UNIVERSIDAD NACIONAL DE COLOMBIA

PROPUESTA DE TRABAJO DE GRADO

Pronóstico Probabilístico Basado en Predicción Conformal para Series Temporales: Comparación con Bootstrapping y DeepAR

Presentado por:

Pedro Jose Leal Mesa

C.C. 1233910198

Correo: plealm@unal.edu.co

Director Propuesto:

Mario E. Arrieta-Prieto, Ph.D.

Departamento de Estadística

Universidad Nacional de Colombia

Facultad de Ciencias

Departamento de Estadística

Maestría en Estadística

Bogotá D.C., Colombia

14 de julio de 2025

Introducción

La predicción es una tarea fundamental en diversas disciplinas, siendo de particular interés en el análisis de series de tiempo. Tradicionalmente, los pronósticos se han centrado en estimaciones puntuales, como el valor esperado de una variable futura. Sin embargo, este enfoque omite información crucial sobre la incertidumbre asociada a la predicción. Problemas clásicos como el del vendedor de periódicos ilustran la necesidad de ir más allá de la media y considerar la distribución completa de los posibles resultados futuros para una toma de decisiones óptima. La predicción probabilística aborda esta necesidad proporcionando no solo un valor central, sino una distribución de probabilidad o cuantiles sobre los valores futuros.

En este contexto, la predicción conformal (Conformal Prediction - CP) emerge como un marco teórico robusto para la construcción de intervalos y regiones de predicción con garantías de cobertura válidas bajo condiciones mínimas, como la intercambiabilidad de los datos. CP ofrece ventajas significativas: posee validez exacta, es aplicable a cualquier distribución de datos (libre de distribución) y a cualquier modelo predictivo subyacente (libre de modelo) (Vovk, Gammerman y Shafer 2005). No obstante, la aplicación directa de CP a series de tiempo enfrenta el desafío de que la suposición de intercambiabilidad a menudo se viola debido a la dependencia temporal inherente. Investigaciones recientes han comenzado a explorar extensiones de CP para escenarios que van más allá de la intercambiabilidad, como aquellos con derivas de distribución o puntos de cambio.

Este estudio adapta y evalúa un método de predicción conformal para generar pronósticos probabilísticos en series de tiempo. Mediante reglas de puntuación que cuantifican propiedades distribucionales, se contrasta su desempeño con enfoques establecidos, incluyendo modelos neuronales (DeepAR) y técnicas clásicas de remuestreo.

Antecedentes

El pronóstico en series de tiempo ha sido abordado mediante diversas metodologías. Los modelos clásicos ARIMA (Autoregressive Integrated Moving Average) y sus variantes han sido pilares fundamentales, proporcionando un marco teórico sólido para modelar dependencias lineales y generar pronósticos puntuales e intervalos bajo supuestos distribucionales (normalmente Gaussianos).

Con el avance computacional, métodos de remuestreo como el Bootstrapping, y en particular el Sieve Bootstrap, se han adaptado para series de tiempo dependientes. Estos métodos permiten generar intervalos de predicción de forma no paramétrica o semi-paramétrica, relajando los supuestos distribucionales estrictos de los modelos clásicos y mostrando un desempeño competitivo, especialmente en la generación de intervalos más estrechos bajo ciertas condiciones (Arrieta Prieto 2017).

En paralelo, el auge del aprendizaje profundo (Deep Learning) ha introducido modelos potentes como DeepAR (Salinas et al. 2020), que utiliza redes neuronales recurrentes para aprender dependencias temporales complejas y generar directamente pronósticos probabilísticos, modelando los parámetros de una distribución de probabilidad (e.g., Gaussiana, Binomial Negativa) condicionados al pasado. Estos modelos han demostrado alta precisión en diversas aplicaciones, aunque pueden carecer de las garantías teóricas formales de cobertura que ofrece CP.

La teoría de la Predicción Conformal (Vovk, Gammerman y Shafer 2005) establece un marco teórico para la construcción de conjuntos de predicción con garantías de cobertura bajo el supuesto de intercambiabilidad. Investigaciones posteriores han extendido este marco para generar distribuciones predictivas (Vovk, Nouretdinov et al. 2017) y abordar datos no intercambiables, como aquellos que presentan deriva distribucional (Barber et al. 2023), aspecto crucial en aplicaciones de series temporales. Pese a los avances recientes en la implementación de predicción conformal para series temporales, estos se han centrado predominantemente en la predicción de intervalos, relegando el desarrollo de métodos para predicción distribucional a progresos limitados. En cuanto a la evaluación de pronósticos probabilísticos, esta requiere métricas especializadas como el Continuous Ranked Probability Score (CRPS), que cuantifica la discrepancia entre la distribución pronosticada y la realización observada (Gneiting y Raftery 2007).

