva $f_T(\cdot)$ de D_T , utilizando el conocimiento aprendido en D_S y $\mathcal{T}_S$, donde $D_S \neq D_T$ y $\mathcal{T}_S = \mathcal{T}_T$.

Tenemos que para los casos de aprendizaje transductivo, presentado en la definición 1.2.2 no hay datos etiquetados en el dominio objetivo, mientras que hay relativa abundancia de datos etiquetados en el dominio fuente. Es necesario contar con datos no etiquetados del dominio objetivo durante el entrenamiento. Además, según las diferencias específicas entre los dominios, esta configuración de transferencia transductiva puede dividirse en dos casos:

- Los espacios de características entre los dominios fuente y objetivo son diferentes, es decir, $X_S \neq X_T$.
- Los espacios de características entre los dominios son iguales, $X_S = X_T$, pero las distribuciones de probabilidad marginal de los datos de entrada son distintas, $P(X_S) \neq P(X_T)$.

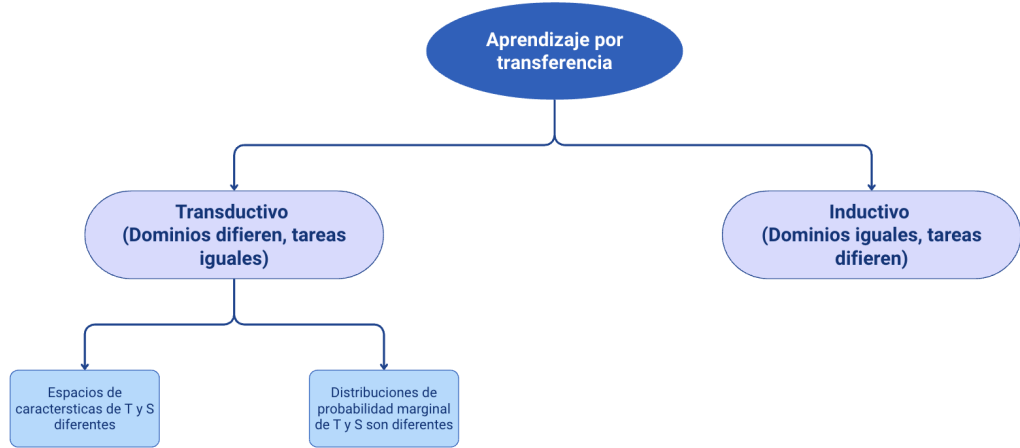


Figura 1.4: Esquema propio de la división entre aprendizaje por transferencia transductivo e inductivo.

Presentamos las diferentes casuísticas en la figura propia 1.4. En el presente trabajo nos enfocaremos en un paradigma de aprendizaje por transferencia transductivo.

1.2.2. Categorización de la solución

En esta sección describimos la segunda categoría dentro del aprendizaje por transferencia presentada en la figura 1.3. Para lograr una transferencia de conocimiento confiable, es esencial responder las siguientes tres preguntas:

1. Qué se transfiere;
2. Cómo se transfiere;
3. Cuándo se transfiere, es decir, en qué circunstancias es beneficiosa la transferencia de conocimiento.

La primera pregunta intenta develar qué conocimiento es el que se transfiere de un dominio a otro. Para esto tenemos principalmente cuatro posibles enfoques, que representamos en la figura 1.3:

i) **Enfoque basado en instancias**, en el cual ciertas partes de los datos del dominio fuente se reutilizan en el dominio objetivo mediante un esquema de ponderación. Un ejemplo destacado es el algoritmo TrAdaBoost, presentado en *Boosting for Transfer Learning* (Dai et al. 2007), que extiende la metodología de AdaBoost. AdaBoost, propuesto por Freund et al. (1999), busca mejorar la precisión de un clasificador ajustando iterativamente los pesos de las instancias de entrenamiento, bajo el supuesto de que las distribuciones de entrenamiento y prueba son idénticas. TrAdaBoost rompe este supuesto, ya que asume que el dominio fuente y el dominio objetivo tienen distribuciones diferentes. Para ello, utiliza una pequeña parte de datos del dominio objetivo con el fin de evaluar la utilidad de cada instancia del dominio fuente. Aquellas instancias fuente que resultan perjudiciales para el desempeño del modelo en el dominio objetivo (por ser demasiado diferentes o estar mal clasificadas debido al cambio de distribución) reciben un menor peso en las iteraciones, debilitando su impacto en el modelo final. De esta forma, se filtran automáticamente los datos que pueden perjudicar el rendimiento del modelo, y se conservan los que aún aportan valor.

ii) **Enfoque basado en representación de características**, donde el objetivo es poder aprender una buena representación de los regresores en el dominio objetivo;

iii) **Enfoque relacional**, en el cual se asume que existen vínculos o dependencias entre los datos de los dominios fuente y objetivo, y es precisamente este tipo de relación el conocimiento que puede transferirse. Un ejemplo de este enfoque se presenta en Mihalkova y Mooney (2008), donde proponen un escenario en el que el dominio destino dispone únicamente de datos sobre una *entidad* central. En este contexto, una entidad es un elemento individual del dominio (por ejemplo, una persona, una organización o un nodo en una red) sobre el cual se registran atributos y relaciones con otras entidades. Los

autores introducen el algoritmo SR2LR, que primero identifica *cláusulas de corto alcance* vinculadas a esa entidad (aquellas que involucran únicamente a la entidad central y que pueden evaluarse directamente con los pocos datos disponibles en el dominio destino) para establecer correspondencias entre predicados (relaciones o atributos que describen propiedades de las entidades o conexiones entre ellas) del dominio fuente y destino. Luego, utiliza dichas correspondencias para transferir *cláusulas de largo alcance*, que describen relaciones más amplias en la red y que no pueden evaluarse directamente, pero cuyas estructuras pueden inferirse a partir de las coincidencias encontradas en el corto alcance. Este procedimiento permite transferir relaciones y patrones entre dominios con estructuras similares, incluso cuando el dominio objetivo cuenta con datos limitados;

iv) **Enfoque basado en parámetros**, en el cual la transferencia se realiza a través de los parámetros del modelo seleccionado. Conceptualmente, según Obst et al. (2022), este tipo de estrategias puede considerarse análogo a un enfoque bayesiano, donde los parámetros aprendidos en el dominio fuente actúan como un prior para el dominio objetivo. En secciones posteriores, nos inspiraremos en esta aproximación para nuestros experimentos, utilizando parámetros estimados en el dominio fuente como inicialización para el entrenamiento en el dominio objetivo, con el fin de evaluar si la transferencia resulta beneficiosa bajo distintas condiciones de similitud entre dominios.

Por último, si los dominios objetivo y fuente no tienen ninguna relación entre sí, existe la posibilidad de que llevar a cabo la transferencia resulte en efectos no deseados e incluso contraproducentes para el rendimiento del nuevo modelo. En estos casos, el modelo estimado directamente en los datos del dominio objetivo ($\hat{f}_T$, de acuerdo a la definición 1.2.3) resulta mejor que el modelo estimado apoyándose en los datos del dominio fuente ($\hat{f}_{TL}$, de acuerdo a la definición 1.2.4), es decir, que la medida de error en el primero es menor que la medida de error en el segundo. Esto se puede encontrar bajo el nombre de *aprendizaje negativo* (*negative transfer*) (Zhang et al. 2022), lo cual profundizaremos en la sección 1.3. Por lo tanto, es relevante responder la tercera pregunta, para considerar la transferencia únicamente para los casos en los que puede ser beneficiosa.

Definición 1.2.3 (Modelo estimado directamente en el dominio objetivo). *Sea f una función objetivo que queremos estimar para un dominio dado. El modelo*

estimado directamente en el dominio objetivo se define como

$$\hat{f}_T = \arg \min_f \mathcal{L}_T(f),$$

donde $\mathcal{L}_T(f)$ es la función de riesgo evaluada únicamente sobre los datos del dominio objetivo.

Definición 1.2.4 (Modelo estimado mediante aprendizaje por transferencia). Sea f una función objetivo que queremos estimar para un dominio dado. El modelo estimado mediante aprendizaje por transferencia se define como

$$\hat{f}_{TL} = \arg \min_f \mathcal{L}_T(f) + \lambda \mathcal{R}_S(f),$$

donde $\mathcal{R}_S(f)$ es un término que incorpora información del dominio fuente y $\lambda \geq 0$ es un hiperparámetro que controla la influencia del dominio fuente en la estimación.

En este trabajo analizamos bajo qué condiciones la transferencia de conocimiento resulta beneficiosa cuando existe dependencia entre los datos. Dado que las observaciones cercanas en el tiempo o en el espacio suelen estar correlacionadas, la independencia estadística no siempre se cumple. Aunque este fenómeno ha sido ampliamente estudiado en distintos ámbitos, su impacto en la transferencia de conocimiento ha recibido poca atención. Este aspecto es especialmente relevante en series temporales, donde el objetivo es modelar la dependencia entre observaciones consecutivas.

1.3. Aprendizaje negativo (Negative transfer)

Múltiples estudios han sido realizados sobre la transferencia negativa en el aprendizaje humano de nuevas tareas, mostrando cómo experiencias previas pueden dificultar la resolución de nuevos problemas al limitar la flexibilidad para encontrar soluciones más eficaces. Estas investigaciones aportan evidencia sobre la transferencia negativa, que puede bloquear la capacidad de resolver problemas de manera eficiente (Alzayat, 2011). El mismo concepto lo trasladamos al aprendizaje por transferencia, estudiado por primera vez en el artículo *To Transfer or Not To Transfer* (Rosenstein et al. 2005).

La efectividad del transfer no está garantizada si no se cumplen ciertas condiciones: (1) las tareas de ambos dominios son similares, (2) sus distribuciones

de datos no son muy diferentes. Si estas condiciones no se cumplen, puede producirse transferencia negativa, donde la información del dominio fuente perjudica el aprendizaje en el dominio objetivo (Zhang et al. 2022). Por ejemplo, si se intenta transferir un modelo de predicción de ventas de una tienda en una región urbana a otra tienda en una zona rural con patrones de consumo muy distintos, la transferencia puede empeorar el desempeño del modelo en la segunda tienda, debido a las diferencias en las distribuciones y tareas.

Entre las preguntas planteadas en la sección 1.2.2, una clave para evitar la transferencia negativa es determinar en qué momento es apropiado realizar la transferencia. Se define formalmente la transferencia negativa en la definición 1.3.1 (Z. Wang et al. 2019).

Definición 1.3.1 (Transferencia negativa). *Sea ϵ_T el error en el conjunto de testeo dentro del dominio objetivo, ϵ_S el error en el conjunto de testeo dentro del dominio fuente, $\theta(S, T)$ un algoritmo de transferencia entre el dominio fuente y objetivo, y $\theta(\emptyset, T)$ el mismo algoritmo, pero sin utilizar el dominio fuente. Entonces, la transferencia negativa (NT) ocurre cuando $\epsilon_T(\theta(S, T)) > \epsilon_T(\theta(\emptyset, T))$.*

El grado de la transferencia negativa se puede evaluar mediante la brecha de transferencia negativa (NTG - negative transfer gap) presentada en la definición 1.3.2. Este valor se puede computar siempre y cuando tengamos datos del conjunto objetivo etiquetados.

Definición 1.3.2 (Brecha de transferencia negativa (NTG)). *Definimos la brecha de transferencia negativa como: $NTG = \epsilon_T(\theta(S, T)) - \epsilon_T(\theta(\emptyset, T))$*

Este concepto se ilustra en la figura 1.5. En el escenario de la izquierda, el modelo se entrena y evalúa exclusivamente con datos del dominio objetivo, alcanzando una precisión en prueba del 70 %. En el escenario de la derecha, se incorpora información adicional proveniente de un dominio fuente con el objetivo de mejorar el rendimiento en el dominio objetivo. Sin embargo, debido a discrepancias entre las distribuciones de datos, diferencias semánticas o sesgos entre ambos dominios, el modelo incorpora patrones no representativos para el dominio objetivo. Como consecuencia, se observa una degradación en el desempeño, reduciendo la precisión de prueba al 60 %. Este efecto, denominado ilustra el hecho de que la transferencia de conocimiento desde dominios fuente no siempre es beneficiosa y que es necesario evaluar cuidadosamente la compatibilidad entre dominios antes de aplicar técnicas de transferencia.

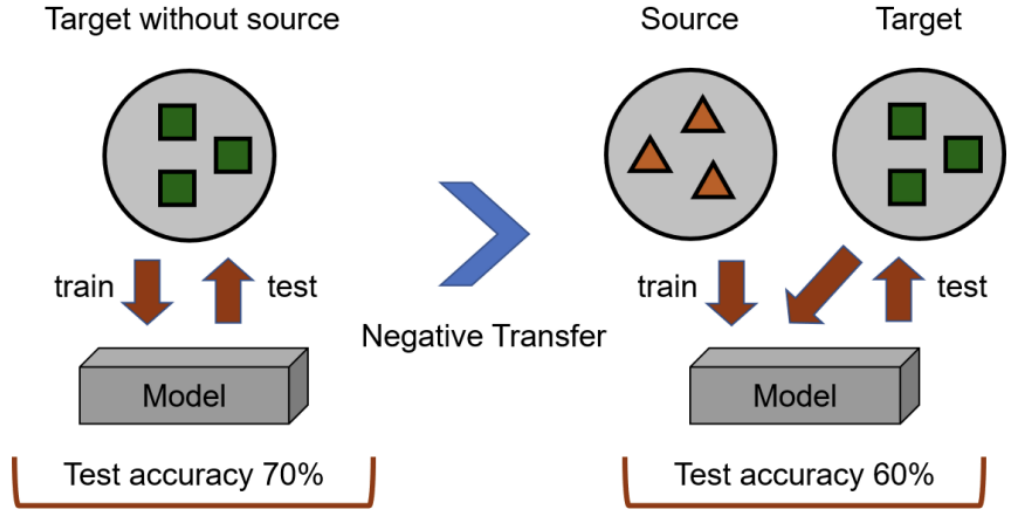


Figura 1.5: Figura tomada del estudio *A Survey on Negative Transfer* (Zhang et al. 2022).

La transferencia negativa puede surgir por la divergencia entre dominios, el algoritmo utilizado o la calidad de los datos (Zhang et al. 2022). Una gran diferencia entre las características o la estructura de los dominios puede hacer que el conocimiento transferido sea inefectivo. Asimismo, un algoritmo de transferencia inadecuado o datos fuente de baja calidad (por ejemplo, con ruido) pueden generar modelos poco confiables. Finalmente, si los datos del dominio objetivo presentan cambios en el tiempo no reflejados en el dominio fuente, el modelo puede volverse inestable y degradar su rendimiento. Por estos motivos, en el presente trabajo nos concentramos en comprender cuándo es beneficioso realizar la transferencia, para lo cual es necesario entender qué factores de similitud entre dominios juegan un rol clave en dicho proceso.

En *A Survey on Negative Transfer* (Zhang et al. 2022) se definen tres enfoques existentes para la estimación de similitud entre dominios:

- **Basados en estadísticas de características:**

Las representaciones de características originales y sus estadísticas, como la media y la covarianza, son fundamentales para medir la discrepancia en la distribución entre dominios. Entre las métricas más utilizadas se encuentran la Máxima Discrepancia de Media (MMD) (Gretton et al. 2012) y la divergencia de Kullback-Leibler (KL) (Kullback y Leibler, 1951), que es utilizada para clasificar dominios según su similitud con el dominio objetivo. Métodos como el criterio de independencia de Hilbert-Schmidt

(HSIC) (Gretton et al. 2005) y la distancia de Wasserstein (Kantorovich, 2006) también se aplican para medir la discrepancia en la distribución de características, mejorando la capacidad de adaptación del modelo entre dominios.

En contextos de dependencia temporal o espacial se suelen utilizar otro tipo de medidas que consideren este tipo de dependencias. Por ejemplo, el índice de Moran (Moran, 1950) es una medida de autocorrelación espacial, que se usa principalmente para evaluar la similitud entre elementos relacionados espacialmente, midiendo dependencia espacial de los valores en un conjunto de datos. Para el caso de datos temporales, la distancia Dynamic Time Wrapping (DTW) es una medida que calcula la disimilitud entre dos series, permitiendo alinearlas de manera no lineal para optimizar su comparación. DTW es útil cuando las series temporales no están alineadas en el tiempo, lo que puede ser común en problemas de series con distintas escalas o longitud (ver sección 5 en el apéndice). Otras medidas como TWED (Time Wrap Edit Distance) o EDR (Edit Distance on Real sequences) serán vistas también en el presente trabajo (ver sección 5 en el apéndice).

- **Basados en el rendimiento en pruebas.**

La similitud entre dominios también puede medirse a partir del rendimiento del modelo en el conjunto de prueba. Si un modelo del dominio fuente logra un buen desempeño en los datos etiquetados del dominio objetivo, entonces se considera que ambos dominios son muy similares. Sin embargo, esto requiere que el dominio objetivo tenga datos etiquetados.

- **Basados en fine tuning.**

Este enfoque es común en aprendizaje profundo para adaptar un modelo de dominio fuente al dominio objetivo, fijando los parámetros de las capas inferiores de las redes neuronales y reajustando los parámetros de las capas superiores. En general, estas aproximaciones alimentan los datos del dominio objetivo en la red neuronal del dominio fuente y utilizan la salida para determinar la similitud entre los dominios.

A continuación presentamos algunos trabajos relacionados con la transferencia de conocimiento, y especialmente en la sección 1.4.1 presentamos estudios con aplicaciones en series de tiempo, que es el foco de esta tesis.

1.4. Trabajos relacionados

El presente apartado tiene como objetivo revisar investigaciones previas vinculadas al aprendizaje por transferencia, con especial atención a aquellas que abordan su aplicación en contextos similares al de este trabajo para poder identificar enfoques, metodologías y hallazgos relevantes en el área.

El aprendizaje por transferencia ha adquirido gran popularidad en los últimos años como tema de investigación dentro de las comunidades de aprendizaje automático y estadística. Esto se debe, en gran parte, a la disponibilidad de modelos previamente entrenados en grandes volúmenes de datos, que luego son ajustados para tareas más específicas.

En el aprendizaje profundo, cuando se dispone de conjuntos de datos pequeños, es habitual emplear técnicas de transferencia de conocimiento a partir de grandes modelos, como los desarrollados con el conjunto de datos ImageNet (Deng et al. 2009) y sus modelos preentrenados, por ejemplo, ResNet50 (He et al. 2016).

Recientemente, el campo del Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés) ha sido revolucionado por la irrupción de modelos preentrenados en extensos conjuntos de datos que abarcan millones de patrones y contextos lingüísticos. Modelos como BERT (Devlin et al. 2019) y GPT (Radford et al. 2018) generan texto imitando el lenguaje humano al predecir la siguiente palabra en una secuencia. Posteriormente, estos modelos sirven como base para adaptarlos a tareas concretas, mejorando tanto el rendimiento como la eficiencia de las predicciones.

Estos modelos preentrenados, conocidos como modelos fundacionales, están diseñados para ser versátiles: su entrenamiento en datos amplios y diversos les permite luego ser ajustados a múltiples tareas específicas (Germán-Morales et al. 2024). Así, la transferencia de conocimiento es el mecanismo que hace posible esta adaptabilidad (Bommasani et al. 2021). Un ejemplo de ello es la denominada “homogeneización”, que ocurre cuando gran parte de los modelos en NLP, como BERT (Devlin et al. 2019), RoBERTa (Y. Liu et al. 1907) o BART (M. Lewis, 2019), se basan en modelos fundacionales para sus aplicaciones.

Aunque muchas investigaciones se centran en aplicaciones de visión por computadora (C. Tan et al. 2018) o procesamiento de lenguaje natural (Al-yafeai et al. 2020), un número creciente de estudios ha comenzado a explorar

campos como las series temporales o la estadística espacial. Estos datos presentan desafíos únicos y una relevancia práctica considerable, lo que ha despertado un notable interés en ámbitos académicos e industriales (Goswami et al. 2024). De hecho, el éxito que han tenido las técnicas de transferencia en visión por computadora sugiere que también pueden ser eficaces en series temporales, dado que estas comparten características generales que pueden ser transferidas entre distintos dominios (Fawaz et al. 2018).

En la sección que sigue abordamos dentro de este campo algunos estudios específicos sobre transferencia en el contexto de series temporales, que será el foco de esta tesis.

1.4.1. Series de tiempo

A continuación, se describen investigaciones relevantes en la transferencia de conocimiento aplicada a series temporales, cuyo objetivo es determinar en qué situaciones esta técnica aporta mejoras significativas.

La transferencia de conocimiento se ha mostrado particularmente efectiva en contextos donde las series temporales del conjunto objetivo contienen pocas observaciones. En este enfoque, se entrena un modelo con uno o varios conjuntos de datos de origen y luego se adapta a la tarea final (a través de fine tuning) o se emplea directamente para hacer predicciones sin ajuste previo (enfoques de tipo *zero-shot*, donde el modelo generaliza a nuevas tareas sin haber sido específicamente entrenado en ellos) (C. Tan et al. 2018). Cuando la tarea es la misma, pero los dominios de origen y destino difieren, resulta esencial evaluar la similitud entre los conjuntos de datos de origen y el conjunto objetivo.

Un ejemplo de esta aplicación es el trabajo presentado en "*Un enfoque de transferencia de conocimiento basada en redes para mejorar la predicción de ventas de nuevos productos*" (Karb et al. 2020). Allí, se aborda el desafío en el comercio minorista de prever las ventas de productos nuevos que no cuentan con datos históricos suficientes. Tradicionalmente, esta carencia de información se compensa con sistemas basados en el conocimiento de expertos, pero estos pueden estar sesgados o depender de la disponibilidad de los profesionales. Frente a ello, la transferencia de conocimiento permite a los modelos de aprendizaje automático aplicar, de manera más sistemática y precisa, lo aprendido en dominios similares (Karb et al. 2020). Los autores concluyen que esta técnica mejora las predicciones, sobre todo en las primeras semanas y durante

períodos de promociones.

Sin embargo, este estudio también evidencia que la similitud entre dominios no garantiza por sí sola un desempeño óptimo. La selección de los conjuntos de datos de origen es un paso clave, ya que no todos los modelos fuente producen mejoras consistentes en todos los casos. Factores como compartir características relevantes, por ejemplo, el rango de precios o el promedio de ventas por hora, pueden hacer que ciertos modelos fuente sean más efectivos que un entrenamiento desde cero. Estos hallazgos subrayan la importancia de considerar múltiples aspectos, más allá de unas pocas características compartidas.

Por otro lado, no todos los conjuntos de origen disponibles son útiles para transferir conocimiento a otro dominio. Comprender cómo las características de los conjuntos de datos y su relación afectan el resultado de la transferencia es, por tanto, fundamental. En su artículo *Transfer learning for time series classification*, Fawaz et al. (2018) emplean más de 7000 redes neuronales como modelos fuente y se demuestra que la efectividad de la transferencia depende del conjunto de datos seleccionado como origen. Se propone, además, utilizar la técnica DTW para medir la similitud entre los conjuntos de datos, logrando mejoras de rendimiento en el 84 % de los casos. Este artículo también destaca que la transferencia de conocimiento puede ayudar a mitigar el sobreajuste en conjuntos de datos muy pequeños, un escenario común en las aplicaciones de series temporales.

En otro estudio, Ye y Dai (2021) introducen un algoritmo que primero elige el conjunto de datos de origen más adecuado de entre varias opciones y luego utiliza redes neuronales convolucionales con causalidad para realizar las predicciones. Para ello, no solo aplican fine tuning, sino que integran la transferencia directamente en el aprendizaje de las características, lo que permite reconocer patrones significativos comunes al dominio de origen. Además, utilizan DTW junto con la divergencia de Jensen-Shannon (Majtey et al. 2005) para comparar conjuntos de datos: DTW alinea las formas de las series y la divergencia mide las distribuciones internas, ofreciendo una evaluación más completa de la similitud entre dominios.

Por último, el artículo *The Impact of Dataset Similarity and Diversity on Transfer Learning* (Ehrig et al. 2024) estudia cómo influyen estas características en la efectividad de la transferencia en series temporales. Se emplean técnicas de zero-shot y fine tuning en cinco conjuntos de datos objetivo, y se utilizan métricas basadas en características y formas para medir la similitud

entre los conjuntos. Los resultados sugieren que la similitud en características ayuda a reducir el sesgo en las predicciones, mientras que la diversidad mejora la precisión y la estimación de incertidumbre. Sin embargo, no encuentran una relación consistente entre el desempeño y las métricas basadas únicamente en la forma, como la DTW.

En conjunto, estos estudios ponen de manifiesto que la transferencia de conocimiento no depende de un solo factor, sino de la interacción entre la similitud entre dominios, las características de cada conjunto de datos y las particularidades de la tarea a resolver. Esto refuerza la importancia de seguir investigando cómo elegir los conjuntos de datos de origen y definir criterios de similitud que permitan optimizar la transferencia de conocimiento, especialmente en escenarios donde los datos son escasos o difíciles de obtener.

Capítulo 2

Aprendizaje por Transferencia en Series de tiempo

En el presente capítulo repasaremos en primer lugar, las definiciones fundamentales de las series de tiempo, con el propósito de establecer las notaciones y conceptos que se emplearán a lo largo del trabajo. En segundo lugar, se detalla la metodología utilizada para llevar a cabo las simulaciones de transferencia que constituyen el núcleo del estudio de la presente sección. Finalmente, se presentan, analizan y discuten los resultados obtenidos a partir de dichas simulaciones.

2.1. Series de Tiempo

Antes de profundizar en las aplicaciones de las técnicas de transferencia de conocimiento para series temporales que hemos venido mencionando, es necesario establecer algunos conceptos fundamentales que establecen las bases de este trabajo. En la definición 2.1.1 presentamos formalmente qué entendemos por serie temporal o serie de tiempo.

Definición 2.1.1 (Serie de tiempo). *Una serie temporal o serie de tiempo es un conjunto de observaciones de una variable aleatoria $\{X_t\}_{t \in \mathbb{Z}}$, medida de forma secuencial en intervalos de tiempo $t = 1, 2, \dots, T^*$, que suelen ser equidistantes.*

En este trabajo, utilizaremos la notación x_t para representar las series temporales definidas anteriormente, y T^* para indicar el límite temporal, con el

propósito de evitar ambigüedades con el subíndice correspondiente al conjunto *target* (que representamos con la letra T). La frecuencia temporal de una serie puede variar según el caso, adoptando intervalos como horas, días, semanas, meses o años, entre otros. Numerosos conjuntos de datos se representan en forma de series de tiempo: desde registros diarios de pacientes en un hospital, pasando por la evolución horaria de precios de acciones, hasta las ventas mensuales de productos en una tienda. Este tipo de datos es común en disciplinas como las ciencias naturales, las ciencias sociales y la ingeniería. Además, una serie de tiempo puede ser: *univariada*, cuando $x_t \in \mathbb{R}$; o *multivariada*, cuando cada x_t es un vector $(x_t^1, \dots, x_t^d) \in \mathbb{R}^d$.

En este contexto, el análisis estadístico de una serie temporal busca comprender la variabilidad de la variable de interés a lo largo del tiempo y, además, generar estimaciones de sus valores futuros, proceso conocido como pronóstico de series temporales (*time series forecasting* (TSF) en inglés). Las observaciones pasadas de la serie se denominan *retardos* (*lags* en inglés), mientras que el número de pasos temporales que se desea anticipar recibe el nombre de horizonte predictivo (*forecast horizon* en inglés).

A continuación, presentamos la definición y conceptualización de los componentes de una serie temporal, así como su categorización.

2.1.1. Componentes de una serie temporal

Comenzaremos analizando los componentes de una serie a través de un ejemplo ilustrativo. En la figura 2.1 se muestran las observaciones correspondientes a la temperatura promedio diaria en la ciudad de Delhi, registradas entre los años 2013 y 2017¹. Se aprecia que, a medida que avanza el tiempo, la temperatura promedio diaria presenta una componente ascendente. Esta serie no es estable en el tiempo, lo que coincide con un patrón común en fenómenos de naturaleza social o económica, donde los niveles de las variables tienden a variar en lugar de mantenerse constantes.

Esta componente creciente que observamos es lo que podemos denominar *tendencia* (en este caso positiva, ya que el "nivel" de la serie aumenta a lo largo del tiempo). Presentamos la definición formal en la definición 2.1.2.

Definición 2.1.2 (Tendencia). *Dada una serie temporal x_t , definimos como tendencia T_t al proceso que describe la evolución lenta y continua del valor*

¹<https://www.kaggle.com/sumanthvrao/daily-climate-time-series-data>

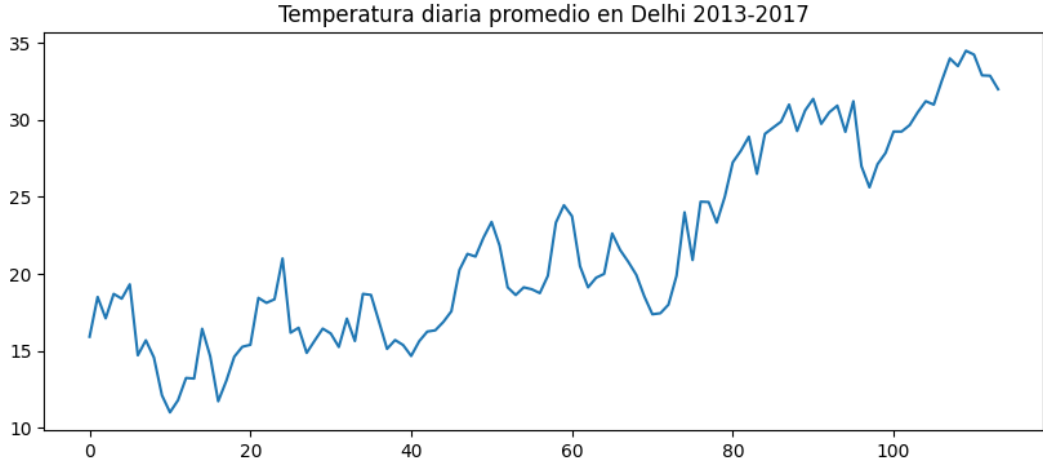


Figura 2.1: Temperatura diaria promedio en Delhi para el período 2013-2017. Datos extraídos de Kaggle.

esperado de x_t a lo largo del tiempo, es decir,

$$E[x_t] = T_t + \varepsilon_t,$$

donde ε_t representa la componente aleatoria.

Toda serie de tiempo puede descomponerse en cuatro componentes:

1. Componente de tendencia (T_t);
2. Componente estacional (E_t);
3. Componente cíclica (C_t), y
4. Componente aleatoria o ruido blanco (ϵ_t)

El componente de tendencia fue definido en 2.1.2. El *componente estacional* (E_t) representa variaciones que se repiten de manera regular en intervalos fijos de tiempo, por ejemplo, cada año, trimestre o mes, y suele estar asociado a patrones climáticos, festivos o de comportamiento del consumidor. La frecuencia de dichos patrones depende del fenómeno que se esté analizando.

El *componente cíclico* (C_t) describe fluctuaciones que no tienen un período fijo y que se relacionan con fenómenos económicos o estructurales de más largo plazo. Por ejemplo, en una serie del PIB (Producto Interno Bruto) trimestral, podrían reflejar ciclos de expansión y recesión que duran varios años, pero cuya duración y amplitud no son constantes.

Finalmente, el *componente aleatorio* o *ruido blanco* (ϵ_t) captura las variaciones impredecibles de la serie que no pueden ser explicadas por los demás

componentes y que se modelan como una sucesión de variables aleatorias independientes con media cero y varianza constante.

Las componentes T_t , E_t y C_t se suponen *determinísticas* (son constantes, aunque pueden variar en el tiempo), por lo que la aleatoriedad de x_t está concentrada por ϵ_t . Las componentes cíclicas van a ser dejadas de lado en los análisis posteriores, porque son más complejas de analizar y dependen de cada serie y contexto en particular.

La relación entre dichos componentes y la variable aleatoria de interés, x_t , puede ser expresada mediante un modelo aditivo o multiplicativo:

$$x_t = T_t + E_t + C_t + \epsilon_t \forall t \quad (2.1)$$

$$x_t = T_t \cdot E_t \cdot C_t \cdot \epsilon_t \quad (2.2)$$

Al modelar series temporales, se suele asumir que los datos no presentan tendencia ni otros componentes que afecten su estabilidad. Para evaluar la invariabilidad del proceso de estudio y determinar si se cumplen las suposiciones básicas de muchos modelos estadísticos, es fundamental estudiar la varianza. Cuando esta no es constante (no hay *homocedasticidad*), los modelos pueden perder precisión y confiabilidad, afectando así la calidad de las predicciones y el análisis.

Si el modelo es multiplicativo, y asumimos independencia entre los elementos de la serie, entonces

$$Var(x_t) = V(T_t E_t C_t \epsilon_t) = T_t^2 E_t^2 C_t^2 \sigma^2,$$

mientras que, si el modelo es aditivo,

$$Var(x_t) = V(T_t + E_t + C_t + \epsilon_t) = \sigma^2.$$

Por lo tanto, en el modelo multiplicativo la varianza (que notamos como $Var(x_t)$ o $V(x_t)$) depende del instante t , mientras que en el modelo aditivo permanece constante para todo valor de t . Esto implica que, en el modelo multiplicativo, la variabilidad o dispersión de los valores de la serie cambia según el nivel en que se encuentre la misma en ese momento del tiempo: cuando esta tiene valores altos, la varianza es mayor, y cuando tiene valores bajos, la

varianza disminuye.

Para identificar esta situación, se puede analizar alguna medida de dispersión, como la desviación estándar, y observar cómo varía en relación con la media de la serie a lo largo del tiempo. En casos donde la varianza no es constante, una práctica común para estabilizarla es aplicar una transformación logarítmica a la serie, lo que permite que las técnicas estadísticas sean más apropiadas y los modelos tengan un mejor desempeño.

A continuación presentamos dos categorías de series de tiempo, basadas en lo mencionado en la presente sección.

2.1.2. Categorización de una serie temporal

En base a los componentes mencionados en la sección 2.1.1, los fenómenos que son representados mediante series de tiempo pueden dividirse en dos categorías:

- *Estacionarios*, lo cual implica que tienen valores estables en el tiempo (media y varianza constantes en el tiempo);
- *No estacionarios*, que pueden mostrar tendencia, estacionalidad y otras variaciones a lo largo del tiempo.

La clasificación en estas dos categorías puede variar según el período de observación, ya que la serie puede ser estacionaria en un período más corto pero no estacionaria si miramos una ventana temporal más amplia. Cabe destacar que la dependencia entre las observaciones tiende a disminuir a medida que aumenta la distancia temporal entre ellas. Esta relación puede analizarse y cuantificarse mediante la *función de autocovarianza*, que mide cómo varía conjuntamente la serie consigo misma en distintos desfases temporales. La planteamos formalmente a continuación.

Definición 2.1.3. *Definimos la autocovarianza de orden p , denotada como γ_p , a la covarianza entre dos observaciones de la serie temporal que están separadas por p períodos. Formalmente, se expresa como:*

$$\gamma_p(t, t - p) = E[(x_{t-p} - \mu)(x_t - \mu)] \quad (2.3)$$

donde:

- x_t representa la serie temporal en el instante t .

- $\mu = E[x_t]$ es la media de la serie.

Como para muchos procesos estocásticos no es posible obtener muestras sucesivas o realizaciones independientes del mismo (solo se puede observar una única realización del proceso), para poder estimar la media y varianza a partir de la evolución de la serie temporal, es necesario suponer que la distribución de las variables en cada instante son estables a lo largo del tiempo. Esto desencadena el concepto de estacionariedad que mencionamos a continuación.

Definición 2.1.4 (Estacionariedad estricta). *Una serie de tiempo x_t se dice estrictamente estacionaria si las distribuciones conjuntas de $(x_{t_1}, \dots, x_{t_p})'$ y $(x_{t_1+h}, \dots, x_{t_p+h})'$ son iguales para todo entero h , y para todo conjunto de tiempos $p = 1, 2, \dots, T^*$ con $p \in \mathbb{Z}$. Es decir,*

$$F(x_i, x_j, \dots, x_p) = F(x_{i+h}, x_{j+h}, \dots, x_{p+h})$$

La distribución conjunta de cualquier subconjunto de variables no se modifica si trasladamos todas las variables en el tiempo.

Es decir, que una serie es estacionaria en sentido estricto si cumple las siguientes condiciones:

- La serie tiene media y varianza constante, es decir, que las *distribuciones marginales* de todas las variables son idénticas.
- La relación entre las variables depende únicamente de sus retardos, y esta relación es la misma sin importar en qué punto del tiempo se observe.

El cumplimiento de estas condiciones implica que la *distribución conjunta* de cualquier conjunto de variables no se modifica si trasladamos las variables en el tiempo. Dicho de otra forma, si un proceso es estacionario, la forma en que las variables están relacionadas entre sí no cambia cuando desplazamos todo el proceso hacia adelante o hacia atrás en el tiempo, es decir, sus distribuciones son iguales. Por lo tanto, para analizar estacionariedad estricta, necesitamos conocer las distribuciones conjuntas para cualquier selección de variables del proceso, lo cual rara vez podemos realizar en la práctica.

Por lo anterior, se define una propiedad de estacionariedad menos restrictiva en la definición siguiente.

Definición 2.1.5 (Estacionariedad débil). *Una serie de tiempo x_t , se dice débilmente estacionaria si cumple las siguientes condiciones:*

- $\mathbb{E}(x_t) = \mu, \quad \forall t \in \mathbb{Z}$ (la media es constante).
- $\mathbb{E}[(x_t - \mu)^2] < \infty, \quad \forall t \in \mathbb{Z}$ (la varianza es finita).
- $\gamma(t, t + p) = \text{Cov}(x_t, x_{t+p}) = \mathbb{E}[(x_t - \mu)(x_{t+p} - \mu)] = \gamma_p, \quad \forall t, p \in \mathbb{Z}$ (la covarianza solo depende del rezago p y no del tiempo t).

Es decir, una serie es estacionaria en sentido débil si la serie es constante en media ($\mu_t = \mu = cte$) y varianza ($\sigma_t^2 = \sigma^2 = cte$), y sus covarianzas solo dependen de la separación entre las variables. El último punto implica que la correlación entre x_t y x_{t-p} , es siempre la misma que entre x_t y x_{t+p} ($\text{Cov}(x_t, x_{t+p}) = \text{Cov}(x_{t+j}, x_{t+j+p}) = \gamma_p$). Es decir, que solo depende de la distancia entre instantes de tiempo y no del instante específico en el que nos encontremos (por ejemplo: $\text{Cov}(x_{101}, x_1) = \text{Cov}(x_{851}, x_{950})$). A lo largo del trabajo, salvo que se indique lo contrario, nos referiremos a un proceso estacionario cuando este sea débilmente estacionario.

Un proceso estacionario muy importante es el denominado *ruido blanco*. Un proceso de ruido blanco es estacionario y se caracteriza porque la esperanza es siempre constante e igual a cero, su varianza es constante y las variables del proceso están incorreladas para todos los retardos. Lo definimos formalmente a continuación.

Definición 2.1.6 (Ruido blanco). *Un proceso estocástico estacionario $\{\varepsilon_t\}$ se denomina ruido blanco si cumple las siguientes condiciones:*

1. $\mathbb{E}(\varepsilon_t) = 0, \quad \forall t \in \mathbb{Z}$.
2. Su función de autocovarianza es:

$$\gamma_p = \begin{cases} \sigma^2 & \text{si } p = 0 \\ 0 & \text{si } p \neq 0 \end{cases}$$

Ejemplos de fenómenos de ruido blanco incluyen lanzar una moneda o generar números aleatorios; en ambos casos, conocer el pasado no nos da información sobre el futuro. El proceso ruido blanco no es necesariamente estacionario en sentido estricto, sólo se impone incorrelación, por tanto, no necesariamente las variables son independientes. Si se impone la condición de que las variables del proceso tienen distribución normal, el proceso será estacionario en sentido estricto y lo denominamos *ruido blanco Gaussiano* (Rodríguez-Collazo, 2024).

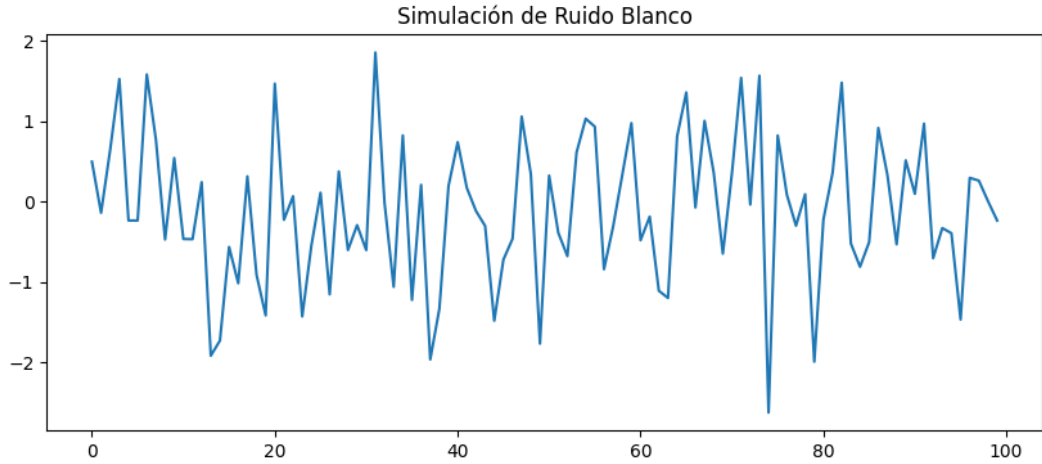


Figura 2.2: Simulación de 100 observaciones de un proceso de ruido blanco Gaussiano.

En la figura 2.2 se muestra una simulación de 100 observaciones de un proceso de ruido blanco Gaussiano, que denotamos como $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$.

En la siguiente sección definimos algunas propiedades que son relevantes en los casos en que trabajamos con series de tiempo.

2.1.3. Propiedades de la dependencia temporal

A diferencia de otros contextos donde se asume que las observaciones son independientes entre sí, en el análisis de series temporales las observaciones están temporalmente correlacionadas. Es decir, el valor de la variable en un instante dado depende, en mayor o menor medida, de los valores que tomó en momentos anteriores. Esta característica viola uno de los supuestos básicos de muchos modelos estadísticos tradicionales: la independencia de las observaciones. Como consecuencia, dichos modelos no pueden aplicarse directamente a series temporales sin ajustar previamente los datos o utilizar herramientas específicas para datos dependientes.

Por ejemplo, en modelos de regresión lineal clásica, se asume que los errores son independientes y no están correlacionados. Cuando se aplican directamente a datos de series temporales, esta suposición suele no cumplirse, lo que puede llevar a estimaciones sesgadas. Además, debido a la complejidad que introduce la dependencia temporal en los datos, la literatura sobre este tipo de enfoques es relativamente limitada. En este sentido, el presente trabajo busca contribuir al estudio y comprensión de estas metodologías.

Si el proceso es *estrictamente estacionario*, la autocovarianza depende únicamente de la distancia p entre observaciones y no del tiempo absoluto t , por lo que se simplifica como:

$$\gamma_p = E[(x_{t-p} - \mu)(x_t - \mu)] \quad (2.4)$$

Esto implica que la relación entre los valores de la serie solo depende del desplazamiento temporal p y no del instante específico en que se mide, propiedad característica de los procesos estacionarios en covarianza. En particular, cuando $p = 0$, la autocovarianza coincide con la varianza de la serie:

$$\gamma_p(t, t) = \text{Var}(x_t) = \sigma_t^2 \quad (2.5)$$

Dado que las funciones de autocovarianza tienen como unidad de medida el cuadrado de la medida de la serie, es útil contar con una medida adimensional de la dependencia lineal entre las observaciones. Para ello, se utiliza el *coeficiente de autocorrelación*, que se define como la autocovarianza normalizada por la varianza de la serie.

Definición 2.1.7 (Coeficiente de autocorrelación). *Sea x_t una serie temporal con función de autocovarianza $\gamma_p(t, t - p) = \text{Cov}(x_t, x_{t-p})$, donde p es el desfase temporal.*

El coeficiente de autocorrelación en el retardo p se define como:

$$\rho_p(t, t - p) = \frac{\text{Cov}(x_t, x_{t-p})}{\text{Var}(x_t)} = \frac{\gamma_p(t, t - p)}{\gamma_0}, \quad (2.6)$$

Este coeficiente toma valores entre -1 y 1 . A continuación lo demostramos.

Cuando el proceso es débilmente estacionario, como definimos en 2.1.5, la media es constante, la varianza es finita y la covarianza depende únicamente del retardo p . En este caso, la expresión anterior se simplifica a:

$$\rho_p = \frac{\gamma_p}{\gamma_0}, \quad (2.7)$$

Por la desigualdad de Cauchy–Schwarz (Hardy et al. 1952) y el supuesto de estacionariedad débil, tenemos que:

$$|\gamma_p| = |\text{Cov}(x_t, x_{t-p})| \leq \sqrt{\gamma_0} \cdot \sqrt{\gamma_0} = \gamma_0.$$

Dividiendo ambos lados de la desigualdad por γ_0 , se obtiene:

$$\left| \frac{\gamma_p}{\gamma_0} \right| \leq 1,$$

es decir,

$$-1 \leq \rho_p \leq 1.$$

Valores cercanos a 1 indican una fuerte correlación positiva, un valor cercano a -1 indica una fuerte correlación negativa, y un valor cercano a 0 indica poca o ninguna correlación. Estos coeficientes se suelen visualizar en lo que denominamos ACF (*Función de autocorrelación simple*, ACF por su sigla en inglés, definida en 2.1.8), que es una representación gráfica de los coeficientes de autocorrelación para diferentes valores de p . Por otro lado, tenemos la PACF (*Función de autocorrelación parcial*, PACF por su sigla en inglés, definida en 2.1.9) que nos muestra la relación entre una observación y sus retardos, aislando los efectos de retardos intermedios.

En general, las gráficas ACF y PACF contienen intervalos de confianza que nos ayudan a determinar qué coeficientes son significativos (cualquier retardo que no esté dentro del intervalo de confianza se puede considerar significativo). A través de estos gráficos podemos detectar también patrones estacionales para rezagos que con determinada frecuencia salgan del intervalo de confianza. A diferencia de la función de autocovarianzas, la función de autocorrelación es independiente de la escala de medida de la serie (Rodríguez-Collazo, 2024).

Definición 2.1.8 (Función de Autocorrelación (ACF)). *La función de autocorrelación del proceso estacionario x_t es una medida normalizada de la función de autocovarianzas y se define como:*

$$\rho(t, t+p) = \frac{\text{Cov}(x_t, x_{t+p})}{\sigma_{\epsilon_t} \sigma_{\epsilon_{t+p}}} = \frac{\gamma(t, t+p)}{\sqrt{\gamma(t, t)} \sqrt{\gamma(t+p, t+p)}} = \rho_p = \frac{\gamma_p}{\gamma_0} \quad (2.8)$$

Definición 2.1.9 (Función de Autocorrelación Parcial (PACF)). *La autocorrelación parcial poblacional m -ésima, denotada $\alpha_m^{(m)}$, se define como la correlación lineal entre X_t y X_{t+m} , habiendo eliminado el efecto de las correlaciones lineales intermedias:*

$$\hat{x}_{t+1|t,t-1,\dots,t-m+1} = \alpha_1^{(m)} x_t + \alpha_2^{(m)} x_{t-1} + \dots + \alpha_m^{(m)} x_{t-m+1}$$

La definición 2.1.9 se puede expresar matricialmente como:

$$\hat{X}_{t+1|t,t-1,\dots,t-m+1} = \begin{pmatrix} \gamma_1 & \gamma_2 & \dots & \gamma_m \end{pmatrix} \begin{pmatrix} \gamma_0 & \gamma_1 & \dots & \gamma_{m-1} \\ \gamma_1 & \gamma_0 & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{m-1} & \dots & \dots & \gamma_0 \end{pmatrix}^{-1} \begin{pmatrix} x_t \\ x_{t-1} \\ \vdots \\ x_{t-m+1} \end{pmatrix}$$

Entonces, $\alpha_m^{(m)}$ corresponde a la última coordenada del siguiente vector:

$$\begin{pmatrix} \alpha_1^{(m)} \\ \alpha_2^{(m)} \\ \vdots \\ \alpha_m^{(m)} \end{pmatrix} = \begin{pmatrix} \gamma_0 & \gamma_1 & \dots & \gamma_{m-1} \\ \gamma_1 & \gamma_0 & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{m-1} & \dots & \dots & \gamma_0 \end{pmatrix}^{-1} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_m \end{pmatrix}$$

La autocorrelación parcial $\alpha_m^{(m)}$ puede entonces interpretarse como la correlación existente entre x_{t+1} y x_{t-m+1} , una vez ajustado el efecto de $x_t, x_{t-1}, \dots, x_{t-m+2}$.

En el caso de un proceso de ruido blanco, se espera que las funciones de autocorrelación y autocorrelación parcial muestren valores cercanos a cero, lo que indica que no hay correlación entre los diferentes retardos de la serie, como muestra la figura 2.3.

Consideramos ahora el caso más general de un proceso estacionario, donde el objetivo es predecir un valor futuro x_{n+h} utilizando una combinación lineal de las primeras n observaciones $(x_1, x_2, \dots, x_n)$. Planteamos este problema como la búsqueda del mejor predictor $\hat{x}_{n+h} = a_1 x_1 + a_2 x_2 + \dots + a_n x_n$ donde los coeficientes $a_1, a_2, \dots, a_n$ se seleccionan para minimizar el error cuadrático medio. En procesos más complejos que el ruido blanco, la ACF y la PACF proporcionan información clave sobre qué observaciones pasadas son relevantes para el pronóstico, ayudando a definir qué términos incluir en el predictor para capturar la estructura de dependencia temporal en la serie.

Una manera natural de representar esta estructura de dependencia es me-

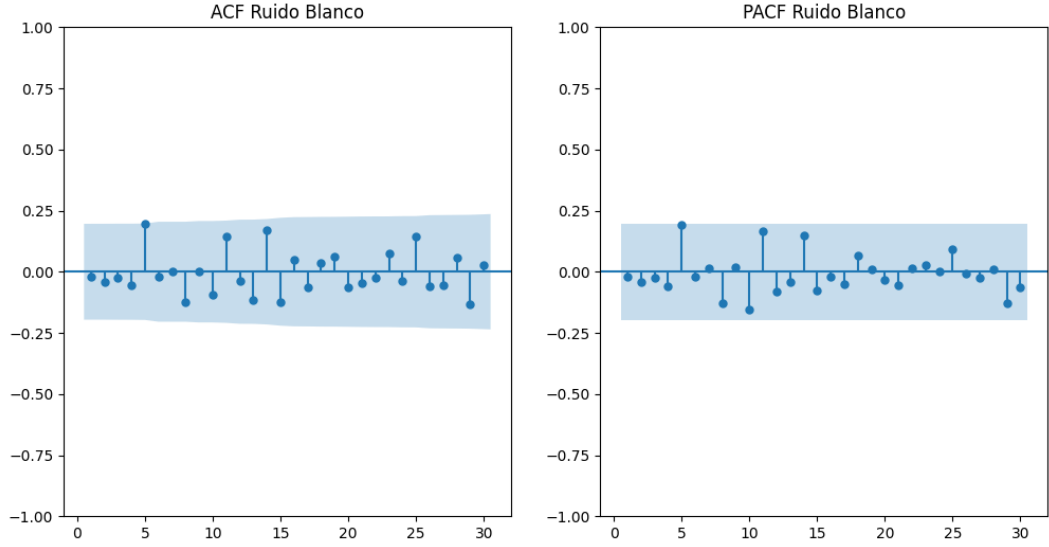


Figura 2.3: ACF y PACF para los primeros 30 retardos del ruido blanco normal simulado en la figura 2.2

diente modelos paramétricos introducidos por Box (2013), como los procesos autorregresivos (AR), de medias móviles (MA) y su combinación en modelos ARMA, que definiremos en la sección 2.1.5 y 2.1.6, respectivamente. Además, para series temporales con componentes estacionales, se utilizan modelos autorregresivos integrados de medias móviles estacionales (SARIMA), que incorporan términos estacionales para captar patrones periódicos. Junto con estos últimos, tenemos también los modelos SARIMAX, que extienden los SARIMA al incluir *variables exógenas* (variables externas al sistema que pueden influir en la serie temporal), permitiendo así modelar la influencia de factores externos sobre la dinámica temporal. Estos enfoques permiten capturar tanto las dependencias de corto como de largo plazo en la serie, ofreciendo herramientas fundamentales para el análisis y la predicción de procesos estacionarios. Los modelos SARIMA y SARIMAX no serán trabajados en el presente estudio.

Los conceptos mencionados constituyen la base del análisis de series temporales y servirán como punto de partida para los modelos de predicción y evaluación que se desarrollan en este estudio. En las secciones 2.1.5 y 2.1.6 se profundizará en estos modelos, explorando sus propiedades, métodos de estimación y aplicaciones en diversos contextos. A continuación, se presentarán las técnicas y criterios para evaluar la estacionariedad de una serie temporal.

2.1.4. Evaluación de estacionariedad

Como se mencionó anteriormente, los modelos clásicos de series temporales asumen que las series analizadas son estacionarias, ya que se basan en las relaciones entre los valores pasados de la propia serie. Por esta razón, antes de aplicar estas técnicas, es importante comprobar si la serie cumple con esta propiedad y, si no lo hace, determinar cuál es la causa de la no estacionariedad. En general, pueden presentarse dos situaciones:

- Tendencia determinística, según lo definimos en 2.1.2.
- *Raíces unitarias*: se presentan cuando el *polinomio autorregresivo* (definido posteriormente en la sección 2.1.5) asociado al modelo tiene una raíz igual a uno. Su presencia en la representación autorregresiva implica que los momentos de segundo orden (como la varianza y la autocovarianza) no se mantienen constantes en el tiempo. Esta propiedad invalida la aplicación de métodos clásicos de inferencia.

En este contexto, se han desarrollado diversas pruebas estadísticas para determinar el grado y tipo de no estacionariedad en una serie. Estas pruebas permiten identificar si el proceso presenta una o más raíces unitarias, si es estacionario en torno a una tendencia determinística, o si cumple con las condiciones de estacionariedad.

Cuando un proceso estocástico presenta una raíz unitaria en su polinomio autorregresivo, se dice que es un *proceso integrado de orden uno*, y se denota como $I(1)$. Aplicando el operador de primera diferencia, definido como $\Delta = (1 - L)$, a una variable con raíz unitaria, se obtiene una nueva serie que es estacionaria en media. En el caso de procesos con más de una raíz unitaria se requiere aplicar el operador de diferencia d veces para alcanzar la estacionariedad, en cuyo caso se dice que el proceso es $I(d)$ (Rodríguez-Collazo, 2024).

Para identificar la presencia de raíces unitarias, se han desarrollado distintos tests, entre los que se destacan las variantes del test de Dickey-Fuller (Dickey y Fuller, 1979), el test de Phillips y Perron (1988), el test ERS (Elliott et al. 1992), y el test KPSS propuesto Kwiatkowski et al. (1992), entre otros.

Adicionalmente, se han propuesto pruebas para detectar cambios estructurales en la serie temporal. Estas comienzan con el trabajo de Phillips y Perron (1988), que considera un momento de cambio exógeno, continúan con el test de Zivot y Andrews (2002), que permite identificar un quiebre de manera en-

dógena, y se amplían con el test de Lumsdaine y Papell (1997), que busca endógenamente dos momentos de quiebre estructural.

Una de las herramientas estadísticas más utilizadas es el *test de Dickey-Fuller Aumentado* (ADF) (Said y Dickey, 1984). Este test es una prueba estadística utilizada para determinar si una serie temporal es estacionaria, evaluando la presencia de una raíz unitaria en un proceso autorregresivo. Se profundiza en más detalle en apéndice 5. En los capítulos siguientes, lo utilizamos para evaluar en datos reales si las series fuente y objetivo que utilizaremos son estacionarias.

En caso de que la serie contenga raíces de módulo 1 en el polinomio autorregresivo estacional, la transformación adecuada es una diferencia estacional, definida como (Rodríguez-Collazo, 2024):

$$\Delta_s = (1 - L^s) \quad (2.9)$$

$$\Delta_s x_t = x_t - x_{t-s} \quad (2.10)$$

Esta transformación implica que hay s raíces de módulo uno.

En las siguientes secciones profundizamos sobre los procesos autoregresivos y de medias móviles antes de adentrarnos en la aplicación de las técnicas de transferencia en simulaciones.

2.1.5. Procesos autoregresivos (AR)

Dados los conceptos anteriores, comenzamos analizando los procesos más simples de series temporales, que son los *modelos autorregresivos* (AR(p)), donde el valor actual de la serie depende linealmente de las p observaciones anteriores. A esto se le incorpora un ruido que abarca toda otra dependencia de x_t , sea con otras variables, o con las mismas variables pero con otro tipo de dependencia no lineal, y lo denotamos como ϵ_t . Solemos también llamar a este término *innovación*.

La ecuación general que describe un proceso AR(p) es:

$$x_t = c + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \epsilon_t \quad (2.11)$$

donde:

- c es una constante;
- $\phi_1, \phi_2, \dots, \phi_p$ son los coeficientes autorregresivos que deben ser estimados, con $-1 < \phi < 1$ (condición necesaria para que el proceso sea estacionario);
- ϵ_t es el término de innovación, que generalmente se asume que sigue una distribución normal $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$.

Si el valor presente depende únicamente del último valor observado, estamos frente a un proceso AR(1). Decimos que una serie x_t sigue un proceso AR(1) si ha sido generada por:

$$x_t = c + \phi x_{t-1} + \epsilon_t \quad (2.12)$$

Para un proceso AR(2), el valor presente depende de los dos valores anteriores, y la ecuación es:

$$x_t = c + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \epsilon_t \quad (2.13)$$

y así sucesivamente para un proceso AR(p), donde el valor presente depende de los p valores anteriores.

En el contexto de modelos autorregresivos (AR), es relevante definir el concepto de *polinomio característico* y *operador de retardo*. El primero es el polinomio asociado a la ecuación de autoregresión del proceso. El segundo, se define como una transformación lineal que desplaza la serie temporal un período hacia atrás y lo llamamos B . Se representa de la siguiente forma:

$$Bx_t = x_{t-1}. \quad (2.14)$$

El polinomio característico se obtiene al representar la ecuación AR en términos del operador de retardo, y es fundamental para analizar la estacionariedad del proceso. Lo definimos como sigue:

$$\phi(B)x_t = \epsilon_t \quad (2.15)$$

Por otra parte, tenemos que la estimación de los parámetros de un proceso AR es muy similar a la estimación por mínimos cuadrados de un modelo de regresión. Supongamos que queremos estimar dichos parámetros por el método de máxima verosimilitud (MV). Tenemos que la función de verosimilitud de forma genérica está dada por:

$$\mathcal{L}(\theta|\vec{x}) = f_{x_1, \dots, x_{T^*}}(\vec{x}|\theta) \quad (2.16)$$

Para esto, necesitamos la función de densidad conjunta y maximizarla respecto a los parámetros θ , siendo estos los parámetros a estimar en la regresión lineal. Retomamos el siguiente razonamiento que plantea Peña (2005) para poder diferenciar el caso de series temporales, donde las variables no son i.i.d., del caso en el que sí lo son. Sabemos que la densidad conjunta de dos variables puede escribirse como una función de la distribución marginal de la primera, multiplicado por la distribución de la segunda condicionada a los valores de la primera como sigue: $f(x, y) = f(x)f(y|x)$.

Consideremos entonces una serie temporal, dada por x_t . Si definimos $x = x_1$ e $y = (x_2, \dots, x_t)$, podemos escribir la función de densidad conjunta como:

$$f(x_t) = f(x_1)f((x_2, \dots, x_t)|x_1) \quad (2.17)$$

Aplicando recursivamente la regla de la probabilidad conjunta a través de factorizaciones en distribuciones condicionales, llegamos a que:

$$f(x_t) = f(x_1)f(x_2|x_1)f(x_3|x_1, x_2)\dots f((x_t)|x_1, \dots, x_{t-1}) \quad (2.18)$$

Esto nos da la posibilidad de escribir la función de verosimilitud de un modelo AR, bajo el supuesto de normalidad conjunta, que nos asegura que todas las distribuciones condicionadas serán normales. Tenemos que la función de verosimilitud va a ser (Peña, 2005):

$$f(x_t) = f(x_1) \prod_{t=2}^{T^*} \sigma^{-1} (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2} \sum_{t=2}^{T^*} (x_t - c - \phi x_{t-1})^2} \quad (2.19)$$

Si la condicionamos a la primera observación, tenemos que la función de soporte tiene la misma expresión que un modelo lineal:

$$f(x_2, \dots, x_{T^*}|x_1) = \prod_{t=2}^{T^*} \sigma^{-1} (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2} \sum_{t=2}^{T^*} (x_t - c - \phi x_{t-1})^2} \quad (2.20)$$

Por lo tanto, suponiendo que $f(x_1)$ es normal con $E(x_1) = \mu$ y $Var(x_1) = \frac{\sigma^2}{1-\phi^2}$ definimos la verosimilitud condicionada, como la asociada a la función 2.20 y el estimador de ϕ como el valor que maximiza dicha función, según se muestra en las expresiones 2.21 y 2.22, respectivamente (Peña, 2005):

$$L_C(\phi, \sigma^2 | x_1) = \frac{-(T^* - 1)}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=2}^{T^*} (x_t - c - \phi x_{t-1}) \quad (2.21)$$

$$\hat{\phi} = \frac{\sum_{t=2}^{T^*} (x_t - \bar{x})(x_{t-1} - \bar{x})}{\sum_{t=2}^{T^*} (x_t - \bar{x})^2} \quad (2.22)$$

donde $\bar{x} = \hat{\mu} = \sum_1^{T^*} x_t / T^*$. Para escribir esta expresión de forma matricial definimos:

$$\mathbf{y} = \begin{bmatrix} x_2 - \hat{\mu} \\ x_3 - \hat{\mu} \\ \vdots \\ x_{T^*} - \hat{\mu} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} x_1 - \hat{\mu} \\ x_2 - \hat{\mu} \\ \vdots \\ x_{T^*-1} - \hat{\mu} \end{bmatrix}$$

Entonces, el estimador presentado en la expresión 2.22 se puede escribir como sigue y es equivalente al estimador MCO para la pendiente de un modelo de regresión lineal simple:

$$\hat{\phi} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \quad (2.23)$$

Los procesos AR no son adecuados para series con memoria corta, donde la autocorrelación cae rápidamente al aumentar el desfase. En estos casos, sólo unos pocos valores pasados influyen en el presente y la autocorrelación decae de forma exponencial (Peña, 2005). Introducimos en la siguiente sección los procesos de medias móviles ARMA(p,q), que incluyen, además de términos autorregresivos, términos de medias móviles.

2.1.6. Procesos de media móvil y ARMA(p,q)

Además de los procesos autorregresivos mencionados, otra clase fundamental de modelos para representar series temporales son los modelos de medias móviles (MA). Mientras que los modelos AR capturan la dependencia temporal en función de valores pasados de la serie, los modelos MA lo hacen en función de perturbaciones pasadas. Tenemos que un proceso MA(q) se define como:

$$\tilde{x}_t = \epsilon_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q} \quad (2.24)$$

donde $\tilde{x}_t = x_t - \mu$.

Estos procesos son siempre estacionarios por ser la combinación lineal de procesos estacionarios.