Este trabajo se sitúa en la intersección de estos enfoques, buscando aprovechar las garantías teóricas de CP adaptadas al dominio temporal y compararlas con la flexibilidad del Bootstrapping y el poder predictivo de modelos de Deep Learning como DeepAR.

Planteamiento del problema

La predicción en series de tiempo enfrenta varios desafíos. Primero, la dependencia temporal viola el supuesto de independencia fundamental en muchos métodos estadísticos estándar. Segundo, la naturaleza de los datos puede cambiar con el tiempo (no estacionariedad, cambios estructurales), lo que complica el modelado y la predicción a largo plazo. Tercero, la necesidad de cuantificar la incertidumbre es crítica en muchas aplicaciones, pero generar pronósticos probabilísticos (intervalos o distribuciones) que sean simultáneamente precisos (calibrados) y eficientes (informativos, e.g., intervalos estrechos) es complejo.

Los métodos existentes tienen limitaciones: los modelos clásicos (ARIMA) dependen de supuestos sobre la estructura del proceso y la distribución del ruido; el Bootstrapping, aunque flexible, puede tener dificultades cerca de límites de no estacionariedad y su validez teórica puede ser asintótica; los modelos de Deep Learning (DeepAR) son potentes pero pueden ser cajas negras y carecer de garantías formales de cobertura de los intervalos generados (Salinas et al. 2020).

La Predicción Conformal (CP) ofrece garantías de cobertura marginal bajo intercambia-

bilidad, pero esta condición es restrictiva para series temporales. Aunque existen extensiones de CP para datos no intercambiables, su aplicación y evaluación sistemática para generar pronósticos temporales y distribucionales en series de tiempo, comparada con referentes relevantes como Bootstrapping y DeepAR, no está completamente explorada dado su reciente auge.

El problema central de este trabajo es: ¿cómo adaptar y evaluar un modelo de predicción probabilística para series de tiempo basado en la teoría conformal, que considere las caracteristicas de dependencia temporal y posible no intercambiabilidad, y cuál es su desempeño comparativo frente a métodos establecidos como Bootstrapping y DeepAR en términos de calidad de predicción probabilística?

Justificación

La cuantificación precisa de la incertidumbre en los pronósticos de series de tiempo es esencial para la toma de decisiones informada en áreas como finanzas, economía, meteorología, gestión de inventarios y planificación de recursos. Los pronósticos puntuales son insuficientes cuando las consecuencias de los errores de predicción son asimétricas en relación con la función de pérdida inherente o cuando se requiere evaluar riesgos.

La Predicción Conformal (CP) ofrece un enfoque atractivo por sus garantías teóricas de cobertura en muestra finita y su aplicabilidad general (libre de distribución y modelo). Adaptar CP al dominio de series de tiempo, abordando el desafío de la dependencia y la posible no intercambiabilidad, representa un avance metodológico importante. Este trabajo busca contribuir a cerrar la brecha entre las garantías teóricas de CP y las necesidades prácticas del pronóstico de series temporales.

Evaluar un modelo CP adaptado frente a métodos de referencia sólidos y diversos como el Bootstrapping (un estándar semi-paramétrico) y DeepAR (un estado del arte en Deep Learning para pronóstico probabilístico) proporcionará una perspectiva clara sobre las fortalezas y debilidades del enfoque conformal en este contexto. La comparación utilizará métricas adecuadas para pronósticos probabilísticos (e.g., CRPS, cobertura y ancho de intervalos).

Los resultados de esta investigación serán de interés tanto para la comunidad académica (aportando al desarrollo de métodos de inferencia confiables para series de tiempo) como para los usuarios que requieren herramientas de pronóstico probabilístico con propiedades teóricas sólidas y buen desempeño empírico. El desarrollo de un modelo CP competitivo podría ofrecer una alternativa valiosa, complementando los enfoques existentes.

Objetivo general y objetivos específicos

Objetivo General:
Formular un modelo fundamentado en la teoría conformal para realizar predicciones probabilísticas en el contexto de series de tiempo, evaluando su desempeño mediante una comparación con modelos establecidos como Bootstrapping y DeepAR.

Objetivos Específicos:

- Desarrollar un modelo de predicción probabilística que integre la teoría de predicción conformal con técnicas de modelado de series temporales.
- Diseñar un entorno de simulación para evaluar el comportamiento del modelo propuesto en diversos escenarios de series temporales, incorporando estructuras temporales y condiciones de ruido variadas.
- Comparar el desempeño del modelo propuesto con modelos de referencia como DeepAR y Bootstrapping en cada escenario simulado, utilizando métricas enfocadas en predicciones probabilísticas.
- Implementar la metodología en un caso de estudio con datos reales, cuantificando su desempeño y relevancia práctica para aplicaciones de pronóstico.