Desde el punto de vista de la estructura de autocorrelación, los AR permiten muchos coeficientes distintos de cero, pero con una pauta de decrecimiento fija, mientras que los MA permiten unos pocos coeficientes distintos de cero, con valores arbitrarios. Los procesos ARMA intentan combinar las propiedades de ambos tipos de modelos, AR y MA, para representar procesos cuyos primeros q coeficientes son cualesquiera, mientras que los siguientes decrecen según leyes simples (Peña, 2005).

Definimos matemáticamente un ARMA(1,1), que es la estructura más simple, como sigue:

$$\tilde{x}_t = \phi_1 \tilde{x}_{t-1} + \epsilon_t - \theta_1 \epsilon_{t-1} \quad (2.25)$$

Donde $\phi_1 \neq \theta_1$ con $|\phi_1| < 1$ y $|\theta_1| < 1$, para que el proceso sea estacionario e invertible. Si ambos coeficientes son iguales, $\tilde{x}_t = \epsilon_t$, es decir, estamos frente a un ruido blanco. En este tipo de procesos, las autocorrelaciones disminuyen geométricamente y se hacen prácticamente cero a los pocos retardos definido en la definición 2.14. De forma general, utilizando el operador de retardos, definimos un proceso ARMA(p, q) como sigue:

$$\phi(B)x_t = \theta(B)\epsilon_t \quad (2.26)$$

Un resultado particularmente interesante es que, cuando se modelan series que son la suma de varias componentes, como suele ocurrir en muchas aplicaciones reales, y al menos una de ellas tiene una estructura AR, es común que la serie resultante presente una estructura ARMA (Peña, 2005). En la práctica, se observa que muchas series pueden ser bien aproximadas mediante modelos puramente AR o MA.

Por estas razones, en las siguientes secciones de este trabajo optamos por utilizar modelos ARMA como herramienta principal para el análisis de las series temporales simuladas y reales. Además de su capacidad para representar dinámicas complejas, estos modelos ofrecen un alto grado de interpretabilidad.

2.2. Simulaciones

Las simulaciones desempeñan un papel fundamental en la validación y demostración de la efectividad de los métodos estudiados. A diferencia de las aplicaciones en datos reales, donde el proceso generador es desconocido, las simulaciones permiten trabajar en un entorno controlado, lo que facilita el análisis de aspectos teóricos y la evaluación del comportamiento de los modelos bajo distintos escenarios. Este capítulo presenta una serie de simulaciones implementadas en Python que ilustran la aplicación práctica de los conceptos teóricos previamente discutidos en esta tesis.

El objetivo principal de los experimentos es identificar qué características en los conjuntos de datos fuente (X_S) y objetivo (X_T) favorecen la transferencia de conocimiento y mejoran la capacidad predictiva de los modelos. Para ello, compararemos dos enfoques: un modelo entrenado únicamente con un número limitado de observaciones del conjunto objetivo ($y_{\text{baseline}} = f_{\text{baseline}}(X_T)$) y un modelo que incorpora información previamente aprendida del conjunto fuente ($y_{\text{TL}} = f_{\text{TL}}(X_T, X_S)$). Todos los resultados referentes al primero, los denotaremos con el subíndice *baseline*, mientras que los resultados asociados al segundo los denotaremos con el subíndice *TL* (Transfer Learning).

Nos centraremos en modelos estadísticos clásicos, como los procesos AR y ARMA, en lugar de enfoques más complejos como redes neuronales. Esta elección responde a la necesidad de interpretabilidad y al interés en explorar la aplicabilidad de la transferencia de conocimiento en modelos paramétricos bien establecidos. Mientras que la transferencia de conocimiento ha sido ampliamente estudiada en redes neuronales, su aplicación en modelos estadísticos tradicionales sigue siendo un campo poco explorado, lo que convierte este trabajo en una posible contribución relevante.

Para la estimación de los parámetros en el conjunto objetivo y la obtención de los parámetros del conjunto fuente que utilizaremos como inicialización, empleamos la librería de Python *statsmodels* (Seabold y Perktold, 2010), específicamente el módulo *tsa*, especializado en el análisis de series temporales. En particular, utilizamos modelos ARMA para estimar los distintos procesos que analizaremos (AR(1), AR(2), ARMA(1,1)), aprovechando las capacidades de la librería mencionada para ajustar estos modelos de manera eficiente.

La Figura 2.4 ilustra un ejemplo representativo de dos series simuladas: una correspondiente al conjunto objetivo definida como $x_t = 0.8x_{t-1} + \epsilon_T$, y

otra del conjunto fuente dada por $x_t = 0.9x_{t-1} + \epsilon_T$, con tamaños de muestra $N_T = 24$ y $N_S = 300$, respectivamente.

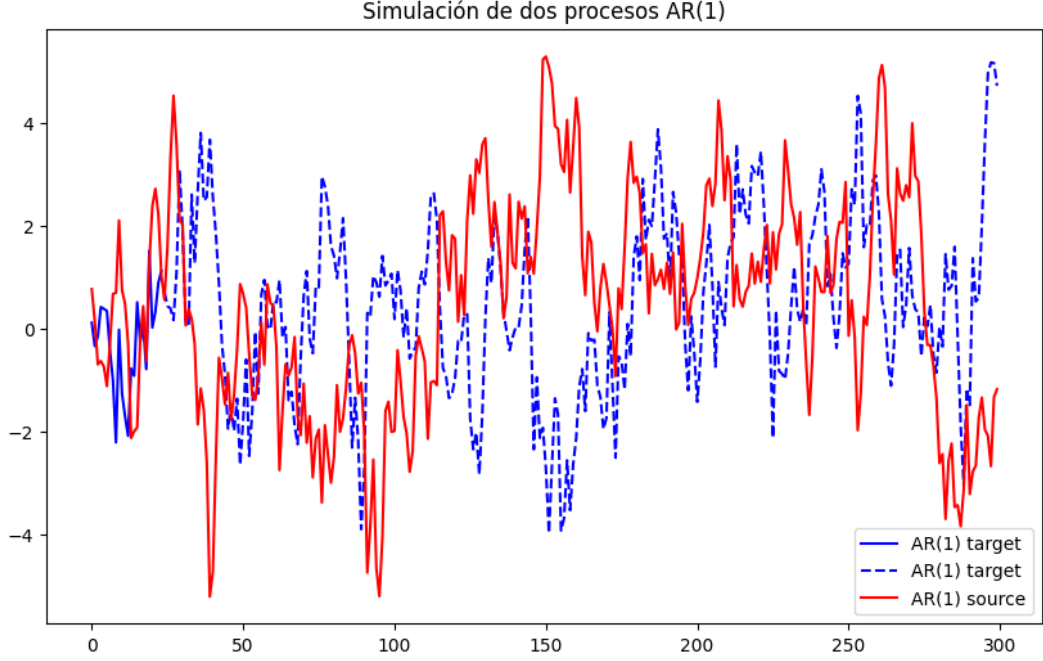


Figura 2.4: Simulación de dos procesos AR(1): $x_t = 0.8x_{t-1} + \epsilon_T$ representada en azul con línea continua y $x_t = 0.9x_{t-1} + \epsilon_T$ representada en rojo. Las observaciones del conjunto objetivo que no tenemos en el momento de entrenar se representan con línea punteada azul.

En la siguiente sección se explicita la metodología utilizada para llevar a cabo los experimentos mencionados.

2.2.1. Metodología

2.2.1.1. Modelos lineales

Para realizar el pasaje de conocimiento nos inspiramos en el enfoque descrito por Obst et al. (2022), donde se propone una metodología de transferencia de conocimiento en un problema de regresión lineal utilizando mínimos cuadrados. En ese trabajo, centrado en la predicción del consumo de demanda energética, los autores utilizan los parámetros estimados a partir del conjunto X_S (con N_S observaciones) como punto de partida para adaptar un modelo sobre el conjunto X_T (con N_T observaciones) mediante descenso por gradiente, donde $N_S > N_T$.

Para realizar la estimación de y_T , los autores se inspiran en el trabajo realizado con redes neuronales y utilizan como metodología fine tuning a través de descenso por gradiente. Esto lo resumimos en el pseudocódigo presentado en el algoritmo 1.

La idea principal en el target es comenzar con $\beta_0 = \hat{\beta}_S = \Sigma_S^{-1} X_S^T Y_S$ y realizar descenso por gradiente con una *tasa de aprendizaje* (*learning rate* en inglés) α intentando minimizar la función objetivo de mínimos cuadrados para el dominio objetivo: $J_T(\beta) = \frac{1}{2} \|Y_T - X_T \beta\|^2$. El proceso se repite durante k iteraciones, hasta alcanzar el número máximo de iteraciones definido (lo denotamos como *iter*) o, en su defecto, hasta que la mejora en la función objetivo sea inferior al umbral de convergencia ϵ . Dicho de otra forma, este proceso pretende incorporar información a las pocas observaciones disponibles del dominio objetivo utilizando la estimación de β del dominio fuente como punto de partida y obtener un estimador $\hat{\beta}_{TL}$ enriquecido. Lo anterior se concentra en qué es lo que se transfiere del dominio fuente al dominio objetivo, lo que se engloba bajo el paradigma de transferencia de parámetros (Mihalkova et al. 2007).

Algorithm 1: Transferencia de parámetros vía descenso por gradiente en el modelo lineal (Obst et al. 2022)

Input: Datos del dominio fuente (X_S, Y_S), datos del dominio objetivo (X_T, Y_T), tasa de aprendizaje α , número máximo de iteraciones *iter*, umbral de convergencia ϵ

Output: Estimador enriquecido $\hat{\beta}_{TL}$

```

1 Calcular  $\beta_0 = \hat{\beta}_S$ ; // Estimación inicial desde el dominio
  fuente por MCO
2 Inicializar  $\beta \leftarrow \beta_0$  y  $k \leftarrow 1$ ;
3 while  $k \leq \text{iter}$  y  $\Delta\beta > \epsilon$  do
4   Calcular el gradiente:  $\nabla J_T(\hat{\beta}_{k-1}) = -X_T^T(Y_T - X_T \hat{\beta}_{k-1})$ ;
5   Actualizar los parámetros:  $\hat{\beta}_k \leftarrow \hat{\beta}_{k-1} - \alpha \cdot \nabla J_T(\hat{\beta}_{k-1})$ ;
6   Calcular la mejora:  $\Delta\hat{\beta} = \|\hat{\beta}_k - \hat{\beta}_{k-1}\|$ ;
7   Incrementar  $k \leftarrow k + 1$ ;
8 return  $\hat{\beta}_{TL} \leftarrow \hat{\beta}_k$ ;

```

Posteriormente en el artículo se estudia cuándo se transfiere, es decir, para qué observaciones $x \in X_T$ es beneficioso utilizar información obtenida previamente en el dominio fuente. Para esto consideran una función de ganancia $\Delta R_k(x)$, donde para un x cualquiera resulta en la siguiente expresión:

$$\Delta R_k(x) = E[(y - \hat{y}_T)^2] - E[(y - \hat{y}_k)^2] = x^T H_k x > 0 \quad (2.27)$$

donde $\hat{y}_T$ es la estimación obtenida utilizando los parámetros del modelo entrenado directamente sobre el dominio target, mientras que $\hat{y}_k$ corresponde a la estimación obtenida mediante transfer, utilizando la metodología presentada.

El criterio que utilizan para comparar ambos predictores es el error cuadrático medio esperado, por lo que si $\Delta R_k(x) > 0$, significa que el error esperado utilizando el predictor con transferencia es menor que el error esperado respecto al predictor entrenado solo en el target, y por ende, para esa observación la transferencia es beneficiosa.

La ecuación 2.27 se simplifica utilizando la matriz H , definida como sigue (Obst et al. 2022):

$$H_k = \sigma_T^2 (\Sigma_T^{-1} - \alpha^2 \Omega_k \Sigma_T \Omega_k) - \sigma_S^2 A^k \Sigma_S^{-1} A^k - A^k B A^k \quad (2.28)$$

donde $A = I_D - \alpha \Sigma_T$, I_D es la matriz identidad de tamaño D , Σ_T^{-1} es la inversa de la matriz de covarianza en el dominio target, $\Omega_k = \frac{1}{\alpha} \Sigma_T^{-1} (I_D - A_k)$ y $B = (\beta_T - \beta_S)(\beta_T - \beta_S)^T$.

La matriz H_k es simétrica y se construye combinando las covarianzas, los parámetros de regresión de los dominios source y target, y los términos de ruido de los datos. En esta construcción, tanto el ruido en los datos como la discrepancia entre los parámetros de regresión determinan de manera fundamental si la transferencia será beneficiosa. Cuando $\|\beta_S - \beta_T\|$ es grande, la transferencia tiende a ser perjudicial. En cambio, si los datos del target son muy ruidosos (σ_T^2 grande), la transferencia puede ser ventajosa, ya que entrenar únicamente con el dominio target puede resultar más difícil.

2.2.1.2. Series de tiempo

En el caso de las series temporales, podemos definir una función de ganancia análoga a la presentada para modelos lineales en la ecuación 2.27. Con fines de simplicidad y claridad en la exposición, desarrollaremos el análisis considerando el caso de un proceso AR(1), presentando una metodología y algoritmo propios.

Dado que en las series temporales cada observación depende de los valores pasados de la propia serie, resulta necesario incorporar esta dependencia

temporal mediante el uso de la esperanza condicional, tal como se muestra a continuación en la ecuación 2.29:

$$\Delta R_k(x_{T-1}) = E[(x_T - \hat{x}_T)^2 | x_{T-1}] - E[(x_T - \hat{x}_k)^2 | x_{T-1}] \quad (2.29)$$

donde $x_T = \phi x_{T-1} + \epsilon_T$, $\hat{x}_T = \hat{\phi}_T x_{T-1}$ y $\hat{x}_k = \hat{\phi}_k x_{T-1}$.

Dado lo anterior, proponemos una metodología para un modelo AR(1) que optimiza el parámetro ϕ mediante descenso por gradiente, utilizando como punto de partida $\phi_0 = \hat{\phi}_S$, donde $\hat{\phi}_S$ corresponde al parámetro estimado sobre el conjunto source mediante el método de máxima verosimilitud. A diferencia del modelo lineal, en el cual la función a minimizar en el descenso por gradiente se basa en el criterio de mínimos cuadrados, en el enfoque propuesto en este trabajo la función objetivo es la log-verosimilitud negativa (Negative Log-Likelihood, en inglés), asociada a la función de verosimilitud 2.19.

$$\begin{aligned} \mathcal{NL}(\phi, \sigma^2) = - & \left[-\frac{T}{2} \log(\sigma^2) + \frac{1}{2} \log(1 - \phi^2) \right. \\ & \left. - \frac{1}{2\sigma^2} \left\{ (1 - \phi^2)(x_1 - \mu)^2 + \sum_{t=2}^T (x_t - c - \phi x_{t-1})^2 \right\} \right] \end{aligned} \quad (2.30)$$

donde $c = \mu(1 - \phi)$ es una constante y $E(x_1) = \mu$. En los procesos que siguen, asumimos $c = 0$.

Por lo tanto, vamos a tener que el gradiente respecto de ϕ va a estar dado por:

$$\nabla \mathcal{NL}(\phi) = \frac{\partial \mathcal{NL}}{\partial \phi} = - \left[\frac{\phi}{1 - \phi^2} + \frac{\phi}{\sigma^2} (x_1 - \mu)^2 + \frac{1}{\sigma^2} \sum_{t=2}^T (x_t - \phi x_{t-1}) x_{t-1} \right]. \quad (2.31)$$

Se realiza el descenso por gradiente con una tasa de aprendizaje α , con el objetivo de minimizar la función presentada en la ecuación 2.30. El procedimiento se ejecuta durante k iteraciones, hasta alcanzar el número máximo de iteraciones definido (*iter*) o, en su defecto, hasta que la mejora en la función objetivo sea inferior al umbral de convergencia ϵ . El parámetro obtenido tras este proceso se denota como $\hat{\phi}_k$, y se utiliza para ubicarse en el instante $t - 1$, calculando la predicción como el producto entre $\hat{\phi}_k$ y x_{t-1} ($\hat{x}_k = \hat{\phi}_k x_{T-1}$). Este

procedimiento se resume en el Algoritmo 2.

Algorithm 2: Transferencia de parámetros vía descenso por gradiente en el modelo AR(1)	
Input: Datos del dominio fuente (X_S), datos del dominio objetivo (X_T), tasa de aprendizaje α , número máximo de iteraciones $iter$, umbral de convergencia ϵ	
Output: Estimador enriquecido $\hat{\phi}_{TL}$	
1	Calcular $\phi_0 = \hat{\phi}_S$; // Estimación inicial desde el dominio fuente por MV
2	Inicializar $\phi \leftarrow \phi_0$ y $k \leftarrow 1$;
3	while $k \leq iter$ y $\Delta\phi > \epsilon$ do
4	Calcular el gradiente para la log-verosimilitud negativa: $\nabla \mathcal{NL}(\hat{\phi}_{k-1})$;
5	Actualizar los parámetros: $\hat{\phi}_k \leftarrow \hat{\phi}_{k-1} - \alpha \cdot \nabla \mathcal{NL}(\hat{\phi}_{k-1})$;
6	Calcular la mejora: $\Delta\phi = \ \hat{\phi}_k - \hat{\phi}_{k-1}\ $;
7	Incrementar $k \leftarrow k + 1$;
8	return $\hat{\phi}_{TL} \leftarrow \hat{\phi}_k$;

Para las simulaciones que siguen, utilizaremos una tasa de aprendizaje $\alpha = 0.001$, un umbral de convergencia $\epsilon = 1 \times 10^{-6}$ y un máximo de iteraciones $i = 200$ para evitar altos costos computacionales. Cabe destacar que la optimización de los parámetros no es el objetivo principal de este trabajo, por lo que no se abordará en estas pruebas y quedará pendiente para trabajo futuro. Sin embargo, se debe tener presente que este último punto puede afectar las conclusiones y el rendimiento de los modelos.

Una vez que tengamos los dos modelos ($M_{baseline}$ y M_{TL}) para cada combinación fuente-objetivo, compararemos la precisión de ambos modelos utilizando el error cuadrático medio (MSE, por sus siglas en inglés), definido como:

$$\text{MSE} = \frac{1}{N_{T'}} \sum_{t=1}^{N_{T'}} (x_t - \hat{x}_t)^2 \quad (2.32)$$

Donde $N_{T'}$ es el número de observaciones del conjunto objetivo que no se utilizan para el entrenamiento, pero que se consideran para realizar las predicciones, cumpliendo que $N_{T'} < N_T$.

Para entrenar el modelo en el conjunto fuente, aplicaremos una división 80-20, destinando el 80% de los datos al entrenamiento y el 20% a la validación. En el caso del conjunto objetivo, la cantidad de datos utilizados para

el entrenamiento dependerá de los experimentos realizados, ya que uno de los aspectos clave que buscamos evaluar es cómo influye el tamaño del conjunto objetivo en la efectividad de la transferencia de conocimiento.

En resumen, primero procedemos a estimar los parámetros de un modelo ARMA con la librería mencionada para cada conjunto de datos fuente, y utilizamos estos parámetros obtenidos para inicializar el descenso por gradiente en los datos objetivo, obteniendo M_{TL} . Por otro lado, entrenamos un modelo ARMA, que denotamos como $M_{baseline}$, directamente en los datos objetivo. Una vez que contemos con ambos modelos, realizamos un análisis para evaluar la efectividad de la transferencia, concluyendo que el transfer es positivo si $MSE_{TL} < MSE_{baseline}$. Finalmente, al repetir este procedimiento para cada combinación de series fuente y objetivo, analizamos las características estadísticas de las series (por ejemplo, sus parámetros, autocorrelaciones y medidas de similitud) y la relación entre ellas, con el fin de responder a la pregunta de cuándo resulta beneficioso transferir conocimiento entre series temporales.

En la siguiente sección presentamos los criterios para comparar y caracterizar las diferentes series de tiempo, y en base a ello concluiremos cuáles de estas características juegan un rol importante en la determinación de la beneficiencia del transfer.

2.2.2. Similitud fuente-objetivo

A partir de los resultados obtenidos, analizaremos distintas propiedades de las series para identificar cuáles influyen en la efectividad de la transferencia de conocimiento. Para asegurar un análisis completo, consideraremos métricas que aborden diferentes aspectos, incluyendo el tamaño de los conjuntos de datos, la estructura y comportamiento de las series, así como sus principales características.

Investigaremos cómo la estructura de las series y la distancia entre ellas afectan la transferencia de conocimiento. Para lo primero extraemos algunas características individuales de las series (ver sección 2.2.2.1) y para lo segundo necesitamos funciones de similitud (ver sección 2.2.2.2).

2.2.2.1. Características individuales de series de tiempo

Para analizar cómo las similitudes en la forma de las series y sus propiedades estadísticas influyen en la transferencia de conocimiento, empleamos el paquete

tsfresh (Christ et al. 2018) para extraer diversas medidas relevantes en términos de escala y forma de las series temporales. En particular, analizaremos las siguientes características:

- Autocorrelación agregada (*agg autocorrelation*)
- Media (*mean*)
- Mediana (*median*)
- Kurtosis (*kurtosis*)
- Oblicuidad (*skewness*)
- Desviación estándar (*standard deviation*)
- Tendencia lineal (*linear trend*)

Para comparar estas propiedades entre los conjuntos fuente y objetivo, seguiremos el procedimiento propuesto en (Ehrig et al. 2024), en donde realizan una investigación estructurada que proporciona medidas de similitud y diversidad para explicar qué características de los datos de origen y destino favorecen el éxito del aprendizaje por transferencia. Evalúan de forma sistemática cómo la similitud entre ambos conjuntos, así como la diversidad del conjunto de origen, afectan los resultados de pronóstico, tanto en escenarios sin entrenamiento previo (zero-shot) como en aquellos con fine-tuning. Analizan estos efectos en términos de precisión, sesgo y estimación de la incertidumbre, utilizando redes neuronales preentrenadas en cinco conjuntos de datos públicos como fuente y aplicándolas al pronóstico de cinco conjuntos de datos objetivo.

2.2.2.2. Medidas de similitud entre series de tiempo

Para cuantificar la distancia entre dos series de tiempo, se emplean funciones o medidas de similitud. Estas son fundamentales para evaluar el grado de relación entre los dominios fuente y objetivo dentro del paradigma de series temporales (Long et al. 2025), siendo uno de los elementos clave en los sistemas de agrupamiento (clustering) y clasificación de series (Serra y Arcos, 2014). Identificar una métrica que capture de manera precisa la similitud entre series y que, además, sea robusta frente a distintos conjuntos de datos, representa un desafío importante.

En años recientes, varios investigadores se han centrado en comparar diferentes medidas de similitud aplicadas a la clasificación de series temporales. Por ejemplo, X. Wang et al. (2013) realizaron un estudio exhaustivo en el

que evaluaron la precisión de clasificación de nueve medidas (junto con cuatro variantes adicionales) sobre 38 conjuntos de datos provenientes de distintos dominios. Sus resultados muestran que, a pesar de haberse propuesto nuevas medidas de similitud con fundamentos teóricos sólidos, en la mayoría de los casos resulta difícil superar métodos clásicos como DTW, que exhibe un rendimiento consistentemente elevado. Además, se destaca la vigencia de la distancia euclidiana debido a su simplicidad y precisión, y se señala que ciertas medidas basadas en distancias de edición, como EDR, también resultan competitivas.

- **Distancia Euclídea:** esta es de las maneras más directas de cuantificar la disimilitud entre dos series temporales es mediante el uso de normas L^n , que se definen como:

$$d_{L^n}(x, y) = \left(\sum_{i=1}^M |x_i - y_i|^n \right)^{1/n} \quad (2.33)$$

donde n es un entero positivo, M representa la longitud de las series temporales, y x_i, y_i son los valores de las series x y y en el instante i . Estas medidas, comparan valores que se encuentran en la misma posición temporal. En (Long et al. 2025) utilizan esta medida para calcular la distancia entre los coeficientes estimados de dos modelos AR. En nuestro caso, como nos interesa la dirección de la diferencia entre los coeficientes, no la utilizaremos para este propósito.

- **DTW (Dynamic Time Warping):** es una métrica que permite evaluar la similitud en la forma de las series temporales. A diferencia de otras medidas como la distancia euclidiana, que comparan valores de manera estrictamente punto a punto, DTW tiene la capacidad de ajustar y alinear series que presentan diferentes patrones de variabilidad temporal, permitiendo identificar similitudes incluso cuando ciertos eventos ocurren en momentos desplazados. Los detalles técnicos de esta medida se desarrollan en el apéndice 5. Para su cálculo, se empleará el paquete estadístico *tslearn* (Tavenard et al. 2020).
- **EDR (Edit Distance for Real Sequences):** es una medida basada en distancias de edición que cuantifica la similitud entre dos series temporales considerando el número mínimo de operaciones (inserciones, eliminaciones o sustituciones) necesarias para transformar una secuencia en otra, adaptada al caso de valores reales. Es, probablemente, la única

medida que logra competir de forma consistente con DTW en términos de precisión, según el estudio mencionado anteriormente. Se presenta un desarrollo más detallado de esta métrica en el apéndice 5. Para su implementación, se empleará el paquete *editdistance* de Python (Yamaguchi, 2022).

- **TWED (Time-Warped Edit Distance)**: es una métrica de similitud que combina elementos de DTW y de las distancias de edición, y representa una de las extensiones más relevantes dentro de los métodos basados en programación dinámica. Por un lado, incluye un parámetro de penalización λ por desajustes, como en las distancias de edición; por otro, incorpora un parámetro de rigidez temporal ν , que regula la elasticidad del alineamiento, similar al enfoque de DTW. Esta combinación permite una mayor flexibilidad al comparar series con variaciones tanto en magnitud como en tiempo. Se detallan los aspectos técnicos de esta medida en el apéndice 5. Para su implementación utilizamos el paquete *pytwed*¹.

En la siguiente sección presentamos los resultados obtenidos a partir de la metodología previamente descrita y se disponen las conclusiones derivadas del análisis de dichos resultados.

2.2.3. Resultados

Consideraremos conjuntos objetivos de tamaños que van desde 24 a 228 observaciones, incrementando de 12 en 12, lo que nos da un total de 18 configuraciones distintas. Para los experimentos propuestos, utilizaremos como objetivo series simuladas bajo un modelo AR(1). Con el propósito de evaluar el impacto de diferentes estructuras en la transferencia de conocimiento, variaremos las series fuente considerando modelos AR(1), AR(2) y ARMA(1,1). Esto nos permitirá analizar cómo las características de las series fuente y su relación con las series objetivo influyen en la efectividad de la técnica.

Se evalúan los desempeños de los modelos base ($M_{baseline}$) y los modelos con transferencia de conocimiento (M_{TL}) para todas las combinaciones posibles, considerando distintas parejas de coeficientes autorregresivos ϕ_S y ϕ_T con el fin de simular series que presentan diversos grados de dependencia. Excluimos

¹<https://github.com/jzumer/pytwed?tab=readme-ov-file>

los casos en los que $\phi_S = \phi_T$ para evitar solapamientos triviales entre fuente y objetivo. En total, realizaremos el análisis sobre 1620 combinaciones de series distintas.

A continuación, se presenta en la tabla 2.1 el resumen tanto de los porcentajes de los resultados obtenidos para cada modelo fuente considerado. Se clasifican los casos en dos categorías: transferencia negativa y transferencia positiva. Sobre el total de casos los experimentos mostraron que, en promedio, 74 % de las combinaciones la transferencia de conocimiento fue positiva, mientras que en el 26 % restante fueron casos de transferencia negativa. Lo que buscamos ahora es comprender si existen características específicas en las series fuente y objetivo, y la relación entre ellas que permitan anticipar cuándo este enfoque de reutilización de aprendizajes previos resultará beneficioso o no.

Tabla 2.1: Porcentajes por modelo fuente y resultado del transfer

Source Model	Transfer Negativo	Transfer Positivo
AR(1)	26.26 %	73.74 %
AR(2)	25.48 %	74.52 %
ARMA(1,1)	25.13 %	74.87 %

En primer lugar, se analiza la relación entre la disponibilidad de datos en el conjunto objetivo y los resultados obtenidos mediante transferencia de conocimiento. La Figura 2.5 ilustra la evolución de los casos de transferencia positiva y negativa en función del número de observaciones en el conjunto objetivo. Cada punto representa la cantidad de casos observados para un valor específico de n_{target} (N'_T), diferenciados según el tipo de transferencia. Se aprecia una tendencia general según la cual, independientemente del valor de n_{target} , los casos de transferencia positiva superan consistentemente a los de transferencia negativa para los tres modelos fuente considerados. Estos resultados indican que el enfoque de transferencia mantiene su efectividad incluso cuando el tamaño de muestra del conjunto objetivo es reducido.

Se realizó un análisis adicional considerando únicamente los casos de transferencia negativa, dado que su menor proporción podría dificultar la visualización de la relación con el número de observaciones. Los resultados presentados en la Figura 2.6 indican que la transferencia negativa no disminuye de manera consistente al incrementarse el tamaño del conjunto objetivo. Asimismo, la cantidad de casos de transferencia negativa varía según el modelo fuente, lo que sugiere que la compatibilidad estructural entre el modelo fuente y el conjunto objetivo influye de forma significativa en la efectividad de la transferencia.

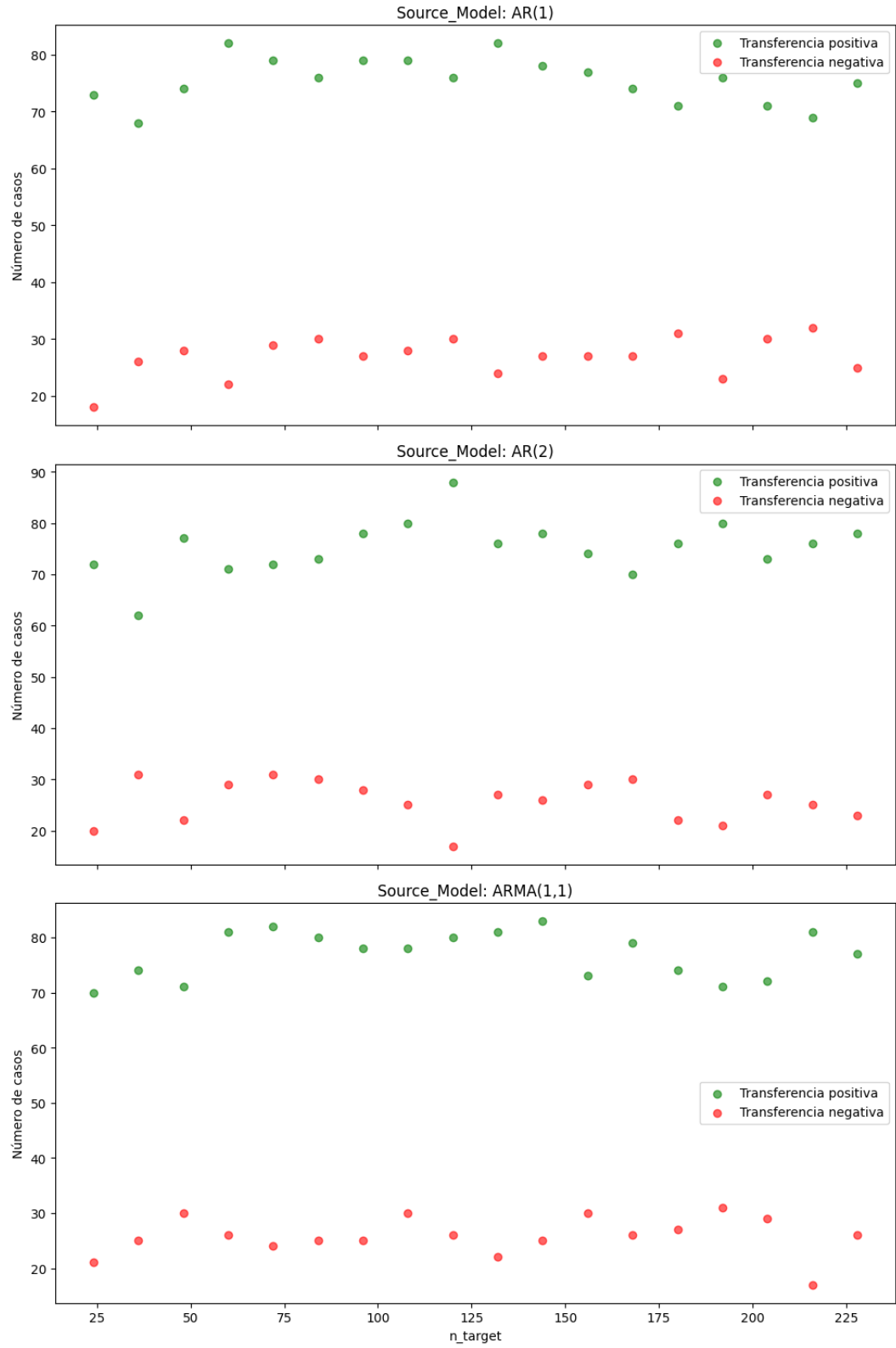


Figura 2.5: Número de casos de transferencia positiva y negativa por tamaño del conjunto objetivo (n_{target}), separados por modelo fuente.

En particular, los modelos AR(1) y AR(2) presentan una mayor variabilidad y un número total ligeramente superior de casos de transferencia negativa en comparación con ARMA(1,1), lo que podría estar asociado a la complejidad relativa entre los modelos fuente y destino.

En el modelo ARMA(1,1) especialmente, se observa una leve tendencia positiva, donde a medida que aumenta el número de observaciones, el número de casos en los que la transferencia es negativa aumenta. Este comportamiento tiene sentido desde un punto de vista estadístico. Cuando el conjunto objetivo es pequeño, el modelo entrenado en el conjunto fuente puede proporcionar una estimación inicial más robusta, reduciendo la varianza y ayudando a estabilizar los coeficientes del modelo. En cambio, cuando hay suficientes observaciones en el objetivo, el modelo puede estimar sus propios parámetros con mayor precisión sin depender de la información del fuente.

Además, se calculó el coeficiente de correlación de Pearson entre ambas variables, obteniéndose un valor de 0.13 en promedio, lo que indica una correlación positiva baja y débil entre el resultado de la transferencia y el número de observaciones. En la tabla 2.2 se muestra el cálculo del coeficiente por modelo, donde observamos que para los modelos AR(1).

Source Model	Correlación Transferencia Negativa n_{target}
ARMA(1,1)	0.0732
AR(1)	0.3670
AR(2)	-0.1668

Tabla 2.2: Correlación entre transferencia negativa y tamaño de target para distintos modelos source

Esto sugiere que, aunque existe una ligera tendencia a que a mayor tamaño de n_{target} aumente el número de casos con transferencia negativa, dicha relación lineal no es significativa ni determinante en el comportamiento observado. En consecuencia, no se puede concluir que el tamaño del conjunto objetivo tenga un efecto fuerte o claro en la probabilidad de resultados adversos en el proceso de transferencia en los casos de estas simulaciones.

En segundo lugar, dado que trabajamos con series simuladas, es posible analizar la relación entre la dependencia temporal de una serie y sus retardos, representada en este caso por los coeficientes autorregresivos ϕ , y cómo esta afecta los resultados de la transferencia. Para ello, consideramos la diferencia entre ϕ_T y ϕ_S en los datos. Esta diferencia permite conocer la dirección de la

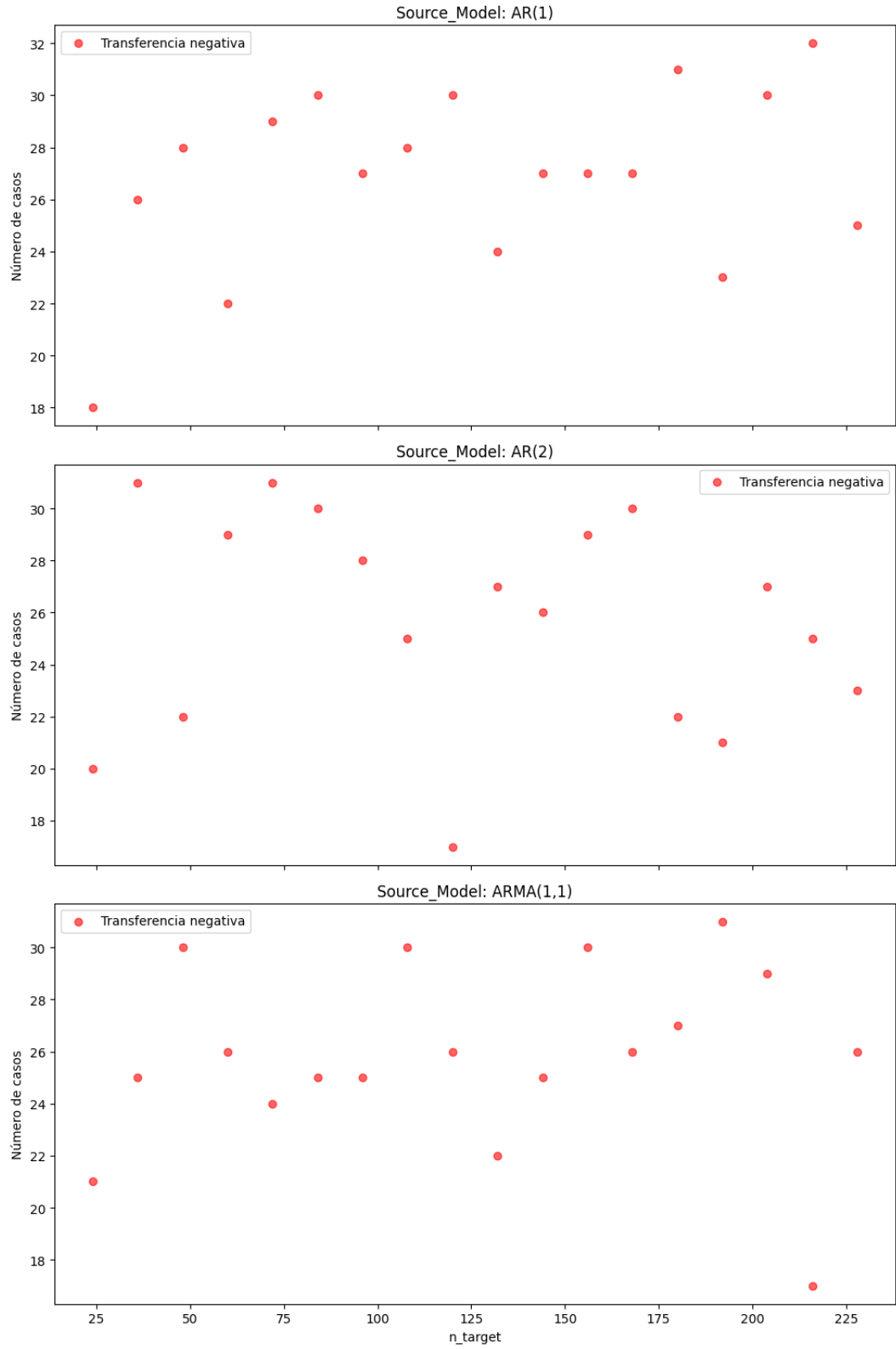


Figura 2.6: Número de casos de transferencia negativa por tamaño del conjunto objetivo (n_{target}), separados por modelo fuente.

variación en la dependencia temporal: cuando es positiva, es decir, $\phi_T > \phi_S$, el conjunto objetivo presenta una mayor dependencia respecto a sus propios valores pasados; cuando es negativa, el conjunto fuente tiene mayor dependencia. Este análisis lo realizamos teniendo en cuenta solamente el primer retardo.

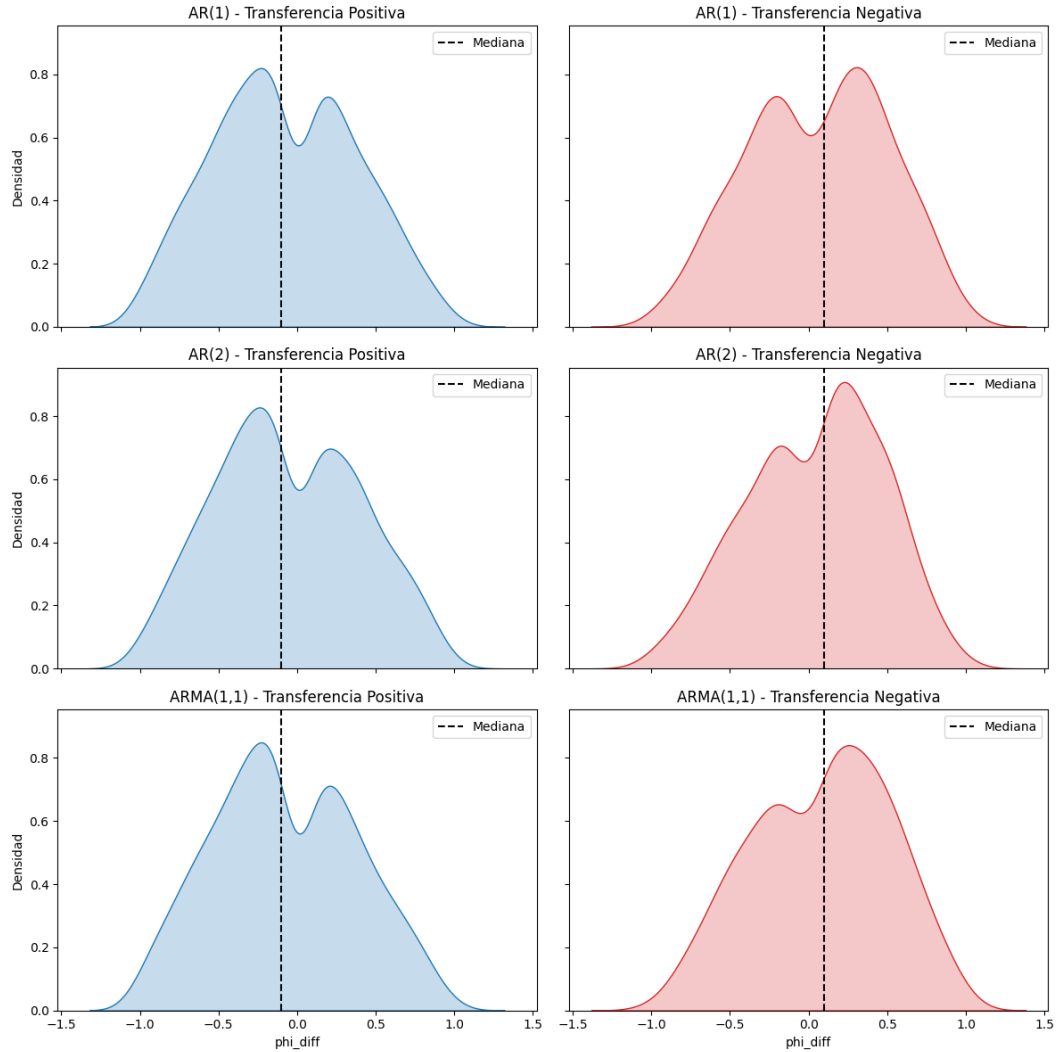


Figura 2.7: Densidades kernel de la diferencia entre los coeficientes ϕ_T y ϕ_S , diferenciadas según el tipo de transferencia: positiva o negativa, por modelo.

La figura 2.7 muestra las distribuciones de densidad del parámetro ϕ_{diff} , definido como la diferencia entre el coeficiente autoregresivo del modelo fuente y el del modelo objetivo. Se presentan por separado los casos de transferencia positiva (paneles de la izquierda) y transferencia negativa (paneles de la derecha), y se analizan los tres tipos de modelos fuente: AR(1), AR(2) y ARMA(1,1).

Observamos en dicha figura una estructura bimodal en las distribuciones de las diferencias de los coeficientes autorregresivos, las cuales difieren notablemente entre los casos de transferencia positiva y negativa. En los casos de transferencia positiva, la distribución se concentra en valores donde $\phi_T < \phi_S$, lo que sugiere que contar con un conjunto fuente con mayor persistencia temporal puede beneficiar la transferencia. Cuando $\phi_T > \phi_S$, es decir, cuando la serie objetivo tiene mayor dependencia temporal que la fuente, la transferencia tiende a fallar, posiblemente porque la estructura más fuerte del target dificulta la generalización de patrones aprendidos en la fuente.

Esta interpretación se ve respaldada por los resultados agregados presentados en la tabla 2.3, donde se observa que la proporción de transferencias positivas es consistentemente mayor cuando $\phi_{\text{diff}} \leq 0$ para todos los modelos evaluados (AR(1), AR(2) y ARMA(1,1)). Por ejemplo, en el modelo ARMA(1,1), el porcentaje de transferencia positiva aumenta de 59.6 % a 80.3 % al pasar de $\phi_{\text{diff}} > 0$ a $\phi_{\text{diff}} \leq 0$, mientras que la transferencia negativa disminuye de 27.7 % a 19.3 %. Resultados similares se observan en los otros modelos.

Esto sugiere que un modelo entrenado en una serie con mayor dependencia temporal no necesariamente se adapta bien a una con menor dependencia, mientras que la transferencia en la dirección opuesta (de una serie más dependiente a una menos dependiente) parece ser más efectiva. Hasta donde sabemos, este comportamiento no ha sido explorado en la literatura existente, por lo que estos resultados podrían aportar una nueva perspectiva al estudio de la transferencia en modelos autorregresivos.

Tabla 2.3: Porcentaje de transferencias positivas y negativas según el signo de ϕ_{diff} y el modelo fuente

Transferencia	AR(1)	AR(2)	ARMA(1,1)
$\phi_{\text{diff}} > 0$			
Transfer positivo (%)	59.60	60.81	59.60
Transfer negativo (%)	26.87	26.46	27.68
$\phi_{\text{diff}} \leq 0$			
Transfer positivo (%)	77.68	75.96	80.30
Transfer negativo (%)	22.02	20.30	19.29

Estos nos permite concluir que la diferencia en la estructura dinámica entre el modelo fuente y el modelo objetivo, medido en este caso a través de ϕ_{diff} es un factor influyente en el éxito o fracaso de la transferencia de conocimiento.

Específicamente, cuando el modelo objetivo tiene un componente autorregresivo significativamente mayor que el del modelo fuente, es más probable que se produzca transferencia negativa. Se subraya entonces la importancia de considerar la similitud estructural entre modelos al seleccionar la fuente para un escenario de transferencia de conocimiento.

En un entorno simulado, donde se conoce la verdadera dinámica del proceso generador, esta relación es más evidente. Sin embargo, en series reales, identificar estos patrones puede resultar más complejo debido a la presencia de ruido, efectos exógenos no modelados y posibles cambios estructurales en los datos, lo que dificulta evaluar de manera clara cómo la dependencia temporal en el conjunto fuente impacta en la transferencia.

Analizamos en tercer lugar las características estructurales de las series objetivo y fuente. Para llevar a cabo este análisis, se examinarán, para cada una de las características de *tsfresh* que presenten resultados relevantes, gráficos de caja diferenciando entre escenarios de transferencia positiva (panel izquierdo) y transferencia negativa (panel derecho). En ambos casos, los resultados se evalúan tanto en el conjunto fuente como en el conjunto objetivo.

En la figura 2.8 se comparan las desviaciones estándar de las series fuente y objetivo, diferenciando los casos de transferencia positiva y negativa. En los escenarios de transferencia positiva, se observa que la desviación estándar en el conjunto objetivo tiende a ser menor que en el conjunto fuente. Esto podría indicar que los modelos entrenados en un dominio más variable logran generalizar mejor cuando se aplican a un dominio más estable. Es decir, una mayor variabilidad en el conjunto fuente puede forzar al modelo a capturar patrones más robustos, lo que facilita su adaptación a contextos menos ruidosos.

Por el contrario, en los casos de transferencia negativa, la desviación estándar del conjunto objetivo suele ser mayor. Esto sugiere que el modelo, al haber sido entrenado en un dominio más simple o menos variable, no logra adaptarse adecuadamente al nuevo dominio, que presenta mayor complejidad o ruido. En este sentido, la transferencia parece ser más efectiva cuando el modelo es entrenado en contextos más diversos y luego se aplica a contextos más estables, y no al revés.

Por último, con el objetivo de comprender en mayor profundidad qué características de la relación entre fuente y objetivo pueden determinar a priori si la transferencia será beneficiosa o no, calculamos las distancias presentadas en la sección 2.2.2.2: DTW, TWED y EDR. En las figuras 2.9, 2.10 y 2.11

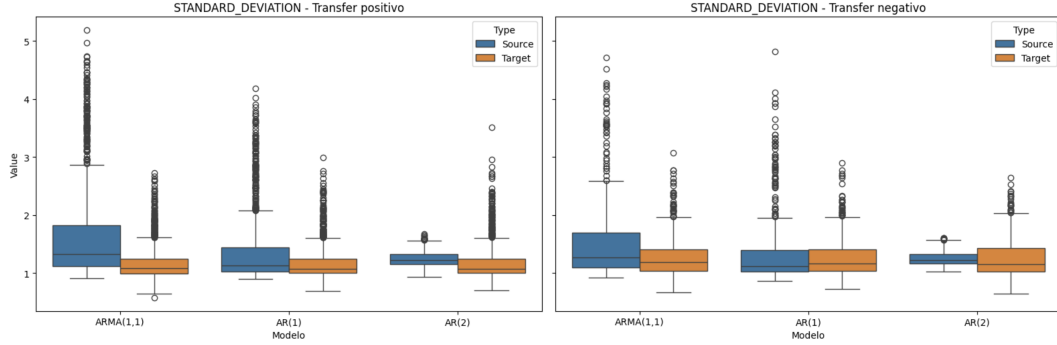


Figura 2.8: Distribuciones de la desviación estándar de los residuos para los modelos ARMA(1,1), AR(1) y AR(2), diferenciando entre conjunto fuente (Source) y conjunto objetivo (Target), en contextos de transferencia positiva (izquierda) y negativa (derecha).

podemos observar las distribuciones de dichas distancias diferenciando por los resultados de la transferencia para los modelos AR(1), AR(2) y ARMA(1,1), respectivamente. En las tres figuras se presentan las distribuciones de distancia normalizada entre pares de series temporales, evaluadas a través de las tres métricas mencionadas. En todos los casos se distinguen dos escenarios: transferencia positiva (curva azul) y transferencia negativa (curva roja).

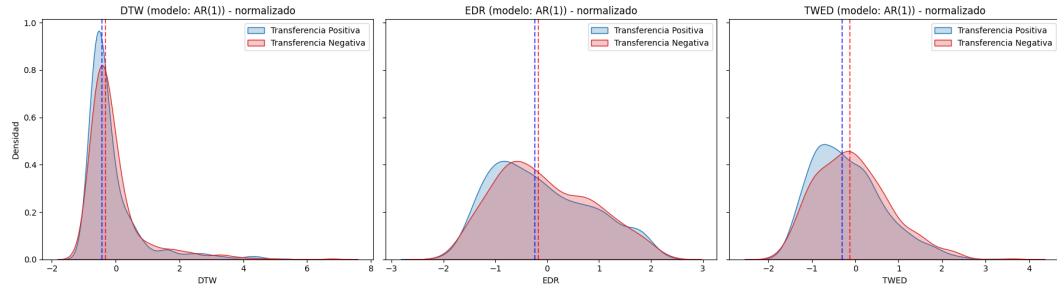


Figura 2.9: Densidades kernel de las distancias DTW, EDR y TWED entre el conjunto objetivo y fuente, diferenciadas según el tipo de transferencia. La visualización corresponde a simulaciones en las que el conjunto fuente y objetivo son ambos generados por un proceso AR(1).

Se observa que, para las tres métricas, las distribuciones asociadas a la transferencia positiva están sistemáticamente desplazadas hacia la izquierda en comparación con aquellas correspondientes a la transferencia negativa, para los tres modelos. Este desplazamiento indica que, en promedio, las distancias entre series que resultan en una transferencia efectiva tienden a ser menores. Además, las medias de las distribuciones (representadas por las líneas punteadas) refuerzan esta diferencia, mostrando que los pares de series más similares

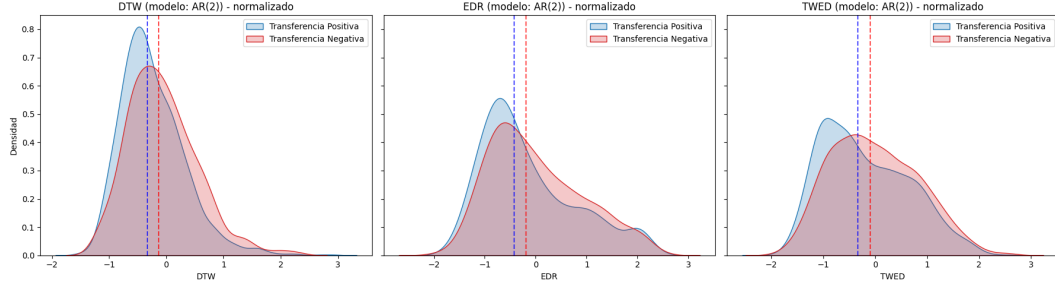


Figura 2.10: Densidades kernel de las distancias DTW, EDR y TWED entre el conjunto objetivo y fuente, diferenciadas según el tipo de transferencia. La visualización corresponde a simulaciones en las que el conjunto fuente y objetivo son ambos generados por un proceso $AR(2)$.

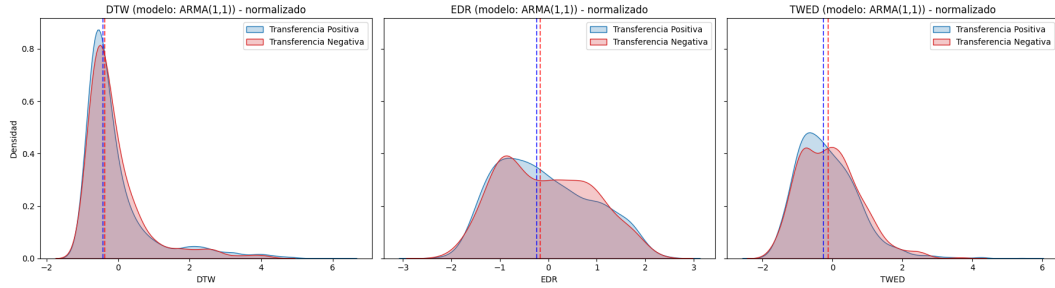


Figura 2.11: Densidades kernel de las distancias DTW, EDR y TWED entre el conjunto objetivo y fuente, diferenciadas según el tipo de transferencia. La visualización corresponde a simulaciones en las que el conjunto fuente y objetivo son ambos generados por un proceso $ARMA(1,1)$.

según estas métricas tienen mayor probabilidad de generar una transferencia positiva.

Este comportamiento sugiere que las métricas de similitud temporal pueden ser utilizadas como indicadores útiles para anticipar el desempeño del transfer learning en contextos de series de tiempo. En particular, podrían servir como criterio de selección de dominios fuente compatibles, lo cual resulta clave para maximizar el éxito de la transferencia. Por tanto, incorporar este tipo de análisis previo puede contribuir a un diseño más robusto y fundamentado de sistemas de transferencia en series temporales. Esa diferencia es más fuerte en el caso del modelo $AR(2)$.

Si bien los resultados sugieren que en los casos de transferencia positiva las distancias entre las series tienden a ser menores, no se identifica un patrón claro que permita utilizar estas métricas como un predictor determinante del éxito de la transferencia. Esto concuerda con hallazgos previos en la literatura, donde estas distancias no siempre resultan suficientes para caracterizar la relación

entre las series en términos de transferencia de información.

2.2.4. Conclusión

El análisis realizado permite identificar las principales características que influyen en el éxito o fracaso de la transferencia de conocimiento entre series temporales simuladas.

En primer lugar, la diferencia en la dependencia temporal entre el conjunto fuente y el conjunto objetivo, definida como $\phi_{\text{diff}} = \phi_T - \phi_S$, se revela como un factor relevante. Los resultados indican que la transferencia resulta más efectiva cuando la serie fuente posee mayor persistencia temporal que la serie objetivo ($\phi_{\text{diff}} \leq 0$). Por el contrario, cuando la serie objetivo muestra una dependencia más fuerte que la fuente ($\phi_{\text{diff}} > 0$), aumentan los casos de transferencia negativa. Esto sugiere que la relación relativa de la dependencia temporal entre fuente y objetivo puede ser un indicador clave para anticipar la efectividad de la transferencia. Hasta donde sabemos, este aspecto no ha sido abordado previamente en la literatura, probablemente porque los modelos de series temporales más complejos no permiten una interpretación directa de la estructura de dependencia, lo que hace que este análisis aporte una perspectiva novedosa sobre los factores que determinan el éxito de la transferencia de conocimiento.

En segundo lugar, el análisis del efecto del tamaño del conjunto objetivo evidenció que, aunque existe una ligera tendencia a que aumente el número de casos de transferencia negativa a medida que n_{target} crece, la correlación es débil y no parece ser determinante. En consecuencia, el tamaño del conjunto objetivo no parece ser un factor crítico para predecir el éxito de la transferencia en estos escenarios simulados.

En tercer lugar, el análisis de las características estructurales de las series, extraídas mediante *tsfresh*, mostró que la variabilidad del conjunto fuente y objetivo también influye en la efectividad de la transferencia. Particularmente, la transferencia positiva se observa con mayor frecuencia cuando el modelo fuente es entrenado en series más variables y se aplica a series objetivo menos ruidosas. Por el contrario, la transferencia negativa tiende a aparecer cuando la serie objetivo presenta mayor complejidad o variabilidad que la fuente.

Finalmente, las métricas de similitud temporal (DTW, TWED y EDR) mostraron que los pares de series que resultaron en transferencias positivas

tienden a ser más cercanos según estas medidas, aunque no se identificó un umbral único que permita predecir con certeza el éxito de la transferencia. Esto sugiere que las métricas de similitud pueden servir como indicadores útiles para la selección de dominios fuente, pero deben complementarse con un análisis de la estructura y características estadísticas de las series.

Los patrones y relaciones identificados en las simulaciones servirán de guía para orientar el análisis de los datos empíricos, permitiendo evaluar si los factores que determinan el éxito de la transferencia de conocimiento en entornos controlados se mantienen en escenarios reales. En la siguiente sección abordaremos la aplicación al IPC de Uruguay.

Capítulo 3

Aplicación: Índice de Precios al Consumo en Uruguay

En la sección siguiente se introducen los datos empíricos que serán utilizados en el presente estudio, sobre los cuales se aplicará la misma metodología empleada en las simulaciones anteriores. En primer lugar, se presentará la motivación de utilizar estos datos y se realizará una descripción detallada de los mismos. Luego, se llevará a cabo un análisis de estacionariedad de las series utilizadas como objetivo y fuente, y evaluaremos diferentes modelos según los criterios presentados en el apéndice 5. Finalmente, se examinarán las características de las series que podrían constituir indicios relevantes para anticipar los resultados obtenidos en el proceso de transferencia de conocimiento.

3.1. Motivación

Tras haber analizado y extraído conclusiones a partir del comportamiento de distintas series simuladas, cuyo proceso generador era conocido, nos enfocamos ahora en el estudio de una situación basada en datos reales. En el análisis de fenómenos socioeconómicos, la discrepancia en la distribución de los datos puede surgir debido a cambios en las políticas económicas, fluctuaciones del mercado, shocks externos o modificaciones en la recopilación de datos. En este contexto, el aprendizaje por transferencia puede ayudar a mitigar el impacto de estos cambios, aprovechando información obtenida en entornos económicos similares. Este enfoque resulta particularmente útil en economías emergentes, donde la disponibilidad de datos puede ser limitada o la dinámica económica

puede cambiar rápidamente, como es el caso de Uruguay.

Cuando se analizan series temporales económicas de distintas regiones, como en este caso para Montevideo y el interior del país, es común encontrar patrones estructurales compartidos, como tendencias, ciclos económicos o respuestas similares ante choques externos (por ejemplo, variaciones en los precios internacionales o fenómenos climáticos). Esta similitud puede ser aprovechada para mejorar la eficiencia en la estimación de modelos econométricos, especialmente cuando una de las regiones presenta una mayor disponibilidad de datos.

En este trabajo, nos centramos en el análisis del Índice de Precios al Consumo (IPC)¹ de Uruguay, una variable macroeconómica fundamental para comprender la evolución del costo de vida, el poder adquisitivo de los hogares y el comportamiento de la inflación. La estructura del IPC refleja tanto la heterogeneidad del consumo regional como la sensibilidad de los precios frente a condiciones locales y globales. Esta complejidad hace que la modelización de su dinámica presente desafíos tanto técnicos como interpretativos.

Uruguay ofrece un contexto empíricamente interesante por su *estructura económica dual*: mientras Montevideo, que representa aproximadamente el 40 % de la población del país², concentra una parte significativa de la actividad económica, el empleo formal y la infraestructura, el interior del país presenta una mayor dispersión geográfica, patrones de consumo diferenciados y, en ocasiones, menor calidad o disponibilidad de datos. Este concepto de economías duales fue introducido por W. A. Lewis et al. (1954), quien planteó la coexistencia de un sector moderno, urbano e intensivo en capital, y otro tradicional, rural e intensivo en trabajo, con menores niveles de productividad. Esta segmentación interna genera asimetrías persistentes que tienen implicancias directas tanto para el análisis económico como para la formulación de políticas públicas.

Desde una perspectiva metodológica, esta heterogeneidad presenta desafíos importantes para el modelado estadístico y econométrico tradicional, particularmente en contextos donde los datos del interior son escasos o ruidosos. En este marco, la transferencia de conocimiento surge como una herramienta prometedora, al permitir aprovechar el conocimiento aprendido en regiones con

¹<https://www.gub.uy/instituto-nacional-estadistica/tematica/ipc-indice-precios-del-consumo>

²<https://www.gub.uy/instituto-nacional-estadistica/censos2023pvh>

mayor disponibilidad de datos (como Montevideo) para mejorar la capacidad predictiva en aquellas con información limitada (Interior).

Aplicado al caso uruguayo, estas técnicas pueden contribuir a reducir las brechas informativas entre regiones, mejorar la precisión de las estimaciones económicas locales y enriquecer el diseño de políticas públicas. A los efectos de este trabajo, se utilizan los datos públicos del IPC mensual publicados por el INE (Instituto Nacional de Estadística), con cobertura desde diciembre de 2010 hasta la fecha más reciente disponible al momento de este estudio (agosto 2025).

La frecuencia de los datos es mensual, y para su cálculo se realiza un relevamiento mensual de los precios de una selección de bienes y servicios, basado en la estructura de consumo de los hogares de las regiones urbanas del país. Esta frecuencia de datos resulta adecuada para el estudio de dinámicas macroeconómicas sin introducir ruido excesivo asociado a fluctuaciones de muy corto plazo.

A continuación, presentaremos los datos con los que vamos a trabajar en el presente capítulo.

3.2. Presentación de los datos

El INE define el IPC como *un indicador del nivel general de precios de los bienes y servicios que los hogares residentes adquieren dentro del territorio nacional y destinan al consumo*¹. El valor del índice en el mes t refleja el nivel de precios de la canasta de referencia en dicho mes expresado como porcentaje respecto al mes base ($t = 0$).

$$I^t = \frac{\sum_{i=1}^N p_i^t q_i^0}{\sum_{i=1}^N p_i^0 q_i^0} \times 100 \quad (3.1)$$

donde I^t es el IPC en el mes t y N es la cantidad de productos de la canasta, p_i^t es el precio del producto i en el mes t y q_i^t es la cantidad del producto i en el mes t .