Marco Teórico

El presente marco teórico se enfoca en los conceptos y metodologías fundamentales que sustentan la propuesta de investigación. Se abordarán la predicción probabilística, la teoría de predicción conformal, y los métodos de bootstrapping y aprendizaje profundo (DeepAR) como benchmarks relevantes.

Bootstrapping para Pronósticos en Series de Tiempo

El bootstrapping es una técnica de remuestreo que permite estimar la distribución de un estimador o construir intervalos de predicción sin realizar supuestos distribucionales fuertes, como la normalidad de los errores. En el contexto de series de tiempo, su aplicación busca capturar la incertidumbre inherente al proceso generador de datos y a la estimación del modelo.

Principios Generales del Bootstrapping de Residuos

Una aproximación básica al bootstrapping en series de tiempo es el **bootstrapping de residuos**. Este método, en su forma más simple, asume que los residuos del modelo son no co-

rrelacionados y presentan varianza constante (homocedásticos) (Hyndman y Athanasopoulos 2021). El procedimiento general es:

1. Ajustar un modelo adecuado a la serie de tiempo histórica (e.g., ARIMA, ETS) y obtener los residuos $e_t = y_t - \hat{y}_{t|t-1}$.

2. Para generar una trayectoria futura, se asume que los errores futuros serán similares a los pasados. Una observación futura y^*_{T+1} se simula como $y^*_{T+1} = \hat{y}_{T+1|T} + e^*_{T+1}$, donde $\hat{y}_{T+1|T}$ es el pronóstico puntual y e^*_{T+1} es un residuo muestreado aleatoriamente (con reemplazo) de la colección de residuos históricos.

3. Este proceso se itera para generar trayectorias completas hasta un horizonte h, utilizando los valores simulados previos para generar los siguientes pronósticos puntuales $\hat{y}^*_{T+k|T+k-1}$.

4. Repitiendo este procedimiento múltiples veces, se obtiene una colección de posibles trayectorias futuras. Los intervalos de predicción se derivan de los percentiles de estas trayectorias simuladas para cada horizonte.

Este enfoque no paramétrico es ventajoso porque los intervalos de predicción resultantes no son necesariamente simétricos y pueden reflejar mejor la verdadera distribución de los errores del pronóstico. Sin embargo, la validez del bootstrapping de residuos simple depende crucialmente de que los residuos sean efectivamente no correlacionados y homocedásticos. Si los residuos exhiben dependencia serial (autocorrelación) o heterocedasticidad, el muestreo i.i.d. de los residuos individuales romperá esta estructura, llevando a intervalos de predicción incorrectos o ineficientes (Härdle, Horowitz y Kreiss 2003). En tales casos, se requieren métodos de bootstrapping más sofisticados que puedan preservar la estructura de dependencia de los datos.

Block Bootstrap

Para abordar la dependencia en series de tiempo, el **Block Bootstrap** es uno de los métodos no paramétricos más conocidos y antiguos (Härdle, Horowitz y Kreiss 2003). En lugar de remuestrear observaciones o residuos individuales, el Block Bootstrap divide la serie de tiempo original en bloques de observaciones consecutivas y luego remuestrea estos bloques con reemplazo para construir las series bootstrap.

1. **División en Bloques:** la serie original $\{X_1, \ldots, X_n\}$ se divide en bloques de una longitud fija ℓ. Pueden ser bloques no superpuestos (e.g., el primer bloque es $\{X_1, \ldots, X_\ell\}$, el segundo $\{X_{\ell+1}, \ldots, X_{2\ell}\}$, etc.) o bloques superpuestos (e.g., el primer bloque es $\{X_1, \ldots, X_\ell\}$, el segundo $\{X_2, \ldots, X_{\ell+1}\}$, etc.). El uso de bloques superpuestos suele preferirse ya que utiliza más eficientemente la información disponible (Härdle, Horowitz y Kreiss 2003; Künsch 1989).

2. **Remuestreo de Bloques:** se seleccionan aleatoriamente n/ℓ bloques (o un número suficiente para construir una serie de longitud n) con reemplazo de la colección de bloques formados.

3. **Construcción de Series Bootstrap:** las series bootstrap se forman concatenando los bloques seleccionados en el orden en que fueron muestreados.

La idea fundamental es que, si la longitud del bloque ℓ es suficientemente grande, la estructura de dependencia dentro de cada bloque se preservará, y si ℓ es pequeña en relación con n, la independencia entre bloques remuestreados será una aproximación razonable de la dependencia débil en la serie original. La elección de la longitud del bloque ℓ es crítica: debe aumentar con n para asegurar la consistencia, pero la tasa de convergencia de los errores de estimación del bootstrap de bloques puede ser relativamente lenta (Härdle, Horowitz y Kreiss 2003). Una variante es el *stationary bootstrap*, donde la longitud de los bloques se muestrea de una distribución geométrica, lo que resulta en series bootstrap que son estacionarias (Politis y Romano 1994). Sin embargo, sus errores pueden ser mayores que los del block bootstrap con longitud fija (Härdle, Horowitz y Kreiss 2003).