Esta fórmula corresponde a un índice de precios de tipo Laspeyres (Laspeyres, 1871), que mantiene constantes las cantidades de la canasta de consumo

¹<https://www.gub.uy/instituto-nacional-estadistica/indice-precios-consumo>

(evaluadas en el período base), permitiendo así medir cómo varían los precios a lo largo del tiempo para un patrón fijo de consumo. Desde el punto de vista económico, este tipo de índice refleja el costo relativo de mantener el mismo nivel de vida del período base, aunque no captura posibles sustituciones entre bienes ante cambios de precios.

La variación del IPC a lo largo del tiempo aproxima la evolución del costo de vida en términos monetarios y mide la inflación en la economía. La *inflación* se entiende como el aumento generalizado de los precios en un período determinado, lo que conlleva una disminución del poder adquisitivo de la moneda y una pérdida de su valor real como medio de intercambio y unidad de medida económica.

La variación del Índice de Precios al Consumo entre los meses $t - 1$ y t se puede calcular como:

$$\Delta I^t = \frac{I^t - I^{t-1}}{I^{t-1}} \times 100 \quad (3.2)$$

donde ΔI^t es la variación porcentual del IPC en el mes t , I^t es el IPC en el mes t y I^{t-1} es el IPC en el mes anterior ($t - 1$).

La tasa de inflación, medida a partir del IPC, es uno de los objetivos de la política macroeconómica, y su evolución está estrechamente vinculada a la política monetaria. Además, el IPC se utiliza para deflactar magnitudes monetarias, ajustar unidades de cuenta como la Unidad Indexada y la Base de Prestaciones y Contribuciones, y para indexar ingresos, prestaciones sociales, contratos, activos financieros, tarifas y tramos impositivos.

En Uruguay, se dispone de datos del IPC desde el año 1914 hasta 2025, con mediciones mensuales. Para el análisis de indicadores económicos como el IPC, es común establecer una *base de referencia*, que corresponde a un período específico elegido como punto de comparación para medir las variaciones a lo largo del tiempo. Este período base es fundamental porque, al fijar los valores de los indicadores en un nivel de referencia (generalmente 100), las variaciones futuras del indicador se expresan en relación con ese punto de partida.

En el caso del IPC uruguayo, existen dos fuentes principales para la definición de esta base de referencia. Desde 1914 hasta 1963, el IPC se construyó a partir de estudios sobre las condiciones de vida de la población. A partir de 1963, el índice se calcula en función de la Encuesta Nacional de Gastos e Ingresos de los Hogares (ENGIH), que se realiza cada diez años y proporciona una actualización más precisa de la estructura del consumo de los hogares. La

última encuesta disponible abarcó el período de noviembre de 2016 a noviembre de 2017, y sirve como la base de referencia más reciente para el cálculo del IPC.

Las Encuestas de Gastos e Ingresos de los Hogares tienen como objetivo principal obtener información para actualizar la base del IPC, permitiendo reflejar con mayor exactitud los cambios en los patrones de consumo de la población. La importancia de la base de referencia radica en que permite contextualizar las variaciones del IPC y hacer comparaciones consistentes a lo largo del tiempo. Sin una actualización periódica de la base, los cambios en el comportamiento del consumidor y en la estructura económica no se reflejarían adecuadamente en el índice, lo que afectaría la precisión y utilidad del indicador como herramienta para la toma de decisiones económicas. En el caso de los datos utilizados en el presente trabajo, la base de referencia es octubre 2022.

Es importante destacar que hasta el año 2010, no era posible distinguir entre las distintas regiones del país, ya que el IPC se calculaba únicamente a nivel nacional. La ausencia de un indicador regional impide capturar la heterogeneidad en las estructuras de consumo y en la evolución de los precios entre distintas zonas, lo que hace que las cifras nacionales oculten diferencias significativas en el costo de vida y reduzcan la representatividad del índice. Esta limitación impacta directamente en varios de sus usos habituales en Uruguay, como los ajustes de precios en contratos, el cálculo de la Unidad Indexada, la estimación de salarios reales, la actualización de remuneraciones y otros ingresos, así como la deflación de la variable consumo en el Sistema de Cuentas Nacionales, dado que todos ellos se realizan sobre un promedio nacional que no necesariamente refleja la realidad de cada región, según lo indicado en la nota metodológica ¹. Con la adopción de la nueva base a partir de 2010, se amplió la cobertura geográfica, se incorporaron fuentes de datos adicionales (además de la ENGIH), como el Sistema de Cuentas Nacionales, y se perfeccionó la clasificación por rubros, tal como se detalla en la nota metodológica ¹.

La importancia de esta distinción regional radica en que, sin la desagregación del IPC por zonas, no sería posible llevar a cabo análisis diferenciados entre regiones ni explorar metodologías de transferencia de información o aprendizaje entre ellas. Contar con indicadores desagregados permite estudiar de manera independiente la evolución de los precios en cada región, comparar dinámicas

¹<https://www.gub.uy/instituto-nacional-estadistica/comunicacion/publicaciones/nota-metodologica-indice-precios-del-consumo-base-diciembre-2010100>

locales y aplicar técnicas de modelado o transferencia que aprovechen patrones o similitudes compartidas, algo que sería inviable si solo se contara con un promedio nacional que uniformiza las variaciones regionales. Cabe destacar también que el relevamiento se realiza en zonas urbanas (tanto en el interior del país como en Montevideo), lo cual introduce cierto sesgo en los resultados, pero al mismo tiempo puede generar una mayor similitud entre regiones, facilitando la transferencia de conocimiento.

En la siguiente sección presentaremos la serie de Montevideo y la serie del interior del país, así como sus características que nos serán útiles para el posterior análisis cuantitativo.

3.3. Caracterización de la serie de tiempo

Con el objetivo de analizar las dinámicas inflacionarias a nivel territorial, en esta sección se presentan las series temporales correspondientes a la variación mensual del Índice de Precios al Consumo para Montevideo y el interior del país (datos obtenidos del INE ¹). Esta desagregación permite observar posibles diferencias regionales en el comportamiento de los precios, así como identificar patrones comunes o divergentes a lo largo del tiempo. A través del análisis visual y posterior modelización, se busca caracterizar la presencia de estacionalidad o episodios específicos que puedan estar vinculados a fenómenos económicos o sociales regionales.

Antes de comenzar a trabajar sobre la variación del IPC en ambas regiones, visualizamos en la figura 3.1 la evolución del índice per se, dado que constituye un punto de partida fundamental para el análisis de la inflación y sus variaciones posteriormente. En el gráfico se observa la trayectoria del IPC tanto para Montevideo como para el interior del país, en el período que va desde diciembre de 2010 hasta agosto de 2025. La serie muestra un crecimiento sostenido en ambos casos, lo cual tiene sentido desde el punto de vista económico ya que la inflación (medida empíricamente con el IPC) se define como el aumento generalizado de los precios (Friedman, 1963).

Graficamos ahora en la figura 3.2 las series de la variación del IPC con los datos tomados, para Montevideo e Interior. Como observamos, la evolución de ΔI en ambas regiones del país presenta fluctuaciones significativas a lo largo

¹<https://www.gub.uy/instituto-nacional-estadistica/datos-y-estadisticas/estadisticas/series-historicas-ipc-base-octubre-2022100>

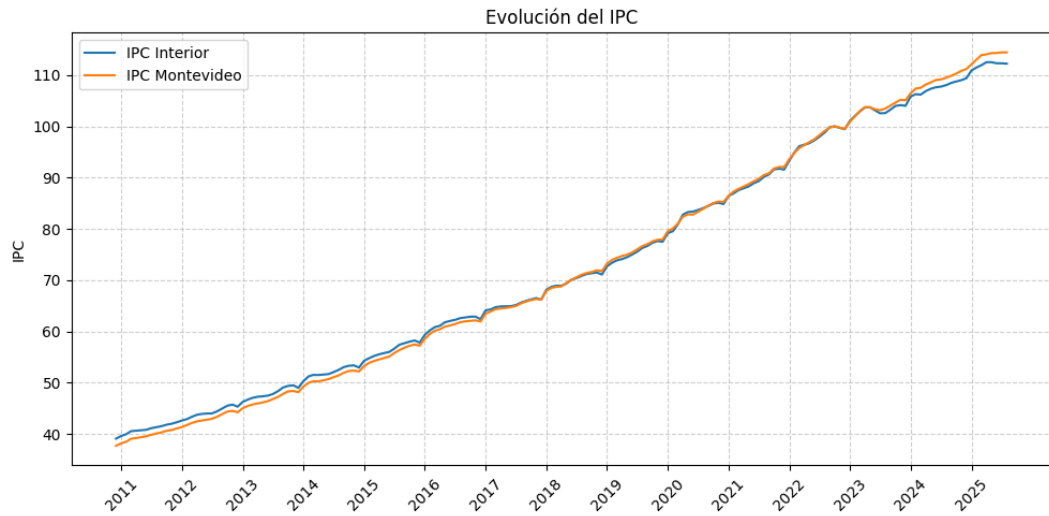


Figura 3.1: Evolución del IPC para Montevideo e Interior (diciembre 2010 - agosto 2025).

del tiempo. No se observa una tendencia clara de crecimiento o disminución sostenida, pero sí existen picos recurrentes que podrían indicar estacionalidad, reflejando ciclos económicos o variaciones en los precios en ciertos períodos del año.

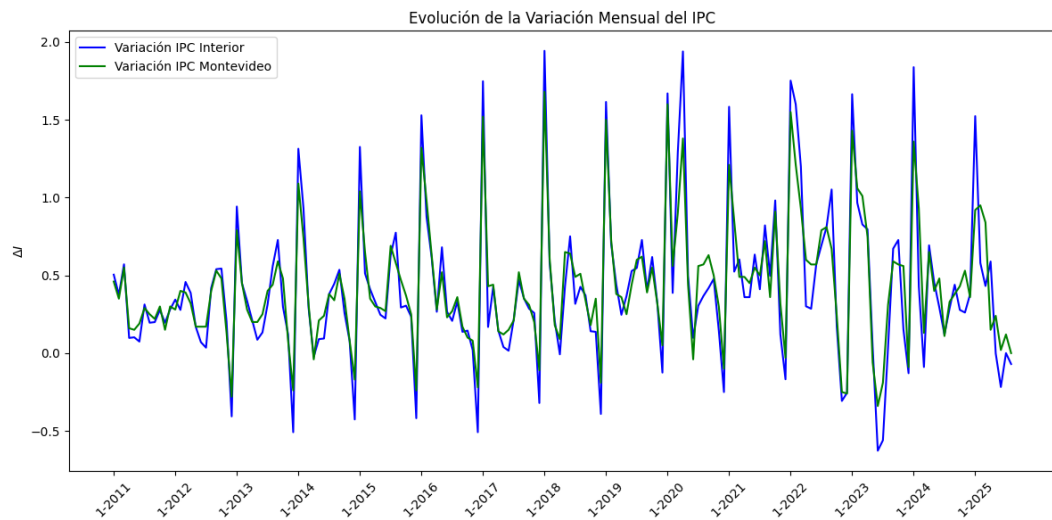


Figura 3.2: Evolución de la variación del IPC (ΔI) para Montevideo e Interior (diciembre 2010 - agosto 2025).

Ambas series parecen oscilar en torno a valores comprendidos entre 0 y 0.5. Esta evolución resulta coherente con lo que habitualmente se observa en la mayoría de las series económicas y sociales, las cuales suelen caracterizarse por

no ser estacionarias y por presentar tendencias persistentes en el tiempo (Peña, 2005). Al realizar una descomposición aditiva de las series en sus componentes de tendencia, podemos observar en la figura 3.3 un patrón creciente durante los períodos 2010-2016 y 2017-2020, mientras que la tendencia se muestra decreciente en los períodos 2016-2017, 2020-2021 y 2022-2025. A pesar de estos períodos de descenso, en el total de la serie se observa un patrón creciente a lo largo de todo el período analizado.

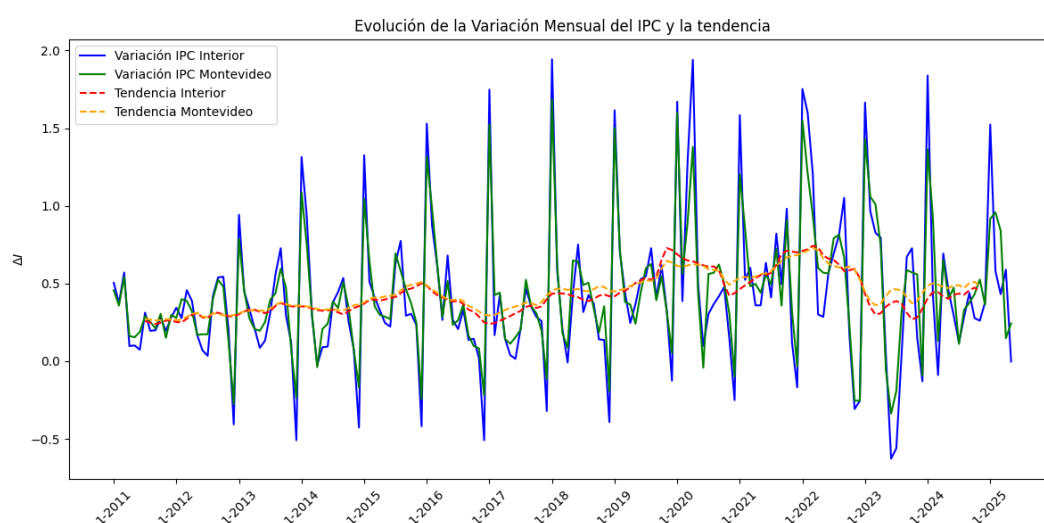


Figura 3.3: Evolución de la variación del IPC (ΔI) para Montevideo e Interior (diciembre 2010 - agosto 2025) y su tendencia.

Desde una perspectiva económica, la evolución observada en las series refleja distintas fases de crecimiento y desaceleración en Uruguay entre 2010 y 2025. Entre 2010 y 2016, la serie muestra una tendencia creciente, coincidiendo con un período de expansión económica impulsado por el buen desempeño de las exportaciones agrícolas y ganaderas y la recuperación global tras la crisis financiera de 2008-2009. A partir de 2016, el precio de las materias primas comenzó a descender a nivel mundial, y se comenzó a recuperar en 2017 ¹. Entre 2017 y 2020, la economía comenzó a desacelerarse debido a la estabilización y caída de los precios internacionales de los *commodities*, lo que redujo el dinamismo económico y moderó la tendencia de crecimiento de la serie. La irrupción de la pandemia de COVID-19 (2019) provocó una contracción de la actividad económica entre 2020 y 2021, afectando la producción, interrumpiendo cadenas de suministro y reduciendo la demanda interna, lo que se reflejó

¹<https://www.gub.uy/presidencia/comunicacion/noticias/instituto-uruguay-xxi-estima-exportaciones-bienes-aumentaran-6-2017>

en una disminución de la variación del IPC durante ese período. Posteriormente, entre 2022 y 2023, eventos climáticos extremos, particularmente una sequía severa, redujeron la producción agropecuaria, incrementando los precios de bienes básicos y generando volatilidad en el Índice de Precios al Consumo, mientras que la actividad económica se mantenía afectada. Finalmente, la serie muestra signos de recuperación hacia 2025, aunque con tendencias todavía volátiles debido a la combinación de shocks externos, eventos climáticos y ajustes post-pandemia.

Con el objetivo de analizar con mayor precisión las particularidades de cada serie, se extrajeron diversas métricas estadísticas de interés, las cuales se muestran en la Tabla 3.1 y fueron presentadas en el capítulo 2.

	Agg Autocorr	Kurtosis	Trend	Mean	Std Dev	Skewness	Var
MVD	-0.014	1.24	0.02	0.44	0.39	0.88	0.15
INT	0.002	1.71	0.11	0.42	0.49	1.04	0.24

Tabla 3.1: Resumen de características `tsfresh` para Interior y Montevideo (diciembre 2010 - agosto 2025).

Desde una perspectiva económica, las diferencias observadas en estas medidas ofrecen indicios relevantes sobre la dinámica inflacionaria en ambas regiones. En primer lugar, la mayor varianza (0.24 frente a 0.15) y desviación estándar (0.49 frente a 0.39) registradas en la serie del Interior reflejan una mayor volatilidad en el comportamiento del IPC, lo cual puede estar vinculado a una mayor sensibilidad de los precios regionales frente a shocks externos o internos, así como una heterogeneidad mayor. La información de precios del interior se releva en 5 localidades diferentes (Ciudad de la Costa, Las Piedras, Maldonado, Salto y Tacuarembó) ¹, lo cual también puede ser una explicación de esta variabilidad.

Asimismo, la curtosis más elevada en el Interior (1.71 frente a 1.24) sugiere una mayor frecuencia de valores extremos, lo que podría reflejar la presencia de episodios puntuales de inflación elevada, por ejemplo, derivados de eventos climáticos adversos o alteraciones en la cadena de suministro, con mayor impacto en regiones menos diversificadas. En este sentido, departamentos como Salto y Tacuarembó, con una marcada especialización en agricultura y ganadería, presentan una vulnerabilidad distinta a la de áreas más urbanizadas y diversificadas como Las Piedras o Ciudad de la Costa. Al estar el IPC del Interior consolidado en un único indicador, estos picos de inflación departamental se

diluyen en el promedio general, aunque igualmente contribuyen a una mayor volatilidad relativa en comparación con el área metropolitana.

Por su parte, la media similar en ambas series, combinada con una asimetría positiva (asimetría cercana a 1.04 en el Interior y 0.88 en Montevideo), indica que los valores tienden a concentrarse por debajo de la media. Esta asimetría positiva refleja que los cambios en los niveles de precio suelen ubicarse en rangos moderados. En cuanto a la tendencia lineal estimada, notablemente más alta en el Interior (0.11 frente a 0.02), esta se corresponde con la observación de una trayectoria inflacionaria creciente en el período analizado, posiblemente influida por factores estructurales como menor acceso a mecanismos de estabilización de precios o una mayor dependencia de productos con precios volátiles (como lo son los commodities).

Además, la métrica de autocorrelación agregada estimada, que resume la dependencia temporal promedio dentro de cada serie, fue ligeramente positiva en Montevideo (0.016) y negativa en el Interior (-0.014). Este resultado refuerza la idea de que los precios en Montevideo tienen una estructura temporal más marcada a lo largo del tiempo, mientras que en el Interior podrían estar más influenciados por factores aleatorios o choques transitorios, coherente con las características de esta zona.

En conjunto, estas métricas respaldan la hipótesis de que existen diferencias estructurales en la dinámica del IPC entre Montevideo y el Interior, lo cual justifica tanto su análisis diferenciado como la aplicación de metodologías que permitan capturar y aprovechar estas heterogeneidades utilizando transferencia de conocimiento. Adicionalmente, tanto lo observado gráficamente en la figura 3.2 como lo discutido en los párrafos precedentes indican que las series presentan un comportamiento marcadamente no estacionario, condición que debe ser corregida para aplicar modelos clásicos de series temporales. Por lo tanto, en la siguiente etapa será necesario aplicar transformaciones que estabilicen la media y la varianza de las series.

3.4. Evaluación de estacionariedad

Cuando el número de series a analizar es pequeño como en este caso, es conveniente realizar la etapa de identificación para comprender y familiarizarse con la estructura dinámica de las series de interés. El objetivo es identificar un conjunto de modelos posibles que sean compatibles con el gráfico de la

serie y sus funciones de autocorrelación simple y parcial. La identificación del modelo requiere identificar la estructura no estacionaria, si existe, y después la estructura ARMA estacionaria (Peña, 2005).

Por ende, las observaciones preliminares de la sección anterior nos habilitan a avanzar hacia un análisis más técnico, en el cual se evalúa formalmente la presencia de no estacionariedad y posibles componentes estacionales mediante el estudio de las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF). Las figuras 3.4 y 3.5 presentan estos gráficos para las series correspondientes a Montevideo y al Interior, respectivamente.

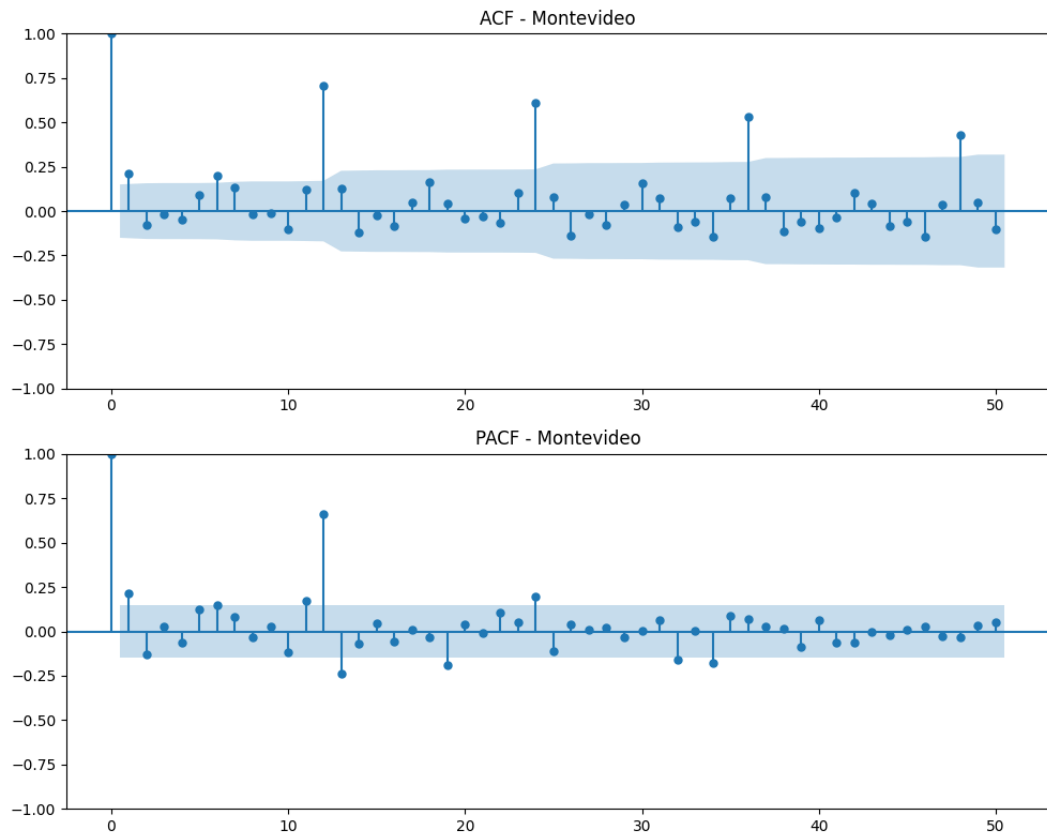


Figura 3.4: Funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) de la variación mensual del IPC para Montevideo.

El análisis de las funciones de autocorrelación y autocorrelación parcial para ambas series refuerza la evidencia de no estacionariedad previamente identificada. En particular, se observa una autocorrelación significativa en múltiples rezagos, especialmente en los múltiplos de 12 en la ACF, lo cual sugiere la presencia de una estacionalidad anual. Por su parte, la PACF muestra un pico destacado en el rezago 1, así como en el rezago 12, lo que indica simultánea-

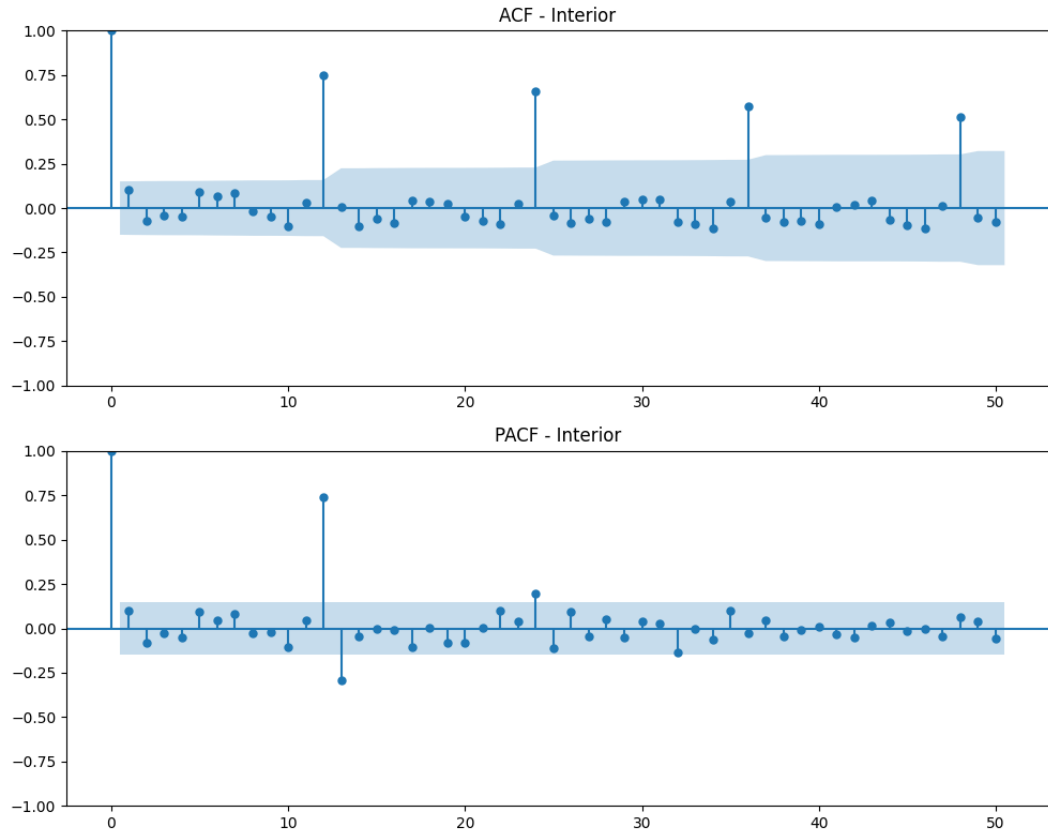


Figura 3.5: Funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) de la variación mensual del IPC para Interior.

mente la existencia de un componente autoregresivo de corto plazo y otro de naturaleza estacional.

Estos patrones son coherentes con la estructura típica de un modelo SARI-MA que incorpora un componente $AR(1)$ junto con una dinámica estacional anual, constituyendo una base adecuada para la especificación inicial de los modelos que se utilizarán en el análisis posterior. La implicancia de esto, es que un modelo que combine un componente autoregresivo simple ($AR(1)$) con un componente estacional anual es el adecuado para capturar la dinámica de estas series. Sin embargo, dado que en nuestro caso como ya hemos estado explorando modelos $AR(1)$, optaremos por desestacionalizar las series y aplicar directamente un $AR(1)$ sobre la serie ajustada, simplificando así la modelización sin perder la capacidad de capturar la dependencia.

La presencia de una fuerte autocorrelación en múltiples rezagos, junto con una clara estacionalidad anual, sugiere que los precios al consumidor no sólo evolucionan de manera persistente a lo largo del tiempo, sino que además están

fuertemente influidos por factores estacionales recurrentes. Esta dinámica es característica de economías donde los precios tienden a ajustarse con cierta rigidez, debido a mecanismos como la indexación salarial, la estacionalidad en la oferta y demanda de ciertos bienes y servicios, y la formación de expectativas inflacionarias basadas en el comportamiento reciente de los precios. En particular, los picos observados en los rezagos 1 y 12 de la PACF pueden interpretarse como el efecto combinado de la inflación reciente y de los ajustes estacionales que se repiten cada año.

Si bien el comportamiento observado en las funciones de autocorrelación y autocorrelación parcial nos da una primera aproximación al diagnóstico de estacionariedad y estacionalidad en ambas series, resulta crucial complementar este análisis con herramientas estadísticas que permitan confirmar dichos patrones de forma más rigurosa. Para ello, utilizamos el test ADF (Said y Dickey, 1984), cuya descripción metodológica se presenta en la sección 2.1.4 y hemos mencionado en el capítulo 2. Los resultados obtenidos a partir de esta prueba se resumen en la tabla 3.2.

	Estadístico	P-Valor	Resultado
Montevideo	-2.22	0.21	No estacionaria
Interior	-2.15	0.22	No estacionaria

Tabla 3.2: Resultados del Test ADF para las series de Montevideo e Interior.

Los resultados del test ADF indican que ninguna de las series, ni la de Montevideo ni la del Interior, cumple con la condición de estacionariedad. En ambos casos, los estadísticos de prueba obtenidos fueron -2.22 y -2.15, respectivamente, mientras que los valores p asociados ascienden a 0.21 y 0.22, superando el nivel de significancia establecido en 0.05. En consecuencia, no se puede rechazar la hipótesis nula de presencia de raíz unitaria, lo que confirma que ambas series presentan características no estacionarias.

Como se explicó en la sección 2.1.4, la condición de estacionariedad es un requisito fundamental para el modelado de series temporales, ya que garantiza que las propiedades estadísticas de la serie, como la media y la varianza, sean constantes a lo largo del tiempo. Dado que hemos comprobado que las series no cumplen con este requisito, es crucial realizar transformaciones que permitan estabilizar sus características estadísticas. La *diferenciación* resulta esencial para eliminar la dependencia temporal y las tendencias presentes en los datos, convirtiendo una serie no estacionaria en estacionaria.

En este trabajo, aplicaremos la diferenciación para suprimir tanto la tendencia como la estacionalidad identificadas en las series. La diferenciación de primer orden, que consiste en calcular las diferencias entre observaciones consecutivas, contribuye a suavizar las series y mitigar tendencias lineales. Adicionalmente, la diferenciación estacional se emplea para eliminar la variación periódica relacionada con la estacionalidad anual detectada, asegurando así que las series estén adecuadamente preparadas para el análisis posterior.

Sea $X_t^{(I)}$ la serie original correspondiente al Interior y $X_t^{(M)}$ la serie original correspondiente a Montevideo. De acuerdo con los resultados presentados en la tabla 3.3, ninguna de estas series es estacionaria, evidenciando la presencia de raíces unitarias. Tras realizar las diferenciaciones requeridas, se obtiene que:

- A la serie del Interior se le aplica una primera diferencia:

$$Y_t^{(I)} = \nabla X_t^{(I)} = X_t^{(I)} - X_{t-1}^{(I)} \Rightarrow X_t^{(I)} = \nabla X_t^{(I)} + X_{t-1}^{(I)}$$

- A la serie de Montevideo se le aplica una primera diferencia:

$$Y_t^{(M)} = \nabla X_t^{(M)} = X_t^{(M)} - X_{t-1}^{(M)} \Rightarrow X_t^{(M)} = \nabla X_t^{(M)} + X_{t-1}^{(M)}$$

Las series transformadas $Y_t^{(I)}$ y $Y_t^{(M)}$ resultan estacionarias. El resultado del test ADF luego de estas transformaciones se presenta en la tabla 3.3. En ambos casos, el estadístico de prueba es menor que los valores críticos al 1 %, 5 % y 10 %. Por lo tanto, se rechaza la hipótesis nula de presencia de una raíz unitaria, concluyéndose que las series no presentan tendencia estocástica y son estacionaria. Presentamos el gráfico de ambas en la figura 3.6.

Serie	Estadístico	P-Valor	Resultado
Montevideo	-9.18	2.2×10^{-15}	Estacionaria
Interior	-8.46	1.53×10^{-13}	Estacionaria

Tabla 3.3: Resultados del Test ADF para las series diferenciadas de Montevideo e Interior.

Se procede a estimar distintos modelos tanto para el conjunto de datos de Montevideo como para el del Interior, con el objetivo de identificar cuál se ajusta mejor a cada serie. A partir de esta evaluación, se seleccionará el modelo óptimo (según los criterios que mencionaremos a continuación) para Montevideo, cuyos parámetros estimados se emplearán posteriormente para

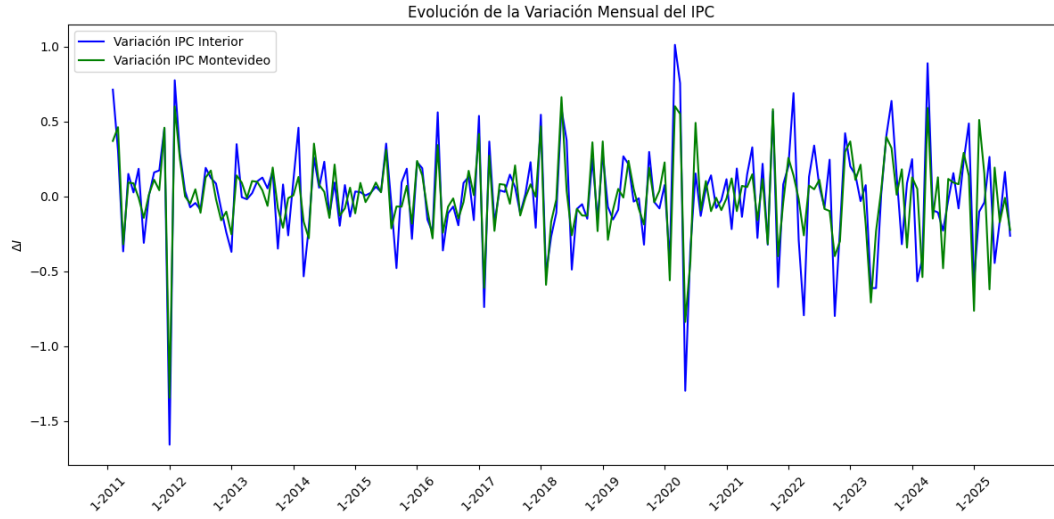


Figura 3.6: Evolución de la variación de la serie del IPC para Montevideo e Interior estacionarias (diciembre 2010 - agosto 2025).

mejorar la estimación de los parámetros en el modelo correspondiente al Interior, aplicando así la estrategia de transferencia de aprendizaje.

3.5. Evaluación de modelos

En esta etapa, se ajustaron modelos ARIMA de orden simple, específicamente AR(1), AR(2) y ARMA(1,1), a ambas series para comparar la calidad del ajuste de modelos con y sin componente de media móvil. Para evaluar los ajustes, se consideraron criterios de información como AIC y BIC, que penalizan la complejidad del modelo y ayudan a seleccionar aquel que mejor explique la dinámica temporal. Además, se empleó la log-verosimilitud para medir la capacidad del modelo de representar los datos observados. Estos resultados permiten identificar qué modelo captura mejor la estructura subyacente de cada serie y fundamentar así las decisiones para la transferencia de aprendizaje.

Los resultados, resumidos en la Tabla 3.4, se obtuvieron a partir de un procedimiento en el que cada serie fue dividida en un conjunto de entrenamiento (80 %) y uno de prueba (20 %). Los modelos fueron ajustados sobre el conjunto de entrenamiento y evaluados en términos de ajuste estadístico (AIC, BIC y Log verosimilitud).

Para la serie de Montevideo, el modelo ARMA(1,1) proporciona el mejor ajuste según los criterios de AIC (-40.57) y BIC (-28.81), incorporando un

componente de media móvil que mejora el ajuste frente a modelos puramente autorregresivos. No obstante, dado que el coeficiente autorregresivo no resulta estadísticamente significativo, se opta por un modelo más simple. En este contexto, el modelo AR(1) constituye una alternativa más parsimoniosa y con un desempeño razonable ($AIC = 10.37$, $BIC = 19.20$), presentando la ventaja de ser más simple, con un solo parámetro, lo que facilita su interpretación y disminuye el riesgo de sobreajuste. Por su parte, el modelo AR(2) incluye un parámetro adicional que permite capturar mayor complejidad y mejorar ligeramente los valores de AIC y BIC, pero a costa de una menor interpretabilidad. Finalmente, para esta serie se selecciona el modelo AR(1) como opción preferente.

En el caso de la serie del Interior, aunque el modelo ARMA(1,1) presenta mejores valores de AIC (41.11) y BIC (52.87), solo el coeficiente autorregresivo resulta estadísticamente significativo, mientras que el coeficiente de media móvil no aporta evidencia relevante. Si bien el modelo AR(1) mantiene simplicidad y facilita la interpretación, se selecciona el modelo AR(2) debido a que mejora significativamente más los criterios de información (AIC y BIC) y la log-verosimilitud, capturando con mayor precisión la dinámica de la serie. Esta elección permite un balance adecuado entre complejidad y ajuste, siendo apropiada para aplicaciones posteriores de transferencia de aprendizaje.

Tabla 3.4: Resultados de ajuste de modelos ARMA para series *MVD* (source) y *INT* (target)

Serie	Modelo	AIC	BIC	LogLik	AR(1)	AR(2)	MA(1)
MVD	ARMA(1,1)	-40.57	-28.81	24.29	×	—	×
MVD	ARMA(2,0)	-4.31	7.45	6.16	✓	✓	—
MVD	ARMA(1,0)	10.37	19.20	-2.19	✓	—	—
INT	ARMA(1,1)	41.11	52.87	-16.55	✓	—	×
INT	ARMA(2,0)	76.63	88.39	-34.31	✓	✓	—
INT	ARMA(1,0)	91.64	100.47	-42.82	✓	—	—

En conclusión, para la serie de Montevideo el modelo AR(1) ofrece el mejor balance entre ajuste y complejidad, presentando mejoras claras en los criterios de información aunque con el término de media móvil no significativo. Para la serie del Interior, si bien tanto el modelo AR(1) como el AR(2) resultan adecuados y preferibles por su simplicidad y por la significancia de los coeficientes autorregresivos, se elige el modelo AR(2) ya que presenta menores valores de AIC, BIC y mayor log-verosimilitud en comparación con el AR(1).

Basándonos en estos modelos seleccionados para cada serie, se procederá a implementar técnicas de transferencia de aprendizaje tal como se ha desarrollado en la sección anterior. Esto permitirá aprovechar la disponibilidad de mejores datos para Montevideo, para entender si podemos mejorar la modelización y predicción de la región de Interior, considerando las características particulares de cada proceso identificado.

3.6. Interior como conjunto objetivo

Una vez analizados los datos, corregida su no estacionariedad y habiendo elegido el modelo que mejor representa el proceso subyacente, procedemos a realizar la transferencia de conocimiento. Para esto, al igual que para el caso de las simulaciones, realizaremos la transferencia del coeficiente autorregresivo desde el modelo *source* al *target*. Procederemos con el modelo AR(1) para la serie de Montevideo, y el modelo AR(2) para la serie de Interior.

El objetivo en esta etapa es evaluar, si los factores que en los experimentos simulados resultaron determinantes para que la transferencia fuera beneficiosa o no, mantienen el mismo impacto cuando se trabaja con series observadas en la práctica. Dentro de estos factores, nos interesa analizar el impacto de: la varianza del conjunto objetivo, el tamaño de este conjunto, la diferencia en la dependencia con los rezagos de ambas series (comparar ϕ_S y ϕ_T) y la distancia entre las series *source* y *target* (utilizando las medidas de DTW, EDR y TWED referenciadas en el anexo 5).

Al igual que en los experimentos con series simuladas, en este caso se variará el tamaño del conjunto de datos de Interior con el fin de representar distintos niveles de disponibilidad de información y analizar cómo esto repercute en los resultados. El tamaño del conjunto se modificará en intervalos anuales (equivalentes a 12 observaciones), abarcando desde un mínimo de 12 observaciones en el conjunto objetivo hasta un máximo de 168. El horizonte temporal utilizado como conjunto de prueba será de 6 observaciones (6 meses), las cuales no se incluyen dentro de las observaciones consideradas para el entrenamiento.

Analizamos a continuación los resultados de la transferencia. En un 71% de los casos la transferencia resultó positiva. Este porcentaje corresponde a que, de las 14 variaciones consideradas para el conjunto objetivo, en 10 de ellas la transferencia mejoró el desempeño del modelo en comparación con la estimación directa, mientras que en las restantes no se observaron beneficios. Es

importante remarcar que, a diferencia del enfoque de Obst et al. (2022), donde la utilidad de la transferencia se mide a nivel de observaciones individuales, en este trabajo la comparación se hace a nivel agregado: se evalúa el desempeño global del modelo en el conjunto objetivo considerando el MSE global.

En primer lugar, evaluamos el impacto del tamaño del dataset objetivo en la efectividad de la transferencia de conocimiento, ya que se ha identificado como un factor clave que influye en los resultados. A partir de los experimentos previos realizados con series simuladas (ver Sección 2.2) y del análisis de la literatura relevante, se ha observado que los beneficios de la transferencia tienden a disminuir a medida que aumenta el número de observaciones en el conjunto de datos objetivo. Este fenómeno sugiere que, cuando se dispone de una cantidad suficiente de datos, el modelo puede aprender de manera más efectiva directamente del conjunto objetivo, reduciendo la dependencia del conocimiento previo transferido.

Esto ocurre para tamaños del conjunto objetivo de 24, 48, 84 y 96 observaciones (meses), correspondientes a las series que abarcan desde diciembre de 2010 hasta noviembre de 2012, noviembre de 2014, noviembre de 2017 y noviembre de 2018, respectivamente. A su vez, para todos estos casos tenemos que $\phi_S < \phi_T$, resaltando que las dinámicas temporales de ambas series difieren. La serie objetivo presenta en dichos casos una dependencia temporal negativa de mayor magnitud, lo que sugiere que sus valores actuales están más influenciados por movimientos inversos respecto al período anterior. Esta diferencia en la estructura temporal puede ser relevante: como se observó también en los experimentos con series simuladas, cuando la serie fuente presenta una dinámica con mayor dependencia temporal absoluta, puede aportar información útil para mejorar la estimación de una serie objetivo más ruidosa mediante transferencia de conocimiento. En caso contrario, la transferencia puede resultar perjudicial.

n_T	ϕ_S	ϕ_T	ϕ_{TL}	MSE_{TL}	MSE_T
24	-0.363	-0.522	-0.422	0.010	0.006
48	-0.363	-0.503	-0.411	0.022	0.020
84	-0.363	-0.576	-0.424	0.192	0.164
96	-0.363	-0.506	-0.406	0.028	0.027

Tabla 3.5: Resultados en los que la transferencia no resulta beneficiosa utilizando Interior como target y Montevideo como source, variando el tamaño del conjunto objetivo.

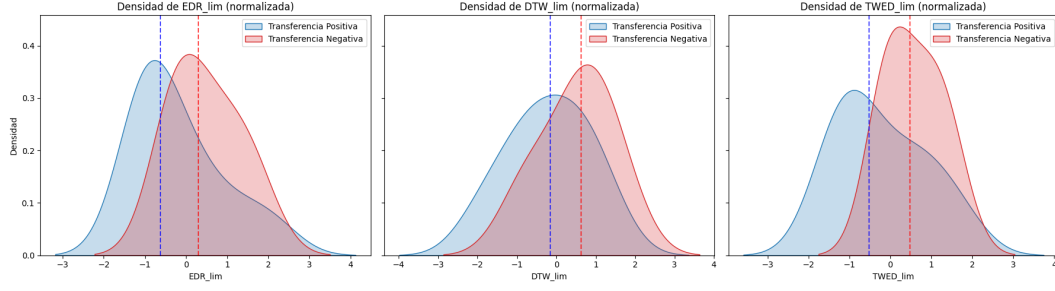


Figura 3.7: Densidades kernel de las distancias DTW, EDR y TWED entre el conjunto objetivo (Interior) y fuente (Montevideo), diferenciadas según el tipo de transferencia.

Respecto a las distancias entre las series, observamos de forma más clara esta relación en la gráfica 3.7. Se comparan las distribuciones normalizadas de tres métricas de distancia entre series temporales (DTW, EDR y TWED), separando los casos según el efecto de la transferencia (si es positiva o negativa). En las tres métricas se observa un patrón consistente, donde los casos de transferencia positiva se concentran en valores de distancia más bajos, mientras que los casos de transferencia negativa se asocian con distancias mayores. Esto indica que la similitud entre las series, evaluada mediante estas tres distancias, podría constituir un buen predictor de la efectividad del transfer learning, confirmando que una mayor semejanza entre las series aumenta la probabilidad de obtener beneficios al transferir conocimiento del modelo source al target.

Desde el punto de vista económico, las observaciones con transferencia negativa coinciden aproximadamente con los años 2014-2018, período en el que Uruguay atravesó distintos shocks económicos y climáticos (mencionados en la sección 3.3) que probablemente afectaron de forma diferencial las dinámicas entre Montevideo e Interior. Estos episodios refuerzan la idea de que la validez de la transferencia no depende solo del tamaño del conjunto de datos del target, sino de cuán similares son las estructuras subyacentes de ambas series en el tiempo considerado. En contextos donde existe una correlación fuerte y estable entre los procesos económicos del Interior y Montevideo, la transferencia puede reducir el error e incrementar la estabilidad del modelo estimado. No obstante, cuando las condiciones económicas divergen, ya sea por shocks asimétricos, políticas públicas diferenciadas o ciclos productivos desincronizados, la utilidad del aprendizaje transferido disminuye drásticamente, incluso induciendo sesgos al modelo objetivo.

3.7. Conclusión

En la tabla 3.6 se resume de manera cuantitativa las condiciones bajo las cuales el transfer learning resultó positivo y negativo. Se observa que los casos de transferencia positiva tienden a asociarse en este caso con series relativamente más largas (mediana de $n_T = 114$) y con valores menores de ϕ_T , lo que indica cierto nivel de dependencia temporal que puede ser aprovechado por el modelo fuente. Sin embargo, este hallazgo difiere parcialmente de lo reportado en la literatura, donde series más cortas a veces presentan mayores beneficios al transferir conocimiento. En contraste, los casos de transferencia negativa presentan series más cortas (mediana de $n_T = 42$) y mayor magnitud en ϕ_T , reflejando series con mayor dependencia temporal, lo que dificulta la transferencia de conocimiento. Además, las métricas de distancia muestran valores medianos más altos en los casos negativos, reforzando que la similitud entre series es un factor importante.

En concreto, los casos en que la transferencia no resultó beneficiosa para la aplicación particular de este trabajo, coinciden con períodos atravesados por shocks externos que pueden haber afectado de manera diferencial a Montevideo y al Interior. Entre 2012 y 2017 se registraron sequías y caídas en los precios internacionales que golpearon con mayor intensidad al sector agropecuario, concentrado en el Interior, generando dinámicas distintas a las de la capital. Estos episodios ilustran que la transferencia pierde efectividad cuando las estructuras económicas se desacoplan por choques externos, aun cuando el tamaño del conjunto objetivo pudiera sugerir lo contrario.

En resumen, esta evidencia empírica respalda la idea de que el aprendizaje por transferencia puede ser especialmente valioso cuando la serie objetivo es más volátil o contiene menos estructura, ya que una serie fuente más estable puede actuar como guía para modelar patrones latentes. Sin embargo, su efectividad es multifactorial y depende críticamente de la similitud estructural entre ambas series. En presencia de shocks externos como sequías, caídas de exportaciones o desacoples regionales, que afectan de forma distinta a Montevideo y al Interior, el valor de la transferencia puede incluso volverse negativo. Por tanto, una correcta aplicación del transfer learning en series temporales económicas requiere no solo evaluar la cantidad de datos disponibles y la similitud entre los datos source y target, sino también considerar el contexto macroeconómico y la relación entre las dinámicas regionales. Integrar esta mi-

rada cualitativa al análisis cuantitativo resulta clave para decidir cuándo y cómo aplicar transferencia de conocimiento de forma efectiva.

Tabla 3.6: Resumen cuantitativo de la efectividad del transfer learning (transferencia positiva y negativa) utilizando Montevideo como fuente e Interior como objetivo.

	Transfer Positivo	Transfer Negativo
Número de casos	10	4
% de casos	71 %	29 %
Mediana DTW	24.694	24.694
Mediana EDR	84.0	106.0
Mediana TWED	75.59	125.17
Mediana $\hat{\phi}_T$ ($\hat{\phi}_S = -0.36$)	-0.47	-0.54
Mediana n_T	114	42

Una extensión metodológicamente interesante de este enfoque consiste en utilizar series temporales de otros países con estructuras económicas similares, como economías pequeñas y abiertas de América Latina, como series fuente para enriquecer la estimación de la dinámica inflacionaria a nivel nacional. Este tipo de transferencia de conocimiento permitiría capitalizar la información contenida en trayectorias inflacionarias de países comparables que enfrentan choques macroeconómicos comunes, tales como fluctuaciones en los precios internacionales, políticas monetarias regionalmente alineadas o eventos climáticos compartidos.

En particular, sería interesante identificar regiones o países cuya estructura productiva, estacionalidad o sensibilidad a shocks externos se asemeje a la del agregado nacional para mejorar las estimaciones. Esta extensión no solo profundizaría la utilidad práctica del enfoque de transferencia en el análisis económico, sino que también aportaría evidencia empírica sobre el grado de integración macroeconómica entre territorios, con potencial para informar el diseño de políticas económicas diferenciadas y más adaptativas a nivel región.

En el siguiente capítulo exploramos otras técnicas más novedosas para el análisis de transferencia de conocimiento en el marco de series de tiempo. Particularmente, exploraremos el uso de modelos de lenguaje de gran escala para evaluar patrones complejos y relaciones no lineales entre series, con el objetivo de mejorar la predicción y la efectividad del transfer learning en contextos económicos, también aplicado al IPC uruguayo.

Capítulo 4

LLM para Series de Tiempo

En capítulos anteriores se presentaron técnicas clásicas para el análisis de series temporales. En este capítulo, en cambio, nos enfocaremos en la aplicación de modelos de lenguaje a gran escala (LLM, por sus siglas en inglés) a este tipo de datos. Se mostrará cómo los avances en aprendizaje profundo, y en particular el desarrollo de estos modelos, han ampliado el enfoque y las capacidades para el análisis de datos temporales. Por último, se realizarán experimentos básicos utilizando los datos de IPC para Uruguay, referenciados en el capítulo 3.

4.1. Introducción

Con el crecimiento exponencial en la disponibilidad de datos y el aumento en la capacidad de cómputo, surgieron nuevos enfoques basados en aprendizaje automático, como las Máquinas de Soporte Vectorial (Cortes y Vapnik, 1995) o XGBoost (T. Chen y Guestrin, 2016), para resolver problemas de series de tiempo. Estos modelos ofrecieron mejoras en precisión y versatilidad frente a los métodos estadísticos clásicos, convirtiéndose en herramientas ampliamente adoptadas. Modelos clásicos como ARIMA (Peña, 2005), ampliamente utilizados por su interpretabilidad, se apoyan únicamente en información endógena, es decir, en los valores históricos de la propia serie. Esta perspectiva resulta limitada en escenarios reales, donde los fenómenos suelen estar influenciados por múltiples variables. La inclusión de información exógena, proveniente de otras series o fuentes, permite capturar interacciones relevantes que los modelos clásicos vistos anteriormente no consideran.

En las últimas décadas, la inteligencia artificial ha impulsado una transformación aún más marcada en el análisis de series temporales, especialmente a través del aprendizaje profundo (Lim y Zohren, 2021). Arquitecturas como las redes neuronales recurrentes (RNN), redes convolucionales aplicadas a secuencias (CNN) y, más recientemente, los modelos basados en Transformers, han demostrado una notable capacidad para modelar relaciones no lineales y dependencias de largo plazo. Estos últimos se benefician especialmente del mecanismo de atención, que aprende de forma automática la conexión entre elementos de una secuencia y permite a la red concentrarse en las partes más relevantes de la entrada para cada paso de predicción, lo que lo hace ideal para este tipo de tareas (Nie et al. 2022, Miller et al. 2024). Estas técnicas han ampliado significativamente las posibilidades de modelado en entornos caracterizados por ruido, estacionalidad y dinámica cambiante, como las finanzas, la meteorología o el ámbito de la salud.

Más recientemente, los modelos de lenguaje a gran escala han emergido como una de las tecnologías más disruptivas del presente. Modelos como GPT-4 (Achiam et al. 2023) y Gemini (Team et al. 2023) no solo han transformado el procesamiento del lenguaje natural, sino que han demostrado una notable adaptabilidad a nuevas modalidades de datos. Entrenados con cantidades masivas de datos, estos modelos han alcanzado niveles de comprensión y generación impresionantes en tareas como traducción automática, análisis semántico, síntesis de información y razonamiento contextual. Esta capacidad ha despertado un creciente interés en explorar su potencial para el análisis de series temporales, abriendo nuevas oportunidades para superar las limitaciones de los métodos tradicionales y los modelos de aprendizaje profundo convencionales.

La motivación principal detrás del uso de modelos de lenguaje a gran escala en series temporales surge de su probado éxito en la modelización de secuencias de texto. Dado que tanto el lenguaje natural como las series temporales comparten una estructura secuencial subyacente, resulta natural explorar si la capacidad de los LLM para capturar dependencias complejas y de largo alcance en texto puede transferirse a tareas tradicionales de análisis temporal, como la clasificación, detección de anomalías o pronóstico. Aunque esta línea de investigación ha ganado tracción recientemente, aún persisten interrogantes claves sobre hasta qué punto los mecanismos que hacen efectivos a los LLM en texto resultan útiles en el contexto de datos temporales, especialmente cuando se trata de dominios como las finanzas o la economía, donde las series presentan

no estacionariedad, alta volatilidad y una señal subyacente difícil de capturar.

En la siguiente sección se presentará una introducción a los modelos de lenguaje, explicando cuáles son sus componentes fundamentales y cómo funcionan, para poder sentar el marco conceptual que permita comprender por qué estas arquitecturas, diseñadas originalmente para el procesamiento de texto, pueden resultar también relevantes en el análisis de series temporales.

4.2. Fundamentos de los modelos de lenguaje

El propósito de esta sección es presentar algunos conceptos fundamentales para comprender el funcionamiento de los modelos de lenguaje, sin entrar en detalle en la arquitectura subyacente que los sustenta.

De manera general, un modelo de lenguaje es un modelo estadístico que tiene como objetivo predecir la próxima palabra, dado un número de palabras anteriores (Almeida y Xexéo, 2019). En tareas de procesamiento de lenguaje natural, es necesario representar a cada unidad textual de forma numérica para que puedan ser trabajadas posteriormente. Para ello, se recurre al proceso de *tokenización*, que consiste en segmentar una secuencia de texto en unidades discretas denominadas *tokens*, que constituyen las unidades mínimas en las que un texto puede dividirse para luego ser procesado.

Cada token puede corresponder a una palabra, una subpalabra o un carácter individual. Inicialmente, a cada token se le asigna un identificador numérico único, que luego es proyectado en un espacio vectorial mediante un mecanismo de embeddings. Un *embedding* constituye una representación continua de dimensión d , que ubica a cada token en un espacio en el que es posible modelar y capturar relaciones semánticas y sintácticas. Para lograr esto, se pueden emplear técnicas más clásicas como Word2Vec (Church, 2017), GloVe (Pennington et al. 2014) o FastText (Bojanowski et al. 2017), que aprenden representaciones distribuidas a partir de grandes corpus de texto. En estos modelos, cada palabra del vocabulario tiene asociado un único vector fijo, independientemente del contexto en el que aparezca.

Más recientemente, los modelos basados en arquitecturas de Transformers (Vaswani et al. 2017) han introducido los llamados *embeddings contextuales*, en los cuales el vector asignado a un token depende de la secuencia completa en la que se encuentra. Esto permite capturar de manera más precisa la polisemia (una palabra puede tener más de un significado) y las dependencias a

largo plazo, generando representaciones dinámicas que se ajustan al significado contextual de cada palabra dentro de la oración. Dentro de los Transformers tenemos distintas arquitecturas de referencia, como BERT (Devlin et al. 2019), GPT (Radford et al. 2018), T5 (Raffel et al. 2020), entre otras.

Una vez que el texto ha sido transformado en tokens y representado mediante embeddings (ya sean estáticos o contextuales), estos vectores constituyen la entrada al modelo de secuencia. En otras palabras, los embeddings funcionan como el nexo entre la representación simbólica del lenguaje y la arquitectura neuronal encargada de procesar la información. Este procedimiento suele implementarse mediante una arquitectura conocida como *encoder-decoder*, la cual se apoya en redes neuronales recurrentes. Estas redes están específicamente diseñadas para el tratamiento de secuencias, ya que incorporan conexiones recurrentes que les permiten retener información de estados previos y modelar dependencias en los datos (Sherstinsky, 2020).

Una arquitectura encoder-decoder está formada por dos componentes principales: la red encoder, encargada de recibir la secuencia de entrada (embeddings) y transformarla en una representación vectorial de dimensión fija, y la red decoder, que utiliza dicha representación para producir la secuencia de salida. Ambos módulos se entrenan de manera conjunta con el objetivo de maximizar la probabilidad condicional de la secuencia de salida dada la secuencia de entrada.

$$\max_{\theta} \frac{1}{N} \sum_{n=1}^N \log p_{\theta}(y^{(n)} | x^{(n)}), \quad (4.1)$$

donde θ representa el conjunto de parámetros del modelo y cada par $(x^{(n)}, y^{(n)})$ corresponde a una pareja (*secuencia de entrada, secuencia de salida*) proveniente del conjunto de entrenamiento. Entre los modelos encoder-decoder más conocidos se encuentran Seq2Seq (Sutskever et al. 2014) y las arquitecturas basadas en Transformers (Vaswani et al. 2017) que son la base de muchos de los LLM.

Durante la fase de generación de texto, el decoder produce la secuencia de salida de forma autoregresiva, es decir, predice un token a la vez condicionado en los tokens previamente generados. En cada paso, el decoder calcula valores denominados *logits*, que corresponden a puntajes no normalizados que indican cuán probable es que cada token sea el siguiente en la secuencia, dado el contexto hasta ese momento, para cada token del vocabulario. Finalmente, los logits se transforman en probabilidades mediante la función softmax, que

convierte estos puntajes sin normalizar en valores entre 0 y 1, que representan la probabilidad.

La estrategia utilizada para seleccionar el siguiente token tiene un impacto significativo en la calidad y características del texto generado (Wiher et al. 2022). Métodos como *beam search* (Freitag y Al-Onaizan, 2017), que buscan maximizar la probabilidad de la secuencia completa, suelen tener buen desempeño en tareas que requieren alta coherencia y precisión, como la traducción automática. Sin embargo, estas estrategias pueden generar resultados repetitivos o incoherentes en tareas más creativas, como la escritura de historias. Por otro lado, técnicas basadas en muestreo, como *top-k sampling* (Holtzman et al. 2019), permiten introducir diversidad en la generación, adaptándose mejor a contextos donde la creatividad y la variabilidad son deseables. Este tipo de técnica, selecciona aleatoriamente la siguiente palabra a partir de las k opciones más probables, en lugar de intentar decodificar texto que maximice la probabilidad. En general, las propiedades de cada estrategia de decodificación no siempre se transfieren de manera uniforme entre diferentes tareas de generación de lenguaje natural, por lo que la elección de la estrategia debe alinearse con los objetivos específicos de cada aplicación.

En la sección que sigue, revisaremos cuáles son los desafíos que los modelos de lenguaje intentan resolver y abordaremos los estudios más destacados que investigan su aplicación en el modelado de series de tiempo. El foco será entender las metodologías empleadas, los principales retos identificados y los avances recientes que definen el estado actual de la disciplina, y motivan el presente estudio.

4.3. Desafíos y motivación

El creciente interés en aplicar LLM al ámbito de las series temporales surge a partir de una serie de propiedades deseables que estos modelos han demostrado en otros dominios (Jin et al. 2023). Entre ellas, destaca su notable capacidad de generalización, evidenciada en contextos de aprendizaje con pocos o incluso ningún ejemplo (few-shot o zero-shot), lo que los vuelve especialmente atractivos para escenarios donde los datos históricos son escasos. Esta eficiencia en el uso de datos contrasta con los métodos tradicionales, que suelen requerir grandes volúmenes de información para lograr buenos desempeños.

Además, los LLM han mostrado habilidades avanzadas de razonamiento y

reconocimiento de patrones complejos, lo cual podría traducirse en predicciones más precisas si se logran aprovechar estos aprendizajes de alto nivel. A medida que estas arquitecturas se vuelven multimodales, también se abre la posibilidad de incorporar información proveniente de diferentes fuentes (como texto o imágenes), enriqueciendo así el proceso de pronóstico (X. Zhang et al. 2024).

Finalmente, una ventaja operativa clave es que estos modelos, una vez entrenados, pueden reutilizarse en nuevas tareas sin necesidad de rediseñar arquitecturas o realizar extensos ajustes de hiperparámetros, lo cual simplifica su adopción frente a los modelos tradicionales.

Sin embargo, su aplicación en series temporales presenta desafíos importantes, especialmente por las diferencias inherentes entre los dominios para los cuales fueron diseñados y la naturaleza de los datos temporales. Mientras que los LLM operan sobre secuencias de tokens discretos, como palabras en lenguaje natural, los datos de series temporales representan fenómenos subyacentes de carácter continuo y toman valores en un espacio numérico también continuo. Esta diferencia estructural introduce una brecha de representación que debe ser abordada mediante técnicas de *cuantización*, segmentación o transformación del espacio continuo al discreto. Dichas transformaciones pueden comprometer la capacidad del modelo para capturar adecuadamente las dinámicas temporales, lo que representa un área crítica de investigación dentro de este campo emergente.

En este sentido, en el artículo *Large Language Models for Time Series: A Survey* (X. Zhang et al. 2024) los autores realizan una revisión de las metodologías empleadas para aprovechar el potencial de los LLM en el análisis de series temporales, destacándose por enfatizar cómo superar la brecha entre datos textuales y series temporales.

Pese a las prometedoras cualidades teóricas de los LLM para tareas de predicción temporal, estudios recientes han comenzado a cuestionar su efectividad práctica en este dominio. En particular, M. Tan et al. (2024) analizan críticamente su utilidad en el trabajo titulado *Are Language Models Actually Useful for Time Series Forecasting?*. En línea con el creciente interés por aplicar estos modelos a series temporales, los autores diseñan experimentos que evalúan de manera sistemática la contribución de cada componente dentro de tres enfoques representativos basados en LLM (OneFitsAll (Zhou et al. 2023), Time-LLM (Jin et al. 2023) y CALF (P. Liu et al. 2025)). Sus resultados muestran que

eliminar el LLM o sustituirlo por una simple capa de atención no sólo no perjudica el desempeño, sino que en muchos casos incluso lo mejora. Asimismo, evidencian que los LLM preentrenados no superan a modelos entrenados desde cero, no logran capturar adecuadamente las dependencias secuenciales propias de las series temporales y no aportan ventajas significativas en entornos con datos limitados.

Es relevante destacar que estos trabajos evalúan su desempeño en distintos conjuntos de datos públicos, incluyendo algunos de tipo económico como series de tipo de cambio. En estos casos, caracterizados por alta volatilidad, los resultados indican que los modelos no basados en LLM tienden a obtener mejores resultados, lo que sugiere que los enfoques actuales aún no superan a métodos más simples en ciertos contextos financieros.

A pesar de que los modelos de lenguaje a gran escala son a menudo percibidos como sistemas de tipo “caja negra” debido a su complejidad y número de parámetros, una de sus ventajas distintivas es su capacidad para generar explicaciones en lenguaje natural (X. Yu et al. 2023). Esta habilidad permite ofrecer interpretaciones coherentes y contextualizadas sobre sus decisiones, lo que puede aportar mayor transparencia en tareas de pronóstico. En particular, mediante técnicas como razonamiento paso a paso (conocido como CoT por su sigla en inglés: Chain-of-Thought), los LLM pueden elaborar explicaciones detalladas que no solo justifican sus predicciones, sino que también ayudan a entender los patrones y relaciones que han sido internalizados por el modelo. Esta propiedad resulta especialmente valiosa en contextos donde la interpretabilidad es clave para la confianza y la adopción de este tipo de herramientas.