Sieve Bootstrap

Cuando la estructura de dependencia de la serie de tiempo es más compleja o se desea un método que pueda ofrecer tasas de convergencia más rápidas bajo ciertas condiciones, el **Sieve Bootstrap**, propuesto por Bühlmann (1997) (desarrollado a partir de Bühlmann (1995) y discutido en Härdle, Horowitz y Kreiss (2003)), ofrece una alternativa. Se basa en la idea de que muchos procesos estacionarios pueden ser aproximados (o tamizados) por un modelo autorregresivo (AR) de orden p, donde p puede aumentar con el tamaño de la muestra n (Härdle, Horowitz y Kreiss 2003; Arrieta Prieto 2017). La metodología general es:

1. **Aproximación AR y Estimación:** Se ajusta un modelo AR(p) a la serie de tiempo. El orden p se selecciona típicamente mediante criterios de información (e.g., AIC, AICC), donde p puede ser una función no decreciente de n, $p(n) = o(n)$ (Bühlmann 1995; Arrieta Prieto 2017). Los parámetros del AR(p) seleccionado se estiman, comúnmente mediante los estimadores de Yule-Walker.

2. **Obtención de Residuos:** Se calculan los residuos del modelo AR(p) ajustado. Estos residuos, $\hat{e}_t = X_t - \sum_{j=1}^{p} \hat{\phi}_j X_{t-j}$, se centran para que tengan media cero. Se obtiene su distribución empírica.

3. **Remuestreo de Residuos y Generación de Series Bootstrap:** Se generan secuencias de residuos e_t^* mediante muestreo i.i.d. con reemplazo de los residuos centrados $\hat{e}_t$. Con cada secuencia de e_t^*, se construye una réplica bootstrap de la serie X_t^* utilizando

la ecuación del modelo AR(p) estimado:

$$X_t^* = \sum_{j=1}^{p} \hat{\phi}_j X_{t-j}^* + e_t^*$$

Los primeros p valores de X_t^* se pueden tomar de los valores originales.

4. **Modificación para Pronóstico Condicional:** Para asegurar que los pronósticos se condicionen adecuadamente a la historia observada, Alonso, Peña y Romo (2002) sugieren reemplazar las últimas p observaciones de cada serie bootstrap X_t^* con los valores originales de la serie X_t (Arrieta Prieto 2017).

5. **Generación de Distribuciones de Pronóstico:** Para cada serie bootstrap X_t^*, se reestiman los parámetros del modelo AR(p) y se generan pronósticos para el horizonte h. Al repetir este proceso un gran número de veces, se obtiene una distribución empírica de los valores futuros X_{T+h}^*. Los intervalos de predicción y, potencialmente, la distribución predictiva completa, se derivan de esta colección de trayectorias simuladas.

El Sieve Bootstrap es particularmente útil porque no requiere que el verdadero proceso generador de datos sea un AR de orden finito; la aproximación AR actúa como un tamiz para capturar la estructura de dependencia. Se ha demostrado que este método puede proporcionar mejoras en la precisión de los intervalos de predicción, especialmente generando intervalos más estrechos en comparación con enfoques clásicos bajo ciertas condiciones (Arrieta Prieto 2017; Härdle, Horowitz y Kreiss 2003). Para procesos lineales, el Sieve Bootstrap es a menudo considerado uno de los mejores métodos de bootstrap (Härdle, Horowitz y Kreiss 2003).

DeepAR para Pronóstico Probabilístico

En los últimos años, los modelos de aprendizaje profundo, y en particular las redes neuronales recurrentes (RNN), han demostrado ser herramientas poderosas para el pronóstico de series de tiempo, especialmente en escenarios con grandes cantidades de series relacionadas. DeepAR, propuesto por Salinas et al. (2020), es una metodología destacada en este ámbito que se enfoca en la generación de pronósticos probabilísticos precisos.