Cabe destacar que muchos de los estudios que analizan la aplicación de modelos de lenguaje en tareas relacionadas con series temporales han empleado arquitecturas como GPT-2 (Radford et al. 2019) o las primeras versiones de LLaMA (Touvron et al. 2023). Aunque estos modelos fueron fundamentales para impulsar la revolución del aprendizaje autosupervisado, el campo ha experimentado un avance acelerado en los últimos años. Hoy en día, existen versiones mucho más potentes y versátiles, como GPT-4, LLaMA 3, e incluso modelos diseñados específicamente para el manejo de datos numéricos y estructurados.

Dado que se trata de un área de investigación altamente dinámica y en plena evolución, no puede descartarse que estudios futuros, que empleen modelos más modernos o incorporen ajustes diseñados específicamente para las carac-

terísticas propias de las series temporales, encuentren mejoras significativas en el desempeño. Por lo tanto, las conclusiones extraídas hasta el momento deben ser interpretadas con cautela, reconociendo que podrían cambiar a medida que se desarrollen nuevas variantes arquitectónicas y enfoques más adecuados al dominio temporal.

En la sección que sigue, nos adentramos en las técnicas utilizadas por los estudios sobre esta temática y posteriormente utilizaremos algunas de ellas para realizar experimentos sobre los datos trabajados en el capítulo anterior.

4.4. Trabajos relacionados

En esta sección presentamos y analizamos los enfoques más relevantes que han utilizado modelos de lenguaje a gran escala para la predicción de series temporales, con especial énfasis en aplicaciones sobre datos financieros y económicos. Este foco resulta particularmente pertinente dado lo discutido en secciones anteriores, donde se ha demostrado que las series económicas poseen una estructura altamente compleja: son no estacionarias, presentan alta volatilidad y patrones que se repiten en múltiples momentos del tiempo.

Podemos dividir las técnicas basadas en modelos de lenguaje en dos grandes categorías Z. Liu y Jia, [2025](#):

- LLM entrenados con retroalimentación del mercado y datos externos. Estos modelos combinan información textual (noticias, relaciones entre acciones, tendencias estadísticas, etc.) con datos históricos, tratando los datos como secuencias de tokens. Sin embargo, estos enfoques dependen de la disponibilidad de datos textuales de alta calidad, lo cual no siempre es viable.
- Adaptación directa de LLM generales a tareas de series temporales. Estos métodos incluyen el uso de instrucciones diseñados para alinear las series con representaciones textuales (*prompting*), o el preentrenamiento con grandes volúmenes de datos temporales para fomentar la transferencia de conocimiento.

Dentro de esta segunda categoría, una línea de trabajo reciente en un marco general de adaptación de estos modelos a series de tiempo, propone usar únicamente instrucciones en lenguaje natural para interactuar con estos modelos, sin necesidad de modificar su estructura ni entrenarlos nuevamente (Xue

y Salim, 2023, Gruver et al. 2023). La propuesta consiste en convertir los datos de series temporales en formato textual y presentarlos al modelo mediante instrucciones específicas. Este enfoque es más simple que otros que requieren cambios en la arquitectura o fine-tuning complejos, lo que lo hace más fácil de aplicar en la práctica. Esto se logra gracias a la capacidad de estos modelos para procesar y comprender tanto información textual como numérica, lo que posibilita reinterpretar tareas clásicas de series temporales como problemas de procesamiento de lenguaje. Un ejemplo típico de esta aproximación es el de Xue y Salim (2023):

Desde t_1 hasta t_{obs} , la temperatura promedio de la región U_m fue de x_{mt} grados cada día. ¿Cuál será la temperatura en t_{obs} ?

Sin embargo, persisten ciertas dudas sobre si estos modelos realmente pueden razonar sobre series temporales con aproximaciones zero-shot en escenarios más complejos. Transformar datos de series temporales en texto también conlleva limitaciones, ya que estos modelos no siempre capturan efectivamente las dependencias temporales características de las series de tiempo (Merrill et al. 2024). Convertir las series en secuencias de tokens textuales puede generar que se pierda la estructura temporal original y, con ello, la capacidad del modelo para reconocer patrones secuenciales de forma efectiva. Por ejemplo, Belyaeva et al. (2023) señalan este desafío en aplicaciones del ámbito de la salud, caracterizadas por datos altamente dimensionales y multimodales, y proponen un marco (HeLM) para mapear dichos datos no textuales al espacio de representaciones tokenizadas. Esta representación permite proporcionar contexto a un LLM para realizar tareas como la predicción de riesgo de enfermedades, sin depender exclusivamente de una conversión textual de los datos y demuestra mejores resultados. Otros enfoques como Chronos, que es un modelo de series temporales basado en transformers y preentrenado como un modelo de lenguaje, tokeniza los valores de la serie mediante normalización y cuantización (Ansari et al. 2024).

Posteriormente, se ha observado que en lugar de formular directamente preguntas complejas con la totalidad de los datos, una estrategia eficaz consiste en descomponer las tareas en pasos más simples y secuenciales, facilitando así el razonamiento del modelo (Wei et al. 2022). En esa dirección, otras investigaciones han incorporado cadenas de pensamiento, que consisten en una serie de pasos intermedios de razonamiento que mejoran significativamente la

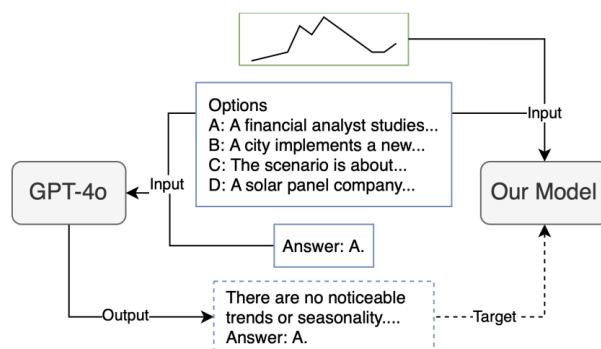


Figura 4.1: Ejemplo de uso de razonamiento con cadenas de pensamiento (CoT) en el modelo GPT-4o para interpretar datos de series temporales y seleccionar la opción más adecuada a partir de un conjunto de alternativas. Imagen tomada de (Chow et al. 2024).

capacidad de los modelos para resolver tareas complejas. Por ejemplo, en el trabajo de Chow et al. (2024) se aplica una estrategia de ampliación con CoT, donde al modelo se le presentan varias opciones de respuesta junto con la opción correcta, y se le pide que genere una explicación que justifique por qué esa respuesta es la correcta (ver figura 4.1). Otros estudios, exploran cómo diseñar instrucciones con CoT que activen explícitamente capacidades de razonamiento a corto y largo plazo en LLM, con el fin de mejorar su desempeño como predictores zero-shot en tareas de series temporales (H. Liu et al. 2024).

Otros enfoques en esta dirección se concentraron en transformar las series temporales en secuencias significativas mediante la segmentación en parches (estrategia conocida como "patching") (Nie et al. 2022), con el objetivo de capturar patrones temporales relevantes y reducir el costo computacional. Estas ideas sirvieron como base para desarrollos posteriores como TST (Bumb et al. 2025), donde los parches son incrustados en vectores latentes que funcionan como representaciones textuales de los datos originales.

Más recientemente, enfoques como PatchInstruct (Bumb et al. 2025) extienden esta idea al combinar la estrategia de parches con instrucciones estructuradas en lenguaje natural, lo que permite guiar al modelo de manera explícita hacia predicciones más precisas. En este trabajo, además incorporan series temporales similares seleccionadas a partir de las series más parecidas en el conjunto de datos histórico, con el objetivo de proporcionar señales contextuales adicionales y patrones recurrentes que podrían no ser completamente observables en la serie objetivo de manera aislada. La incorporación de series

vecinas puede mejorar la capacidad del modelo para inferir tendencias más precisas, al aportar información contextual que va más allá de la serie objetivo. Sin embargo, esto no siempre es beneficioso: concluyen que cuando las series vecinas presentan baja correlación, pueden introducir ruido o confusión en lugar de contribuir con contexto útil, lo cual vimos en secciones anteriores bajo el concepto de transferencia negativa.

Aunque todavía son pocos los trabajos que aplican LLM a series temporales financieras y económicas, existe un enfoque destacado que combina múltiples estrategias y fuentes de datos externas para predecir precios de acciones del NASDAQ-100 (X. Yu et al. 2023). Este enfoque integra precios históricos, perfiles de empresas y noticias financieras, las cuales son resumidas mediante instrucciones específicas diseñadas para GPT-4. En los experimentos, se incluyen ejemplos dentro de las instrucciones que combinan secuencias de precios, perfiles empresariales y resúmenes de noticias, lo que permite al modelo realizar razonamiento entre distintas secuencias (cross-sequence reasoning). Además, se observa que la incorporación de cadenas de pensamiento mejora significativamente el rendimiento predictivo del modelo.

Un aspecto especialmente crítico es que los enfoques actuales basados en modelos de lenguaje de gran escala tienden a aplicar divisiones temporales uniformes, ignorando las propiedades multiescala y no estacionarias que caracterizan a las series temporales financieras. A diferencia de otras series, como las de consumo eléctrico, los datos financieros presentan patrones complejos que emergen en distintas escalas temporales y cuya estructura frecuencial varía a lo largo del tiempo. Estas particularidades estructurales requieren representaciones que conserven la integridad informacional de los datos.

Con el fin de abordar estas limitaciones, en el trabajo *LLM4FTS: Enhancing Large Language Models for Financial Time Series Prediction* (Z. Liu y Jia, 2025) los autores proponen una arquitectura que integra un módulo de reconocimiento de patrones invariante a la escala, junto con una estrategia dinámica y adaptable de segmentación en parches, la cual se ve potenciada mediante operadores convolucionales. Esta combinación permite modelar dependencias temporales en múltiples resoluciones, facilitando así una extracción jerárquica de características y un análisis sistemático de patrones complejos presentes en series financieras.

Para la segmentación, se emplea el algoritmo de K vecinos más cercanos en conjunto con la métrica de distancia DTW, lo que permite agrupar seg-

mentos de distintas longitudes en función de su forma, una vez normalizados en duración y magnitud. Esta estrategia posibilita identificar similitudes entre segmentos incluso si difieren en escala, y seleccionar de forma local la longitud de segmentación óptima, en lugar de imponer una única longitud fija para toda la serie. Esto permite preservar las correlaciones internas propias de cada escala, mejorar la detección de patrones y reforzar la capacidad predictiva del modelo. Asimismo, esta metodología podría servir como estrategia para la selección de conjuntos objetivo en enfoques de transferencia de conocimiento aplicados a series temporales mencionadas anteriormente en el trabajo.

En la sección siguiente realizaremos algunas pruebas utilizando modelos de lenguaje de gran escala con las series de tiempo de IPC utilizadas en el capítulo anterior. En particular, exploraremos cómo modelos como BERT (Devlin et al. 2019), GPT-2 (Radford et al. 2019) y T5 (Raffel et al. 2020) pueden emplearse para realizar predicciones, generación de texto y análisis semántico sobre descripciones asociadas a las series temporales.

4.5. Aplicación: LLM para el Índice de Precios al Consumo en Uruguay

El objetivo de esta sección es realizar una serie de experimentos exploratorios que permitan ilustrar la aplicación de técnicas recientes basadas en grandes modelos de lenguaje al problema del pronóstico del IPC trabajado en el capítulo 3. Estas técnicas, discutidas en la sección anterior, se evalúan aquí en tareas complementarias al modelado estadístico tradicional, haciendo foco en su capacidad de generar pronósticos en modo sin requerir reentrenamiento sobre datos específicos del dominio.

Es importante señalar, que hasta ahora gran parte de este trabajo se ha centrado en modelos clásicos de series temporales, combinando simulaciones y análisis de datos reales para identificar en qué escenarios el transfer learning resulta beneficioso, en función de las características de las series source y target y la relación entre ellas. A continuación, se propone un cambio de enfoque hacia modelos de lenguaje a gran escala, no con el objetivo principal de determinar cuándo transferir conocimiento, sino de explorar técnicas específicas de transferencia y adaptación que estos modelos permiten (que responden a la pregunta de cómo transferir). En esta etapa, la aproximación consistirá únicamente en

la variación sistemática de *prompts* como mecanismo de adaptación.

Este enfoque es relevante porque estos modelos ofrecen capacidades de aprendizaje contextual y representación de secuencias mucho más flexibles que los modelos clásicos, permitiendo analizar cómo se puede adaptar un modelo preentrenado a diferentes series temporales sin necesidad de diseñar manualmente cada escenario de transferencia. De esta manera, se pretende complementar la perspectiva anterior con un enfoque centrado en las estrategias de transferencia más que en los factores que la hacen efectiva.

Como caso de estudio, se utilizó la serie del IPC correspondiente al interior del país para los años 2023 y 2024, con el objetivo de generar predicciones para los primeros seis meses de 2025. Para esto, se consideraron tres variantes de instrucciones que buscan incorporar progresivamente más información contextual o razonamiento:

- PromptCast (Xue y Salim, 2023): donde los valores numéricos de entrada se convierten directamente en una secuencia textual. El modelo genera predicciones extrapolando esa secuencia sin contexto adicional.
- Incorporación de Montevideo: agrega también la serie del IPC de Montevideo como secuencia auxiliar, de modo que el modelo pueda inferir patrones compartidos y usarlos como referencia para mejorar la predicción del Interior.
- LSTPrompt (H. Liu et al. 2024): aquí, se proveen instrucciones explícitas al modelo para que divida la tarea de pronóstico en componentes de corto y largo plazo, diseñe mecanismos de predicción distintos para cada uno, y avance paso a paso en la generación de cada valor. Este enfoque busca inducir razonamiento estructurado en la predicción.

En los tres escenarios anteriores se utilizaron únicamente datos de 2023 y 2024 como insumo. Para el primer caso, utilizamos la interfaz de OpenAI de ChatGPT ¹ utilizando el modelo GPT-4o (Hurst et al. 2024) y Chronos. Para los casos 2 y 3, como Chronos no permite el agregado de una instrucción, utilizamos solamente la interfaz mencionada.

Chronos (Ansari et al. 2024) es un modelo de series temporales preentrenado bajo un marco de modelado del lenguaje, capaz de realizar pronósticos *zero-shot*, es decir, sin necesidad de reentrenamiento con datos específicos. Este

¹<https://chatgpt.com/>

enfoque implica que el modelo no requiere ser reentrenado o ajustado específicamente sobre los datos de la serie de interés para generar predicciones. En otras palabras, gracias a su entrenamiento previo en un amplio conjunto de series heterogéneas, Chronos puede generalizar y aplicar directamente el conocimiento adquirido a nuevas series, incluso cuando no ha sido expuesto a ellas con anterioridad. Para nuestros experimentos, utilizamos la implementación disponible en HuggingFace² para generar predicciones de la serie del Interior con un horizonte de 6 meses (6 observaciones).

En este tipo de modelos, el número de parámetros es un factor relevante, ya que generalmente un mayor número de parámetros permite al modelo capturar patrones más complejos y aprender representaciones más ricas de los datos, aunque también implica un mayor costo computacional. Evaluamos variantes de Chronos con diferente tamaño de modelo (small, base y large con 46, 200 y 710 millones de parámetros, respectivamente) y observamos que sus desempeños fueron comparables en nuestro caso de estudio. Por esta razón, seleccionamos la versión large para el análisis de las conclusiones, dado que ofrece la mayor capacidad de representación sin comprometer la eficiencia de los resultados.

En el apéndice 5 se presentan las diferentes instrucciones utilizadas en este experimento. Los resultados los presentamos en la tabla 4.1.

Mes	Real	Chronos	PromptCast	+ MVD	LSTPrompt
2025-01	110.91	109.68	109.65	109.70	109.60
2025-02	111.49	109.68	110.01	110.12	109.94
2025-03	111.92	110.46	110.34	110.47	110.45
2025-04	112.51	110.46	110.63	110.83	110.80
2025-05	112.51	110.46	110.91	111.21	111.15
2025-06	112.29	110.46	110.98	112.17	111.78
MSE	–	3.13	2.35	1.66	1.89

Tabla 4.1: Predicciones de IPC Interior (enero-junio 2025) y errores cuadráticos medios (MSE) para cada enfoque: Chronos, PromptCast, PromptCast con contexto de Montevideo (+ MVD), y LSTPrompt con razonamiento paso a paso.

El enfoque que alcanzó el mejor desempeño fue PromptCast con contexto de Montevideo, con un MSE de 1.66. Este resultado sugiere que, incluso sin un entrenamiento específico, un LLM puede beneficiarse de señales contextuales

²<https://huggingface.co/amazon/chronos-t5-small>

externas siempre que estas se presenten de manera adecuada en el prompt. Cabe señalar que las predicciones del IPC para 2025 se basaron en la información de los años 2023 y 2024, periodos en los cuales las series mostraban una alta similitud estructural y una baja distancia bajo métricas como DTW, EDR y TWED, tal como se analizó en el capítulo 3.

Por su parte, la incorporación de razonamiento paso a paso mediante LST-Prompt también permitió una mejora respecto a PromptCast y Chronos sin contexto, alcanzando un MSE de 1.89. Sin embargo, este resultado no llegó a superar el desempeño obtenido en el escenario con señales contextuales de Montevideo.

Por otro lado, Chronos, a pesar de estar especializado en series temporales, mostró un comportamiento más conservador, con una marcada tendencia a repetir valores, lo que contribuyó a un mayor error de predicción. Asimismo, funciona como una caja negra: no ofrece posibilidades de intervención en el diseño de las instrucciones ni en la arquitectura de razonamiento, lo que limita su interpretabilidad. Su carácter estocástico genera resultados distintos en ejecuciones sucesivas, lo que dificulta identificar con precisión los factores que explican los casos en los que el pronóstico es más acertado. Finalmente, cabe destacar que Chronos está entrenado para predecir valores numéricos de manera directa.

En resumen, los experimentos realizados evidencian el potencial de los LLM como herramientas de pronóstico flexibles, particularmente cuando se aplican técnicas de prompting apropiadas y se incorpora razonamiento estructurado. Una línea de trabajo futura consiste en optimizar las instrucciones y explorar otros modelos disponibles. Es importante señalar que, gracias a su elevada capacidad de representación y generalización, estos modelos no requieren grandes volúmenes de datos de transferencia para adaptarse a nuevas tareas. Asimismo, su habilidad para generar resultados en lenguaje natural permite acceder a información y patrones que serían menos evidentes mediante modelos clásicos.

No obstante, es fundamental tener en cuenta los riesgos asociados al uso de LLM en tareas de pronóstico de series temporales. En primer lugar, estos modelos funcionan como cajas negras, lo que significa que es difícil comprender cómo se generan las predicciones o qué factores específicos influyen en los resultados. Esta falta de interpretabilidad limita la capacidad de validar y explicar los pronósticos, lo cual es una desventaja frente a modelos clásicos. En segundo lugar, su desempeño depende de manera crítica de la formulación de

los prompts, en donde pequeñas variaciones en la redacción pueden producir resultados significativamente distintos. Esto exige un proceso cuidadoso de diseño y ajuste de las instrucciones para asegurar consistencia y relevancia en las predicciones. En tercer lugar, aunque los LLM son capaces de generalizar, no siempre capturan de manera precisa la dinámica temporal de las series, dado que su arquitectura y entrenamiento están orientados principalmente al procesamiento de texto. Esto puede dar lugar a la imposición de patrones uniformes que no reflejan la realidad subyacente de la serie, especialmente en contextos altamente volátiles. Finalmente, persiste el riesgo de las denominadas *alucinaciones*, en las cuales el modelo puede producir información incorrecta, inexacta o incluso ficticia, aunque la presente de forma coherente y convincente, lo que disminuye su confiabilidad y utilidad (Kalai et al. [2025](#)). La combinación de estos factores resalta la necesidad de un uso crítico y supervisado de los LLM, complementando sus predicciones con métodos tradicionales y validación empírica.

Desde la perspectiva de este trabajo, enfocado en la transferencia de conocimiento en series temporales, los LLM ofrecen un marco metodológico potente para explorar nuevas formas de transferencia de conocimiento, planteando la posibilidad de definir conjuntos de tareas fuente y objetivo más afines desde un punto de vista estructural, facilitando una transferencia más efectiva y fundamentada. Es un área que debe seguir explorándose y que tiene un potencial muy valioso para contribuir con la predicción de series de tiempo, aunque se debe tener en consideración qué riesgos subyacentes se presentan en estas aplicaciones a series de tiempo.

Capítulo 5

Conclusiones y trabajo futuro

En este trabajo se exploró el aprendizaje por transferencia en contextos con dependencia temporal, combinando el análisis de modelos clásicos de series de tiempo con la aplicación de enfoques más recientes, como los grandes modelos de lenguaje (LLM). A partir de una revisión conceptual y experimental que abarcó simulaciones controladas y el estudio de un caso real con datos del IPC de Uruguay, se pretendió ofrecer una visión crítica sobre cuándo la transferencia de conocimiento puede resultar efectiva. Dentro de las conclusiones sobre lo anterior, tenemos que la transferencia de conocimiento desde un dominio fuente no garantiza mejoras sistemáticas en el dominio objetivo.

Esto último resultó posible gracias al uso de modelos clásicos de series de tiempo, cuya estructura paramétrica permitió una interpretación directa del proceso de transferencia y sus efectos sobre la dinámica del modelo objetivo. Este enfoque pretendió aportar una perspectiva novedosa dentro de la literatura existente, donde los trabajos sobre transferencia en series temporales, hasta lo que conocemos, se apoyan principalmente en modelos de caja negra, sin ofrecer un análisis interpretativo del mecanismo subyacente.

Los resultados de las simulaciones, en línea con hallazgos de otras investigaciones, muestran que a medida que se dispone de un mayor número de observaciones en el conjunto objetivo, el beneficio de la transferencia tiende a disminuir, aunque esta relación positiva no resultó tan fuerte. Dado que en este caso trabajamos con series simuladas y conocíamos el proceso generador de los datos, fue posible identificar que cuando la dependencia temporal del target con sus propios rezagos es mayor que la del source, la transferencia no resulta ventajosa. Asimismo, al emplear medidas de distancia entre series (EDR,

TWED y DTW), se observó que los casos de transferencia negativa coinciden con aquellos en los que la distancia entre la serie fuente y la serie objetivo es más elevada. Para esto último, sería interesante continuar investigando si es posible determinar un umbral utilizando estas distancias, a partir del cual el transfer deja de resultar beneficioso.

En línea con lo anterior, resultaría también relevante realizar una descomposición de la expresión de ganancia, así como se realiza en el caso de modelos con independencia de datos (Obst et al. 2022). Esto nos permitiría entender de manera analítica los mecanismos que determinan el éxito o fracaso de la transferencia en contextos con dependencia temporal, análogo a lo que se ha hecho en contextos independientes, pero considerando explícitamente la estructura de correlación propia de las series de tiempo. Consecuentemente, abriría la posibilidad de derivar condiciones teóricas de transferibilidad y de cuantificar el impacto del factor temporal en la ganancia obtenida.

En el análisis con los datos del IPC, a diferencia de las simulaciones, no contamos con certeza sobre el proceso generador, aunque se realizó un estudio previo de las series para identificar los modelos más adecuados. En este caso no se observó que la utilidad de la transferencia disminuyera a medida que aumentaba el número de observaciones del conjunto objetivo. Sin embargo, sí se constató que cuando la serie objetivo presenta una mayor dependencia temporal que la serie fuente, la transferencia puede no resultar beneficiosa. Además, se advirtió que los episodios de transferencia negativa coincidían con períodos en los que las series de Montevideo e Interior mostraban mayores distancias, lo que subraya la relevancia de la similitud entre series para la efectividad del proceso.

Particularmente, los casos en los que la transferencia no aportó mejoras coincidieron con intervalos atravesados por shocks externos que afectaron de forma diferenciada a ambas regiones. Entre 2012 y 2017, por ejemplo, se registraron sequías y caídas en los precios internacionales que impactaron con mayor fuerza en el sector agropecuario, concentrado en el Interior, generando dinámicas divergentes respecto a la capital. Estos resultados ilustran que la transferencia pierde efectividad cuando las estructuras económicas se desacoplan por choques externos, aun en escenarios en los que el tamaño del conjunto objetivo podría sugerir lo contrario.

Tras el análisis con modelos clásicos, se realizó una revisión de la literatura y se exploraron técnicas basadas en grandes modelos de lenguaje. En este ca-

so, el objetivo no era determinar bajo qué condiciones la transferencia resulta beneficiosa, sino evaluar, mediante pruebas con diferentes estrategias de instrucción, si la inclusión de información de la serie fuente agregaba valor en la predicción de la serie objetivo. Para ello se utilizaron nuevamente los datos del IPC de Uruguay. Los resultados disponibles permitieron observar que agregar contexto de la serie de Montevideo al predecir la serie del Interior generaba mejoras en la predicción, sugiriendo que, incluso con enfoques simples, los LLM pueden aprovechar información adicional de series relacionadas.

Los LLM ofrecen ciertas ventajas respecto a otro tipo de modelos. Su capacidad para trabajar con descripciones en lenguaje natural permite integrar información contextual o cualitativa que los modelos clásicos o redes neuronales no capturan. Además, pueden adaptarse a distintos dominios mediante técnicas de *fine-tuning* o *prompting*, facilitando la generalización del conocimiento aprendido. Asimismo, su arquitectura flexible permite combinar datos estructurados y no estructurados, lo que abre nuevas posibilidades para el modelado temporal y la interpretación de fenómenos complejos.

No obstante, este enfoque presenta limitaciones importantes, como los desafíos computacionales, la menor interpretabilidad de los modelos y la escasez de estudios previos sobre la aplicación de LLM en transferencia de conocimiento para series temporales. Esto indica que el campo sigue en desarrollo y ofrece oportunidades significativas para futuras investigaciones.

En línea con la posibilidad de continuar esta investigación, sería valioso explorar las técnicas y hallazgos de la literatura en otros contextos de dependencia, más allá de la temporal, como por ejemplo la dependencia espacial. En cuanto a los modelos clásicos, se podrían experimentar combinaciones más diversas de enfoques, prestando especial atención al fine-tuning de los parámetros para optimizar su desempeño. Respecto al análisis del IPC, sería interesante incorporar otras series regionales además de Montevideo e Interior, lo que permitiría estudiar la transferencia en un contexto más amplio y heterogéneo. Por último, en el ámbito de los LLM, se podrían evaluar otros modelos y técnicas que permitan un análisis más profundo de los métodos aplicables a series temporales, incluyendo estrategias avanzadas de instrucción, agregación de contexto y posibles ajustes de entrenamiento que maximicen la capacidad de los modelos para aprovechar información proveniente de series relacionadas.

La articulación de enfoques clásicos con LLM se perfila como una vía prometedora, siempre que se mitiguen riesgos y se preserve la validación empírica

y la interpretabilidad. Este trabajo busca aportar a metodologías de transferencia más sólidas y efectivas, ofreciendo un marco práctico que facilite su aplicación en contextos reales.

Referencias bibliográficas

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Almeida, F., y Xexéo, G. (2019). Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*.
- Alyafeai, Z., AlShaibani, M. S., y Ahmad, I. (2020). A survey on transfer learning in natural language processing. *arXiv preprint arXiv:2007.04239*.
- Alzayat, A. (2011). *Positive and Negative Analogical Transfer in Problem Solving* [Tesis de maestría, University of Waterloo].
- Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., et al. (2024). Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*.
- Arnold, Andrew, Nallapati, Ramesh, Cohen y W, W. (2007). A comparative study of methods for transductive transfer learning. *Seventh IEEE international conference on data mining workshops (ICDMW 2007)*, 77-82.
- Belyaeva, A., Cosentino, J., Hormozdiari, F., Eswaran, K., Shetty, S., Corrado, G., Carroll, A., McLean, C. Y., y Furlotte, N. A. (2023). Multimodal llms for health grounded in individual-specific data. *Workshop on Machine Learning for Multimodal Healthcare Data*, 86-102.
- Bojanowski, P., Grave, E., Joulin, A., y Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5, 135-146.
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

- Box, G. (2013). Box and Jenkins: time series analysis, forecasting and control. En *A Very British Affair: Six Britons and the Development of Time Series Analysis During the 20th Century* (pp. 161-215). Springer.
- Bozinovski, S. (2020). Reminder of the first paper on transfer learning in neural networks, 1976. *Informatica*, 44(3).
- Bumb, M., Vemulapalli, A., Jella, S. H. V. P., Gupta, A., La, A., Rossi, R. A., Chen, H., Dernoncourt, F., Ahmed, N. K., y Wang, Y. (2025). Forecasting Time Series with LLMs via Patch-Based Prompting and Decomposition. *arXiv preprint arXiv:2506.12953*.
- Chen, L., Keogh, E. J., Tang, J., Sidiropoulos, N., y Príncipe, J. C. (2005). Edit distance for real sequences. *Proceedings of the 2005 SIAM International Conference on Data Mining*, 29-40.
- Chen, T., y Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785-794.
- Chen, Z., y Daehler, M. W. (1989). Positive and negative transfer in analogical problem solving by 6-year-old children. *Cognitive Development*, 4(4), 327-344.
- Chow, W., Gardiner, L., Hallgrímsson, H. T., Xu, M. A., y Ren, S. Y. (2024). Towards time series reasoning with llms. *arXiv preprint arXiv:2409.11376*.
- Christ, M., Braun, N., Neuffer, J., y Kempa-Liehr, A. W. (2018). Time series feature extraction on basis of scalable hypothesis tests (tsfresh—a python package). *Neurocomputing*, 307, 72-77.
- Church, K. W. (2017). Word2Vec. *Natural Language Engineering*, 23(1), 155-162.
- Cortes, C., y Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.
- Dai, W., Yang, Q., Xue, G.-R., y Yu, Y. (2007). Boosting for transfer learning. *Proceedings of the 24th international conference on Machine learning*, 193-200.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., y Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248-255.
- Devlin, J., Chang, M.-W., Lee, K., y Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

- Dickey, D. A., y Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a), 427-431.
- Ehrig, C., Sonnleitner, B., Neumann, U., Cleophas, C., y Forestier, G. (2024). The impact of data set similarity and diversity on transfer learning success in time series forecasting. *arXiv preprint arXiv:2404.06198*.
- Elliott, G., Rothenberg, T. J., y Stock, J. H. (1992). Efficient tests for an autoregressive unit root.
- Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., y Muller, P.-A. (2018). Transfer learning for time series classification. *2018 IEEE international conference on big data (Big Data)*, 1367-1376.
- Freitag, M., y Al-Onaizan, Y. (2017). Beam search strategies for neural machine translation. *arXiv preprint arXiv:1702.01806*.
- Freund, Y., Schapire, R., y Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780), 1612.
- Friedman, M. (1963). Inflation: Causes and consequences. (*No Title*).
- Gamerman, A., Vovk, V., y Vapnik, V. (2013). Learning by transduction. *arXiv preprint arXiv:1301.7375*.
- Germán-Morales, M., Rivera-Rivas, A., Díaz, M., y Carmona, C. (2024). Transfer Learning with Foundational Models for Time Series Forecasting using Low-Rank Adaptations. *arXiv preprint arXiv:2410.11539*.
- Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., y Dubrawski, A. (2024). Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*.
- Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., y Smola, A. (2012). A kernel two-sample test. *The journal of machine learning research*, 13(1), 723-773.
- Gretton, A., Bousquet, O., Smola, A., y Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. *International conference on algorithmic learning theory*, 63-77.
- Gruver, N., Finzi, M., Qiu, S., y Wilson, A. G. (2023). Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36, 19622-19635.
- Hardy, G. H., Littlewood, J. E., y Pólya, G. (1952). *Inequalities*. Cambridge university press.

- He, K., Zhang, X., Ren, S., y Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778.
- Holtzman, A., Buys, J., Du, L., Forbes, M., y Choi, Y. (2019). The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.
- Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. (2024). Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., et al. (2023). Time-llm: Time series forecasting by re-programming large language models. *arXiv preprint arXiv:2310.01728*.
- Kalai, A. T., Nachum, O., Vempala, S. S., y Zhang, E. (2025). Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.
- Kantorovich, L. V. (2006). On the Translocation of Masses. *Journal of mathematical sciences*, 133(4).
- Karb, T., Kühl, N., Hirt, R., y Glivici-Cotruta, V. (2020). A network-based transfer learning approach to improve sales forecasting of new products. *arXiv preprint arXiv:2005.06978*.
- Kullback, S., y Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1), 79-86.
- Kurzweil, R., Richter, R., Kurzweil, R., y Schneider, M. L. (1990). *The age of intelligent machines* (Vol. 579). MIT press Cambridge.
- Kwiatkowski, D., Phillips, P. C., Schmidt, P., y Shin, Y. (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*, 54(1-3), 159-178.
- Laspeyres, K. (1871). IX. Die Berechnung einer mittleren Waarenpreissteigerung. *Jahrbücher für Nationalökonomie und Statistik*, 16 (1), 296–318.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8), 707-710.
- Lewis, M. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Lewis, W. A., et al. (1954). Economic development with unlimited supplies of labour.

- Lim, B., y Zohren, S. (2021). Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194), 20200209.
- Liu, H., Zhao, Z., Wang, J., Kamarthi, H., y Prakash, B. A. (2024). Lst-prompt: Large language models as zero-shot time series forecasters by long-short-term prompting. *arXiv preprint arXiv:2402.16132*.
- Liu, P., Guo, H., Dai, T., Li, N., Bao, J., Ren, X., Jiang, Y., y Xia, S.-T. (2025). Calf: Aligning llms for time series forecasting via cross-modal fine-tuning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18), 18915-18923.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., y Stoyanov, V. (1907). RoBERTa: A robustly optimized BERT pretraining approach. arXiv [Preprint](2019). *arXiv preprint arXiv:1907.11692*.
- Liu, Z., y Jia, R. (2025). LLM4FTS: Enhancing Large Language Models for Financial Time Series Prediction. *arXiv preprint arXiv:2505.02880*.
- Long, H.-W., Ho, O.-I., He, Q.-Q., y Si, Y.-W. (2025). Transfer Learning in Financial Time Series with Gramian Angular Field. *arXiv preprint arXiv:2504.00378*.
- Lumsdaine, R. L., y Papell, D. H. (1997). Multiple trend breaks and the unit-root hypothesis. *Review of economics and Statistics*, 79(2), 212-218.
- Majtey, A. P., Lamberti, P. W., y Prato, D. P. (2005). Jensen-Shannon divergence as a measure of distinguishability between mixed quantum states. *Physical Review A—Atomic, Molecular, and Optical Physics*, 72(5), 052310.
- Malhotra, P., Tv, V., Vig, L., Agarwal, P., y Shroff, G. (2017). TimeNet: Pre-trained deep recurrent neural network for time series classification. *arXiv preprint arXiv:1706.08838*.
- Marteau, P.-F. (2009). Time Warp Edit Distance with Stiffness Adjustment for Time Series Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2), 306-318. <https://doi.org/10.1109/TPAMI.2008.76>
- Merrill, M. A., Tan, M., Gupta, V., Hartvigsen, T., y Althoff, T. (2024). Language models still struggle to zero-shot reason about time series. *arXiv preprint arXiv:2404.11757*.

- Michie, D., Spiegelhalter, D. J., Taylor, C. C., y Campbell, J. (1995). *Machine learning, neural and statistical classification*. Ellis Horwood.
- Mihalkova, L., Huynh, T., y Mooney, R. J. (2007). Mapping and revising markov logic networks for transfer learning. *Aaai*, 7, 608-614.
- Mihalkova, L., y Mooney, R. J. (2008). Transfer learning by mapping with minimal target data. *Proceedings of the AAAI-08 workshop on transfer learning for complex tasks*, 31-36.
- Miller, J. A., Aldosari, M., Saeed, F., Barna, N. H., Rana, S., Arpinar, I. B., y Liu, N. (2024). A survey of deep learning and foundation models for time series forecasting. *arXiv preprint arXiv:2401.13912*.
- Mitchell, T. M. (1980). The need for biases in learning generalizations.
- Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2), 17-23.
- Mou, L., Meng, Z., Yan, R., Li, G., Xu, Y., Zhang, L., y Jin, Z. (2016). How transferable are neural networks in NLP applications? *Proceedings of the 2016 conference on empirical methods in natural language processing*, 479-489.
- Müller, M. (2007). *Information retrieval for music and motion* (Vol. 2). Springer.
- Nie, Y., Nguyen, N. H., Sinthong, P., y Kalagnanam, J. (2022). A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*.
- Obst, D., Ghattas, B., Claudel, S., Cugliari, J., Goude, Y., y Oppenheim, G. (2022). Improved linear regression prediction by transfer learning. *Computational Statistics & Data Analysis*, 174, 107499.
- Oh, K.-S., y Jung, K. (2004). GPU implementation of neural networks. *Pattern Recognition*, 37(6), 1311-1314.
- Pan, S. J. (2020). Transfer learning. *Learning*, 21, 1-2.
- Pan, S. J., y Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10), 1345-1359.
- Pennington, J., Socher, R., y Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532-1543.
- Peña, D. (2005). *Análisis de series temporales*. Alianza.
- Peterson, K. T., y Van Duijn, B. (2024). Leveraging artificial intelligence for enhanced seed testing: Current trends and future prospects.

- Phillips, P. C., y Perron, P. (1988). Testing for a unit root in time series regression. *biometrika*, 75(2), 335-346.
- Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., y Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 9. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., y Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140), 1-67.
- Rodriguez-Collazo. (2024). Apuntes del curso de Series Cronológicas [Apuntes del curso de “Series Cronológicas”, Facultad de Ciencias Económicas y de Administración, Udelar].
- Rosenstein, M. T., Marx, Z., y Kaelbling, L. P. (2005). Transfer Learning with Partial Observability. *Proceedings of the International Workshop on Inductive Transfer: 10 Years Later*. https://socrates.acadiau.ca/courses/comp/dsilver/ITWS05_Submitted_Papers/ITWS10-RosensteinM05_ITWS.pdf
- Said, S. E., y Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3), 599-607.
- Seabold, S., y Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.
- Serra, J., y Arcos, J. L. (2014). An empirical evaluation of similarity measures for time series classification. *Knowledge-Based Systems*, 67, 305-314.
- Shavlik, J. W., y Towell, G. G. (1989). Combining explanation-based learning and artificial neural networks. *Proceedings of the sixth international workshop on Machine learning*, 90-92.
- Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404, 132306.
- Sutskever, I., Vinyals, O., y Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

- Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., y Liu, C. (2018). A survey on deep transfer learning. *Artificial Neural Networks and Machine Learning-ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, 270-279.
- Tan, M., Merrill, M. A., Gupta, V., Althoff, T., y Hartvigsen, T. (2024). Are language models actually useful for time series forecasting? *URL* <https://arxiv.org/abs/2406.16964>.
- Tavenard, R., Faouzi, J., Vandewiele, G., Divo, F., Androz, G., Holtz, C., Payne, M., Yurchak, R., Rußwurm, M., Kolar, K., y Woods, E. (2020). Tslearn, A Machine Learning Toolkit for Time Series Data. *Journal of Machine Learning Research*, 21(118), 1-6. <http://jmlr.org/papers/v21/20-091.html>
- Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Roziere, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*. <https://arxiv.org/abs/2302.13971>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., y Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., y Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, 26, 275-309.
- Wang, Z., Dai, Z., Póczos, B., y Carbonell, J. (2019). Characterizing and avoiding negative transfer. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11293-11302.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824-24837.
- Weiss, K., Khoshgoftaar, T. M., y Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, 3, 1-40.

- Wiher, G., Meister, C., y Cotterell, R. (2022). On decoding strategies for neural text generators. *Transactions of the Association for Computational Linguistics*, 10, 997-1012.
- Xue, H., y Salim, F. D. (2023). Promptcast: A new prompt-based learning paradigm for time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 36(11), 6851-6864.
- Yamaguchi, M. (2022). editdistance: Fast implementation of edit distance (Levenshtein distance) [Accessed: 2025-05-03].
- Yang, Q., y Wu, X. (2006). 10 Challenging Problems in Data Mining Research.
- Ye, R., y Dai, Q. (2021). Implementing transfer learning across different datasets for time series forecasting. *Pattern Recognition*, 109, 107617.
- Yu, F., Cui, L., Chen, H., Cao, Y., Liu, N., Huang, W., Xu, Y., y Lu, H. (2022). Healthnet: A health progression network via heterogeneous medical information fusion. *IEEE Transactions on Neural Networks and Learning Systems*, 34(10), 6940-6954.
- Yu, X., Chen, Z., Ling, Y., Dong, S., Liu, Z., y Lu, Y. (2023). Temporal data meets LLM—explainable financial time series forecasting. *arXiv preprint arXiv:2306.11025*.
- Zhang, Wen, Deng, L., Zhang, Lei, Wu y Dongrui. (2022). A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*.
- Zhang, X., Chowdhury, R. R., Gupta, R. K., y Shang, J. (2024). Large language models for time series: A survey. *arXiv preprint arXiv:2402.01801*.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).
- Zhou, T., Niu, P., Sun, L., Jin, R., et al. (2023). One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36, 43322-43355.
- Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., y He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1), 43-76.
- Zivot, E., y Andrews, D. W. K. (2002). Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *Journal of business & economic statistics*, 20(1), 25-44.

APÉNDICES

Medidas de similitud

En las secciones dentro de este capítulo presentamos diferentes medidas de distancia para comparar series de tiempo. Estos conceptos permiten cuantificar el grado de similitud o disimilitud entre series, lo cual resulta fundamental para evaluar la efectividad de la transferencia de conocimiento entre dominios. Estas distancias sirven no solo como herramientas diagnósticas, sino también como posibles criterios para predecir de antemano la viabilidad de una transferencia.

Dynamic Time Wrapping (DTW)

En las aplicaciones prácticas no siempre es trivial la selección del conjunto de datos fuente dentro de los potenciales datos disponibles. Por esta razón, es necesario contar con alguna medida de similitud entre series temporales para poder realizar esta selección.

En el contexto de series temporales, introducimos entonces Dynamic Time Wrapping, cuyo objetivo es realizar una alineación temporal entre dos series. Originalmente esta medida fue introducida por Hiroaki Sakoe y Seibi Chiba en 1978 [46, Sakoe y Chiba, 1978], en el contexto de la investigación del reconocimiento de palabras habladas, donde la velocidad de habla provoca cambios no lineales de estos patrones en la fluctuación en el eje temporal. Las estrategias de reconocimiento que han tenido éxito se basan en la capacidad de hacer coincidir aproximadamente las palabras a pesar de las variaciones en el tiempo y la pronunciación.

En el ámbito de las series de tiempo económicas, donde dos economías pueden estar defasadas en términos de los ciclos económicos, resulta relevante poder realizar comparaciones que tengan en cuenta que eventos no recurrentes pueden ocurrir a diferentes velocidades y en distintos momentos en el tiempo. Un claro ejemplo de esto es la pandemia Covid-19.

Una medida de similaridad muy utilizada para comparar distancias es la distancia Euclídeana. Sin embargo, tenemos algunas limitaciones a la hora de utilizarla para comparar series de tiempo, como la imposibilidad de evaluar dos series de diferente longitud. Además, esta distancia no considera la alineación temporal, ya que lo que hace es un mapeo punto a punto de los datos en base al tiempo, independientemente de los valores de las características. Actualmente, DTW se considera el principal punto de referencia frente al cual

deben compararse las nuevas medidas de similitud (X. Wang et al. 2013). Son pocas las propuestas que logran superarlo de forma consistente en distintos tipos de datos, y aquellas que lo hacen suelen ser más complejas y requerir una afinación cuidadosa de uno o varios parámetros.

En este contexto, introducimos DTW. En la figura 1 se muestra de forma visual la alineación entre puntos de la serie X y la serie Y que no necesariamente están alineados temporalmente.

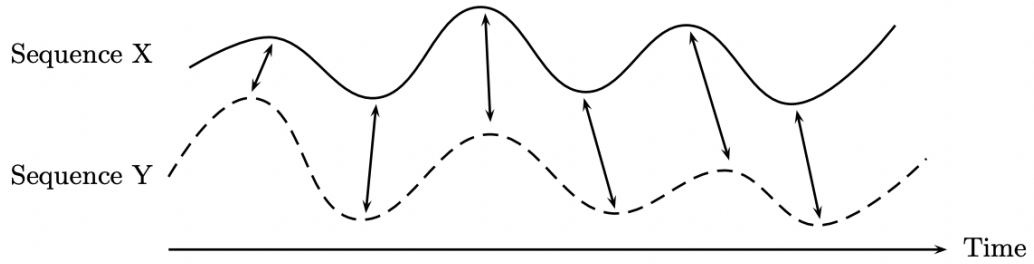


Figura 1: Figura tomada de Muller

Comenzamos definiendo dos series temporales $X = x_1, x_2, \dots, x_N$ y $Y = y_1, y_2, \dots, y_M$, donde $X \in D_T$ y $Y \in D_S$. Notemos que para el cálculo de esta medida N y M no tienen por que ser iguales, pero para el presente trabajo vamos a considerar $N = M$. Independientemente de lo anterior, se debe cumplir que los intervalos temporales sean puntos equidistantes en el tiempo. Si X es una serie medida en intervalos mensuales, Y debe medirse en el mismo intervalo temporal.

El objetivo de esta medida es encontrar la alineación temporal que minimice la distancia entre las dos series. Para esto necesitamos definir una medida de distancia local punto a punto, que llamaremos función de costo local y denotamos como $c(x_n, y_n) = \|x_n - y_n\|$. Cuanto más pequeño sea $c(x_n, y_n)$, más similares resultan x e y (Müller, 2007). A partir del cálculo de esta medida para cada uno de los puntos de las series, definimos una matriz de costos, $C \in \mathbf{R}^{N \times M} / C(n, m) := c(x_n, y_m)$, donde cada entrada de la matriz corresponde con la función de costo local entre dos puntos.

Una vez definida la matriz de costo, vamos a querer encontrar la alineación entre X e Y que minimice el costo total, es decir, la distancia entre las dos series. Definimos entonces $P = (p_1, \dots, p_L)$ donde $p = (n_l, m_l) \in ([1 : N] \times [1 : M]) \forall l \in [1 : L]$, como la trayectoria que corresponde al alineamiento punto a punto de la distancia euclidiana. Por lo tanto, P va a ser una secuencia de

puntos donde cada p se corresponde a un punto $(n, m)_k$.

Cada trayectoria debe cumplir las siguientes restricciones:

1. **Condición de borde:** el primer punto de X debe alinearse con el primer punto de Y , así como el último punto de cada una de ellas también debe estar alineado. Es decir: $p_1 = (1, 1)$ y $p_L = (N, M)$.
2. **Condición de monotonía:** implica que si un elemento en X precede a otro, entonces los elementos correspondientes en Y deben seguir el mismo orden, y viceversa. Es decir: $n_1 \leq n_2 \leq \dots \leq n_L$ y $m_1 \leq m_2 \leq \dots \leq m_L$.
3. **Condición de continuidad:** implica que no se pueden omitir elementos de ninguna de las dos series y todos los pares de índices contenidos en la trayectoria son diferentes. Es decir: $p_{L+1} - p_L \in (1, 0), (0, 1), (1, 1)$ for $\in [1 : L - 1]$.

A continuación, en (Müller, 2007) se ilustra en primer lugar una secuencia que cumple todas las condiciones mencionadas anteriormente (a). En las siguientes b, c y d, se violan las condiciones de borde, monotonía y continuidad, respectivamente.

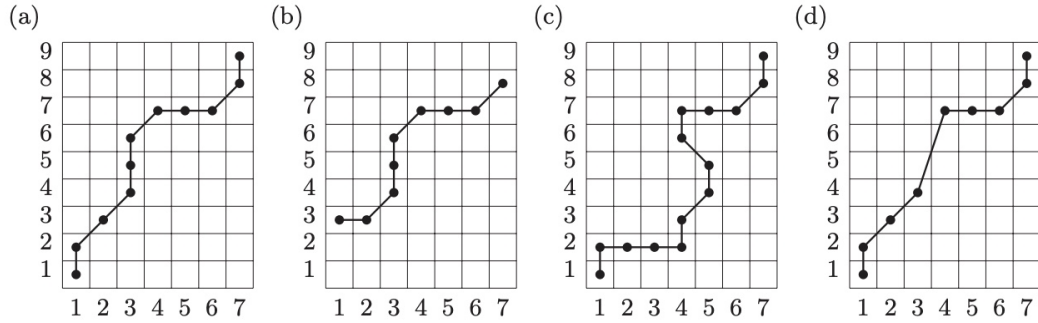


Figura 2: Secuencias de trayectorias que cumplen con las restricciones mencionadas. Figura 4.3 tomada de Muller

Cumplíéndose las condiciones mencionadas, tenemos que para cada trayectoria p el costo total está dado por:

$$c_p(X, Y) := \sum_{l=1}^L c(x_{nl}, y_{nl}) \quad (1)$$

Por lo tanto, la trayectoria óptima p^* va a estar definida por aquella trayectoria que minimice $c_p(X, Y)$. Definimos entonces a la medida de similaridad DTW como:

$$DTW(X, Y) := c_p^*(X, Y) = \min c_p(X, Y) \quad (2)$$

Definido lo anterior, encontrar la trayectoria óptima implicaría calcular todas las trayectorias posibles entre X e Y para luego encontrar aquella que minimiza el costo total. Esto resulta computacionalmente complejo y costoso, creciendo de forma exponencial con el aumento de instancias en cada una de las series. Por lo tanto, es relevante utilizar algoritmos de programación dinámica que resuelvan este problema.

Mediante los algoritmos de programación dinámica, donde se busca descomponer un problema en subproblemas más simples y luego combinar las soluciones de estos subproblemas para llegar a una solución global, se busca derivar una trayectoria de alineación óptima para las secuencias originales a partir de las trayectorias de alineación óptimas para subsecuencias truncadas.

Tenemos que el costo o distancia acumulado, se puede ver como la suma de la distancia del elemento actual y el mínimo de la distancia acumulada de los elementos adyacentes. Es decir, podemos definir $\gamma(n, m)$ como la distancia acumulada entre dichos puntos a partir de la siguiente definición:

$$\gamma(n, m) = c(n, m) + \min \gamma(n-1, m-1), \gamma(n-1, m), \gamma(n, m-1) \quad (3)$$

Por lo tanto,

$$p_{l-1} := \begin{cases} (1, m-1) & \text{si } n = 1 \\ (n-1, 1) & \text{si } m = 1 \\ \operatorname{argmin} D(n-1, m-1), D(n-1, m), D(n, m-1) & \text{en otro caso} \end{cases} \quad (4)$$

La trayectoria óptima p^* se puede calcular a partir de la matriz de costo acumulada, que llamaremos $D(N \times M)$. Dadas las sub secuencias de nuestras series temporales, $X(1:n) := (x_1, \dots, x_n) \forall n \in [1:N]$ y $Y(1:m) := (y_1, \dots, y_m) \forall m \in [1:M]$, tenemos que:

$$D(n, m) := DTW(X(1:n), Y(1:m)) \quad (5)$$

Entonces, la manera de calcular la matriz D , es ir completando cada columna de abajo hacia arriba, comenzando por la columna de más a la izquierda y continuando hacia la derecha, comenzando con $p_L = (N, M)$. Para cada punto

buscamos en cuál de los 3 adyacentes ubicados a la izquierda y abajo hasta llegar a p_1 . Por lo tanto:

En el caso de transfer learning, va a ser de interés calcular la distancia entre nuestra serie *target* y potenciales series *source*. Por lo tanto, se hace necesario buscar algoritmos que nos permitan acelerar los cálculos de la DTW.

Distancia de edición para secuencias reales (EDR)

La distancia de edición para secuencias reales (EDR) corresponde a una extensión de la distancia de Levenshtein original (Levenshtein, 1966) aplicada a series temporales con valores reales. Como señalaron (L. Chen et al. 2005), la EDR superó a las variantes anteriores de la distancia de edición en cuanto a la medición de similitud entre series temporales.

El cálculo de la EDR puede formalizarse mediante un enfoque de programación dinámica. Específicamente, se calcula así:

$$D_{i,j} = \begin{cases} D_{i-1,j-1} & \text{si } m(x_i, y_j) = 1 \\ 1 + \min(D_{i,j-1}, D_{i-1,j}, D_{i-1,j-1}) & \text{si } m(x_i, y_j) = 0 \end{cases} \quad (6)$$

para $i = 1, \dots, M$ y $j = 1, \dots, N$. La función de coincidencia utilizada es:

$$m(x_i, y_j) = \Theta(\epsilon - f(x_i, y_j)), \quad (7)$$

donde $\Theta()$ es la función escalón de Heaviside, tal que $\Theta(z) = 1$ si $z \geq 0$ y 0 en caso contrario, y $\epsilon \in [0, \infty)$ es un parámetro umbral adecuadamente elegido que controla el grado de semejanza entre dos muestras de series temporales que se consideran como coincidencia. La primera fila de D se inicializa con $D_{i,0} = i$ para $i = 0, 1, \dots, M$, y la primera columna con $D_{0,j} = j$ para $j = 0, 1, \dots, N$.

Siguiendo a (L. Chen et al. 2005), quienes reportaron inicialmente algunas mejoras en la precisión de EDR sobre DTW, se establece la función de costo local $f()$ como la diferencia absoluta entre los valores de las muestras, es decir, $f(x_i, y_j) = |x_i - y_j|$. Esto tiene la ventaja adicional de que se puede relacionar fácilmente ϵ con la desviación estándar de las series temporales.

Time-Warped Edit Distance (TWED)

Para series temporales uniformemente muestreadas, la *Time-Warped Edit Distance* (Marteau, 2009) se define recursivamente como:

$$D_{i,j} = \min\{D_{i,j} + \Gamma_{xy}, D_{i-1,j} + \Gamma_x, D_{i,j-1} + \Gamma_y\}, \quad (8)$$

para $i = 1, \dots, M$ y $j = 1, \dots, N$, donde:

$$\Gamma_{xy} = f(x_i, y_j) + f(x_{i-1}, y_{j-1}) + 2\nu|i - j|, \quad (9)$$

$$\Gamma_x = f(x_i, x_{i-1}) + \nu + \lambda, \quad (10)$$

$$\Gamma_y = f(y_j, y_{j-1}) + \nu + \lambda, \quad (11)$$

y $f()$ es una métrica tipo L^n . Como en EDR, seguimos la sugerencia de (Marteau, 2009) y utilizamos $f(x_i, y_j) = |x_i - y_j|$.

El valor final de disimilitud se toma como $d_{\text{TWED}}(x, y) = D_{M,N}$.

TWED presenta dos propiedades distintivas: primero, incorpora diferencias en las marcas temporales, permitiendo comparar series con distintas tasas de muestreo, como aquellas submuestreadas. Segundo, TWED cumple con las propiedades de una métrica (incluyendo la desigualdad triangular), lo que permite utilizar cotas inferiores para acelerar búsquedas por vecinos más cercanos en espacios métricos.

Criterios para la selección de modelos

En esta sección se describen los principales criterios estadísticos utilizados para evaluar y comparar modelos de series temporales, específicamente el AIC (Criterio de Información de Akaike), el BIC (Criterio de Información Bayesiano) y la log-verosimilitud. Estos criterios son fundamentales para seleccionar el modelo que mejor se ajuste a los datos, balanceando la calidad del ajuste y la complejidad.

La log-verosimilitud por sí sola mide la calidad del ajuste, pero tiende a favorecer modelos más complejos, por eso AIC y BIC incluyen términos que penalizan el número de parámetros. En la práctica, es común observar que el AIC favorece modelos ligeramente más complejos que el BIC.

La elección entre AIC y BIC depende del objetivo: para predicción y ajuste flexible se suele preferir AIC, mientras que para identificar el modelo “verdadero” con menos parámetros suele recomendarse BIC. En el presente trabajo, se consideraron ambos criterios para evaluar los modelos ajustados a las series

temporales y seleccionar aquellos que ofrecen un mejor balance entre ajuste y parsimonia, favoreciendo la posterior aplicación de transferencia de aprendizaje.

Log-verosimilitud

La log-verosimilitud es una medida que evalúa qué tan probable es que los datos observados hayan sido generados por un modelo específico con ciertos parámetros. Matemáticamente, si $\mathcal{L}(\theta)$ es la función de verosimilitud para los parámetros θ , la log-verosimilitud es

$$\log \mathcal{L}(\theta) = \sum_{t=1}^n \log f(y_t | \theta),$$

donde $f(y_t | \theta)$ es la función de densidad o probabilidad condicional de la observación y_t dado el modelo y sus parámetros. Un mayor valor de la log-verosimilitud indica un mejor ajuste del modelo a los datos.

Criterio de Información de Akaike (AIC)

El AIC es un criterio que busca equilibrar el ajuste del modelo y su complejidad. Se define como

$$\text{AIC} = 2k - 2 \log \mathcal{L}(\hat{\theta}) \quad (12)$$

donde k es el número de parámetros estimados en el modelo y $\log \mathcal{L}(\hat{\theta})$ es la log-verosimilitud máxima del modelo. El término $2k$ penaliza modelos con más parámetros para evitar el sobreajuste.

Al comparar modelos, se prefiere aquel que minimice el valor de AIC, ya que indica un buen equilibrio entre ajuste y parsimonia.

Criterio de Información Bayesiano (BIC)

El BIC es un criterio similar al AIC pero con una penalización más estricta a la complejidad del modelo, especialmente cuando el tamaño de la muestra es grande. Se calcula como:

$$\text{BIC} = k \log(n) - 2 \log \mathcal{L}(\hat{\theta}) \quad (13)$$

donde n es el número de observaciones. Dado que $\log(n) > 2$ para muestras grandes, el BIC tiende a favorecer modelos más simples que el AIC.

Al igual que con el AIC, se selecciona el modelo con el menor valor de BIC.

Tests de estacionariedad

En la presente sección detallamos algunos tests de estacionariedad relevantes para el presente trabajo.

ADF - test de Dickey-Fuller Aumentado

Para analizar la estacionariedad de una serie temporal, es fundamental determinar si esta posee raíces unitarias, ya que la presencia de una raíz unitaria implica que la serie es no estacionaria y que sus propiedades estadísticas cambian a lo largo del tiempo. Una de las pruebas más utilizadas para este propósito es el test de Dickey-Fuller aumentado (ADF, por sus siglas en inglés) (Said y Dickey, 1984). Este test permite evaluar formalmente si una serie temporal tiene una raíz unitaria, incorporando posibles rezagos de las diferencias para corregir la autocorrelación en los errores. A continuación, se presenta el modelo de regresión utilizado en el test ADF y se explica la interpretación de sus coeficientes, así como la formulación de las hipótesis y la forma de tomar decisiones basadas en los resultados obtenidos.

El modelo de regresión de un test ADF se expresa como:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^p \delta_i \Delta y_{t-i} + \epsilon_t$$

donde:

- y_t es el valor de la serie en el tiempo t ,
- $\Delta y_t = y_t - y_{t-1}$ es la primera diferencia de la serie,
- α es una constante (término independiente),
- βt es un término de tendencia determinística (si se asume que la serie tiene una tendencia),
- γ es el coeficiente de la serie rezagada y_{t-1} ,
- δ_i son los coeficientes de los rezagos de las diferencias de y_t ,
- p es el número de rezagos incluidos en el modelo (determinado por el test de autocorrelación),

- ϵ_t es el término de error blanco.

El test ADF evalúa la presencia de raíz unitaria mediante la hipótesis sobre γ :

$$H_0 : \gamma = 0 \quad (\text{la serie tiene raíz unitaria, no es estacionaria}) \quad (14)$$

$$H_1 : \gamma < 0 \quad (\text{la serie es estacionaria}) \quad (15)$$

A continuación explicamos por qué esa es la hipótesis nula. Consideremos un proceso AR(1) simple según la definición 2.12. Si reescribimos el modelo en diferencias, restando x_{t-1} en ambos lados de la ecuación (y definiendo $\Delta x_t = x_t - x_{t-1}$), se obtiene:

$$\Delta x_t = c + (\phi - 1)x_{t-1} + \epsilon_t. \quad (16)$$

definimos $\gamma = \phi - 1$, entonces:

$$\Delta x_t = c + \gamma x_{t-1} + \epsilon_t \quad (17)$$

Si $\gamma = 0$, implica que la serie diferenciada no depende de su valor pasado, por lo que el proceso es no estacionario con raíz unitaria. El test se basa en la estimación del estadístico γ , y se compara con los valores críticos para determinar si se puede rechazar la hipótesis nula. Si el estadístico de prueba γ es más negativo que el valor crítico, se rechaza la hipótesis nula y se concluye que la serie es estacionaria. El valor p asociado con el estadístico se utiliza para determinar la significancia del test, y si este es menor que el umbral de significancia (típicamente 0.05), se rechaza la hipótesis nula de raíz unitaria.

Prompts - Aplicación: LLM para el Índice de Precios al Consumo en Uruguay

En esta sección se documentan las instrucciones utilizadas en la aplicación de modelos de lenguaje a gran escala al caso del Índice de Precios al Consumo en Uruguay. El propósito es mostrar de manera transparente los prompts utilizados en los experimentos, ya que la adaptación de los LLM a tareas específicas depende en gran medida de la formulación de las instrucciones en

lenguaje natural. Notemos que las instrucciones fueron utilizadas en inglés, dado que los modelos de lenguaje a gran escala alcanzan un mayor desempeño en este idioma, producto de la predominancia de corpus en inglés durante su entrenamiento.

A continuación se presentan las tres instrucciones empleadas: la primera considera únicamente los valores de la serie correspondiente al Interior; la segunda incorpora, además, la serie de Montevideo; y la tercera incluye una cadena de razonamiento junto con los valores de ambas series. Para el primer caso la instrucción fue la siguiente:

```
Input sequence: 101.10, 102.07, 102.90, 103.69, 103.73,
                104.24, 104.54, 104.89, 105.06, 105.54, 106.18, 106.67,
                107.28, 107.59, 107.82, 107.63, 107.71, 107.83,
                108.03, 108.47, 108.75, 109.01, 109.39
```

Para el segundo caso, la instrucción fue la siguiente:

```
Context (Montevideo): 100.92, 101.98, 102.99, 103.74,
                      103.68, 104.58, 104.99, 105.38, 105.87, 106.44, 107.06,
                      107.47, 108.31, 108.53, 108.91, 108.71, 108.85,
                      109.05, 109.48, 109.86, 110.29, 110.82, 111.18
Input sequence (Interior): 101.10, 102.07, 102.90, 103.69,
                           103.73, 104.24, 104.54, 104.89, 105.06, 105.54,
                           106.18, 106.67, 107.28, 107.59, 107.82, 107.63, 107.71,
                           107.83, 108.03, 108.47, 108.75, 109.01, 109.39
```

Por último, utilizamos la instrucción de LSTPrompt (H. Liu et al. [2024](#)):

```
Please continue the following input sequence by addressing
    the task of forecasting IPC Interior. You should break
    down the task into short-term and long-term
    predictions, following a three-step plan.

First, adaptively and reasonably identify the ranges for
    short-term and long-term predictions.
Then, design distinct and correct forecasting mechanisms
    for both short-term and long-term prediction tasks.

For short-term predictions, focus on trends and the last
    few steps of the input sequence.
```

For long-term predictions, emphasize cyclical patterns and statistical properties of the entire input sequence. You may further optimize the forecasting mechanisms based on your observations and domain knowledge.

Finally, correctly implement the forecasting mechanisms, completing predictions one-time step at a time. Remember to take a deep breath after every 2 time steps of prediction.

The input sequence is as follows:

101.10, 102.07, 102.90, 103.69, 103.73, 104.24, 104.54,
104.89, 105.06, 105.54, 106.18, 106.67, 107.28, 107.59,
107.82, 107.63, 107.71, 107.83, 108.03, 108.47,
108.75, 109.01, 109.39