El objetivo de DeepAR es modelar la distribución condicional $P(z_{i,t_0:T}|z_{i,1:t_0-1}, \mathbf{x}_{i,1:T})$, donde $z_{i,t_0:T} = \{z_{i,t_0}, \ldots, z_{i,T}\}$ representa los valores futuros de la serie i desde el tiempo t_0 hasta T, $z_{i,1:t_0-1}$ es su pasado conocido, y $\mathbf{x}_{i,1:T}$ son covariables (features) conocidas para todos los instantes. DeepAR logra esto factorizando la distribución conjunta como un producto de verosimilitudes condicionales univariadas:

$$Q_\Theta(z_{i,t_0:T}|z_{i,1:t_0-1}, \mathbf{x}_{i,1:T}) = \prod_{t=t_0}^{T} Q_\Theta(z_{i,t}|z_{i,1:t-1}, \mathbf{x}_{i,1:t}) = \prod_{t=t_0}^{T} \ell(z_{i,t}|\theta(h_{i,t}, \Theta)) \tag{1}$$

donde $\ell(z_{i,t}|\theta)$ es una función de verosimilitud fija (e.g., Gaussiana, Binomial Negativa) cuyos parámetros θ son la salida de una red neuronal. La red neuronal en sí misma es un modelo

autorregresivo recurrente, donde el estado oculto $h_{i,t}$ se actualiza en cada paso de tiempo:

$$h_{i,t} = h(h_{i,t-1}, z_{i,t-1}, \mathbf{x}_{i,t}, \Theta) \tag{2}$$

Aquí, h es una función implementada por una red neuronal recurrente multicapa (comúnmente con celdas LSTM), Θ representa todos los parámetros del modelo (pesos de la RNN, embeddings, etc.), $z_{i,t-1}$ es el valor real de la serie en el instante anterior (durante el entrenamiento) o un valor muestreado de la predicción anterior (durante la inferencia), y $\mathbf{x}_{i,t}$ son las covariables en el tiempo t.

Las ideas fundamentales de DeepAR son:

1. **Modelo Autoregresivo Recurrente:** Como se ve en la Ecuación 2, el estado oculto $h_{i,t}$ (que resume la información histórica relevante) se calcula recurrentemente. Este estado $h_{i,t}$ se utiliza luego para calcular los parámetros $\theta(h_{i,t}, \Theta)$ de la distribución de probabilidad de $z_{i,t}$ (Ecuación 1).

2. **Pronóstico Probabilístico Directo:** DeepAR modela directamente los parámetros de una función de verosimilitud $\ell(z|\theta)$ para cada punto futuro. Por ejemplo, para datos reales, se puede usar una verosimilitud Gaussiana $\mathcal{N}(z|\mu, \sigma)$, donde los parámetros $\mu(h_{i,t})$ y $\sigma(h_{i,t})$ son salidas de la red (e.g., $\mu(h_{i,t}) = \mathbf{w}_\mu^T h_{i,t} + b_\mu$ y $\sigma(h_{i,t}) = \text{softplus}(\mathbf{w}_\sigma^T h_{i,t} + b_\sigma)$ para asegurar $\sigma > 0$). Para datos de conteo positivos, como la demanda, se utiliza frecuentemente la distribución Binomial Negativa $\text{NB}(z|\mu, \alpha)$, parametrizada por su media μ y un parámetro de forma (o sobredispersión) α. En este caso, $\mu(h_{i,t}) = \text{softplus}(\mathbf{w}_\mu^T h_{i,t} + b_\mu)$ y $\alpha(h_{i,t}) = \text{softplus}(\mathbf{w}_\alpha^T h_{i,t} + b_\alpha)$ (Salinas et al. 2020).

3. **Aprendizaje Global y Manejo de Heterogeneidad:** Se entrena un único conjunto de parámetros Θ utilizando todas las series de tiempo disponibles. Esto permite transferir conocimiento entre series. Para manejar la variabilidad en las escalas, el input autorregresivo $z_{i,t-1}$ se divide por un factor de escala ν_i (e.g., el promedio de $z_{i,t}$ en el rango de condicionamiento), y los parámetros de la distribución (como μ) se multiplican por ν_i antes de calcular la verosimilitud. Adicionalmente, se puede usar un muestreo ponderado durante el entrenamiento, donde la probabilidad de seleccionar una ventana de una serie i es proporcional a su escala ν_i.

4. **Generación de Muestras de Monte Carlo:** Para la predicción, después de procesar el rango de condicionamiento $z_{i,1:t_0-1}$ para obtener h_{i,t_0-1}, se generan trayectorias futuras $\tilde{z}_{i,t_0:T}$. Para cada paso $t \geq t_0$, se calcula $h_{i,t} = h(h_{i,t-1}, \tilde{z}_{i,t-1}, \mathbf{x}_{i,t}, \Theta)$ (con $\tilde{z}_{i,t_0-1} = z_{i,t_0-1}$) y luego se muestrea $\tilde{z}_{i,t} \sim \ell(\cdot|\theta(h_{i,t}, \Theta))$. Repitiendo este proceso se obtienen múltiples trayectorias, de las cuales se pueden derivar cuantiles o la distribución predictiva completa de cualquier funcional de interés.

5. **Incorporación de Covariables y Características Fijas:** Las covariables dependientes del tiempo $\mathbf{x}_{i,t}$ se incorporan directamente en la RNN. Características fijas de

cada serie (e.g., categoría de producto) se pueden incorporar aprendiendo un vector de embedding para cada categoría y alimentándolo como una entrada adicional a la RNN en cada paso.

El entrenamiento se realiza maximizando la log-verosimilitud total sobre todas las series y todos los instantes de tiempo en las ventanas de entrenamiento:

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{t=t_0}^{T} \log \ell(z_{i,t} | \theta(h_{i,t}, \Theta)) \tag{3}$$

DeepAR ha demostrado mejoras significativas en la precisión de los pronósticos probabilísticos en diversos conjuntos de datos del mundo real (Salinas et al. 2020), destacando la capacidad de las técnicas de aprendizaje profundo para superar desafíos de los enfoques clásicos en el pronóstico de grandes volúmenes de series de tiempo relacionadas.

Predicción Conformal (Conformal Prediction)

La predicción conformal (CP) es un marco metodológico que permite construir regiones de predicción con garantías de cobertura válidas bajo supuestos mínimos sobre los datos, siendo el más común la intercambiabilidad. A diferencia de muchos métodos tradicionales que dependen de supuestos distribucionales específicos o de la correcta especificación del modelo, CP ofrece robustez y aplicabilidad general (Vovk, Gammerman y Shafer 2005).

Suposiciones Fundamentales y Validez

El supuesto estándar en CP es que las observaciones (ejemplos) $z_1, z_2, \ldots, z_n, z_{n+1}$ son **intercambiables**. Esto significa que la distribución de probabilidad conjunta de la secuencia es invariante bajo cualquier permutación de sus elementos. Una condición más fuerte, pero comúnmente utilizada como caso particular que implica intercambiabilidad, es que las observaciones sean independientes e idénticamente distribuidas (i.i.d.) de alguna distribución desconocida Q (Vovk, Gammerman y Shafer 2005, Cap. 2).

Un **predictor de confianza (confidence predictor)** Γ es una función que, dados los ejemplos de entrenamiento $z_1, \ldots, z_{n-1}$ y un nuevo objeto x_n, produce un conjunto de predicción Γ_n^ϵ para la etiqueta desconocida y_n a un nivel de significancia $\epsilon \in (0, 1)$. El nivel de confianza es $1 - \epsilon$. Se requiere que estos conjuntos de predicción sean anidados: $\Gamma_n^{\epsilon_1} \subseteq \Gamma_n^{\epsilon_2}$ si $\epsilon_1 \geq \epsilon_2$.

La **validez** de un predictor de confianza es una propiedad crucial. Un predictor Γ es *conservativamente válido* si la probabilidad de cometer un error (es decir, $y_n \notin \Gamma_n^\epsilon$) no excede ϵ en cada paso n, bajo cualquier distribución intercambiable P para la secuencia infinita de ejemplos. Formalmente, si $err_n^\epsilon(\Gamma)$ es la variable aleatoria indicadora del error en el paso n al nivel ϵ:

$$err_n^\epsilon(\Gamma) := \mathbb{1}_{y_n \notin \Gamma^\epsilon(z_1, \ldots, z_{n-1}, x_n)} \tag{4}$$

Entonces, la validez conservativa implica que $P(err_n^\epsilon(\Gamma) = 1) \leq \epsilon$. Un predictor Γ es *exactamente válido* si la secuencia de errores $err_1^\epsilon(\Gamma), err_2^\epsilon(\Gamma), \ldots$ es una secuencia de variables aleatorias de Bernoulli independientes con parámetro ϵ (Vovk, Gammerman y Shafer 2005, Cap. 2.1.3). Esto implica que $P(err_n^\epsilon(\Gamma) = 1) = \epsilon$.

Medidas de No Conformidad y Valores p

El núcleo de la predicción conformal reside en el concepto de una **medida de no conformidad** (o inconformidad). Una medida de no conformidad A es una función que asigna una puntuación numérica $A(\mathcal{Z}, z)$ a un ejemplo z, dado un conjunto (o bolsa) de ejemplos $\mathcal{Z}$. Esta puntuación cuantifica cuán extraño o atípico es el ejemplo z en comparación con los ejemplos en $\mathcal{Z}$. Cuanto mayor sea la puntuación, más no conforme se considera el ejemplo.

Dado un conjunto de entrenamiento $z_1, \ldots, z_{n-1}$ y un nuevo objeto x_n con una etiqueta postulada y, se forma una secuencia aumentada $z_1, \ldots, z_{n-1}, z_n = (x_n, y)$. Para cada ejemplo z_i en esta secuencia aumentada (incluyendo el ejemplo de prueba postulado z_n), se calcula una puntuación de no conformidad α_i:

$$\alpha_i := A(\{z_1, \ldots, z_n\}, z_i) \tag{5}$$

Una forma común de definir $A(\mathcal{Z}, z_i)$ es, por ejemplo, la distancia de z_i a su k-ésimo vecino más cercano dentro de $\mathcal{Z}$, o el residuo de z_i respecto a un modelo ajustado sobre $\mathcal{Z} \setminus \{z_i\}$ (deleted residual).

El **valor p** (o p-value) para un ejemplo z_k (en el contexto de la secuencia aumentada $z_1, \ldots, z_n$) se define como la proporción de ejemplos en la secuencia que son al menos tan no conformes como z_k:

$$p_k := \frac{|\{j = 1, \ldots, n : \alpha_j \geq \alpha_k\}|}{n} \tag{6}$$

Predictor Conformal (Transductivo)

Un **predictor conformal (transductivo o completo)** Γ se define a partir de una medida de no conformidad A. Para un nivel de significancia ϵ dado, el conjunto de predicción para y_n (asociado al objeto x_n) es el conjunto de todas las etiquetas posibles $y \in \mathcal{Y}$ tales que el valor p de $z_n = (x_n, y)$ es mayor que ϵ:

$$\Gamma_n^\epsilon(z_1, \ldots, z_{n-1}, x_n) := \left\{ y \in \mathcal{Y} : \frac{|\{i = 1, \ldots, n-1 : \alpha_i \geq \alpha_n^y\}| + 1}{n} > \epsilon \right\} \tag{7}$$

donde $\alpha_i = A(\{z_1, \ldots, z_{n-1}, (x_n, y)\}, z_i)$ para $i = 1, \ldots, n-1$, y $\alpha_n^y = A(\{z_1, \ldots, z_{n-1}, (x_n, y)\}, (x_n, y))$. La adición de '+1' en el numerador es una forma de manejar los empates y asegurar la validez conservativa. Se puede demostrar que los predictores conformales definidos de esta manera son conservativamente válidos (Vovk, Gammerman y Shafer 2005, Prop. 2.3).

Predictores Conformales Suavizados (Smoothed Conformal Predictors)

Para obtener validez exacta, se introduce una aleatorización en el cálculo del valor p. El **valor p suavizado** se define como:

$$p_k^{smooth} := \frac{|\{j = 1, \ldots, n : \alpha_j > \alpha_k\}| + \tau_k |\{j = 1, \ldots, n : \alpha_j = \alpha_k\}|}{n} \tag{8}$$

donde τ_k es una variable aleatoria auxiliar muestreada uniformemente de $[0, 1]$, independiente de todo lo demás. Un **predictor conformal suavizado** Γ_{smooth} utiliza estos p-valores suavizados:

$$\Gamma^{\epsilon}_{n,smooth} := \{y \in \mathcal{Y} : p_n^{smooth}(y) > \epsilon\} \tag{9}$$

donde $p_n^{smooth}(y)$ es el p-valor suavizado calculado para $z_n = (x_n, y)$ en la secuencia aumentada. La propiedad fundamental de los predictores conformales suavizados es que son **exactamente válidos** (Vovk, Gammerman y Shafer 2005, Prop. 2.4). Esto significa que la probabilidad de error es exactamente ϵ en cada paso, y los errores ocurren independientemente entre los pasos, siempre que las variables aleatorias auxiliares τ se generen independientemente en cada paso.

Predictors Conformales Inductivos (ICPs)

Los predictores conformales completos (transductivos), descritos anteriormente, requieren recalcular las puntuaciones de no conformidad para todos los ejemplos de entrenamiento cada vez que se postula una nueva etiqueta para el objeto de prueba. Esto puede ser computacionalmente intensivo, especialmente con grandes conjuntos de datos o medidas de no conformidad complejas. Los **Predictors Conformales Inductivos (ICPs)**, también conocidos como *split conformal predictors*, ofrecen una alternativa computacionalmente más eficiente dividiendo el proceso en etapas de entrenamiento y calibración distintas (Vovk, Gammerman y Shafer 2005, Cap. 4).

El procedimiento general de un ICP es el siguiente:

1. **División de Datos:** El conjunto de datos disponible se divide en dos partes disjuntas:
   - Un **conjunto de entrenamiento propio (proper training set)**: $z_1, \ldots, z_m$.
   - Un **conjunto de calibración (calibration set)**: $z_{m+1}, \ldots, z_l$.

   El tamaño del conjunto de entrenamiento propio es $m < l$, y el del conjunto de calibración es $l - m$.

2. **Entrenamiento del Modelo Subyacente:** Se utiliza una medida de no conformidad inductiva $A(\mathcal{Z}_{train}, z)$, donde $\mathcal{Z}_{train}$ es el conjunto de entrenamiento propio. Esto usualmente implica entrenar un modelo predictivo (e.g., regresión, clasificador) $f_{\mathcal{Z}_{train}}$ únicamente sobre $\mathcal{Z}_{train} = \{z_1, \ldots, z_m\}$. La medida de no conformidad para un ejemplo $z = (x, y)$ es entonces una función de y y $f_{\mathcal{Z}_{train}}(x)$. Por ejemplo, $A(\{z_1, \ldots, z_m\}, (x, y)) = |y - f_{\{z_1,\ldots,z_m\}}(x)|$.

3. **Cálculo de Puntuaciones de No Conformidad de Calibración:** Para cada ejemplo $z_j = (x_j, y_j)$ en el conjunto de calibración ($j = m+1, \ldots, l$), se calcula su puntuación de no conformidad utilizando el modelo entrenado en el paso anterior:

$$\alpha_j := A(\{z_1, \ldots, z_m\}, z_j) \tag{10}$$

   Estas $l - m$ puntuaciones forman el conjunto de referencia para la calibración.

4. **Predicción para un Nuevo Objeto:** Para un nuevo objeto de prueba x_{l+1} y una etiqueta postulada y:

   - Se calcula la puntuación de no conformidad del ejemplo de prueba postulado (x_{l+1}, y) utilizando el mismo modelo $f_{\{z_1,\ldots,z_m\}}$ entrenado con el conjunto de entrenamiento propio:

$$\alpha_{l+1}^y := A(\{z_1, \ldots, z_m\}, (x_{l+1}, y)) \tag{11}$$

   - El valor p (suavizado, para validez exacta) para esta etiqueta postulada y se calcula comparando α_{l+1}^y con las puntuaciones del conjunto de calibración:

$$p^y := \frac{|\{j = m+1, \ldots, l : \alpha_j > \alpha_{l+1}^y\}| + \tau|\{j = m+1, \ldots, l : \alpha_j = \alpha_{l+1}^y\}|}{l - m + 1} \tag{12}$$

     donde $\tau \sim U[0, 1]$ es una variable aleatoria auxiliar. El denominador incluye el ejemplo de prueba postulado.

5. **Construcción del Conjunto de Predicción:** El conjunto de predicción Γ_{l+1}^ϵ para y_{l+1} al nivel de significancia ϵ se forma con todas las etiquetas $y \in \mathcal{Y}$ cuyo valor p p^y es mayor que ϵ:

$$\Gamma_{l+1}^\epsilon := \{y \in \mathcal{Y} : p^y > \epsilon\} \tag{13}$$

La ventaja computacional de los ICPs radica en que el modelo subyacente $f_{\mathcal{Z}_{train}}$ se entrena solo una vez sobre el conjunto de entrenamiento propio. Luego, para cada nuevo objeto de prueba, solo se necesita calcular su puntuación de no conformidad y compararla con el conjunto fijo de puntuaciones de calibración, lo cual es significativamente más rápido que el enfoque transductivo.

Los ICPs suavizados son **exactamente válidos** bajo el supuesto de intercambiabilidad de las $l + 1$ observaciones $(z_1, \ldots, z_l, z_{l+1})$ (Vovk, Gammerman y Shafer 2005, Prop. 4.1). Si no se usa la aleatorización τ (es decir, se usa un p-valor estándar como en la Ecuación 7 pero aplicado al conjunto de calibración), los ICPs son conservativamente válidos.

Sistemas Predictivos Conformales (CPS) y Distribuciones Predictivas

Más allá de los conjuntos de predicción para un nivel de significancia ϵ fijo, la predicción conformal puede extenderse para producir una **distribución predictiva conformal**

(Conformal Predictive Distribution - CPD) completa para la etiqueta de un nuevo objeto. Un CPD es esencialmente una colección de p-valores para todas las etiquetas posibles $y \in \mathcal{Y}$, organizados como una función de distribución acumulada (CDF) o una función de distribución de probabilidad (PDF/PMF) si el espacio de etiquetas es continuo o discreto, respectivamente (Vovk, Gammerman y Shafer 2005, Cap. 7).

Un **Sistema Predictivo Conformal (CPS)** es una función que, dada una secuencia de entrenamiento y un nuevo objeto, produce una CPD. Formalmente, para un conjunto de entrenamiento $z_1, \ldots, z_{n-1}$ y un objeto de prueba x_n, un CPS Π define una función (aleatorizada, si se usan p-valores suavizados) $\Pi_n(y, \tau)$ que para cada etiqueta potencial y y un número aleatorio $\tau \sim U[0,1]$ (independiente de los datos), da el valor $p_n(y, \tau)$. Este $p_n(y, \tau)$ es el p-valor (suavizado) asociado a la hipótesis de que $y_n = y$. La CPD se define como:

$$\Pi_n(y, \tau) := \frac{|\{i = 1, \ldots, n : \alpha_i(y) < \alpha_n(y)\}| + \tau|\{i = 1, \ldots, n : \alpha_i(y) = \alpha_n(y)\}|}{n} \tag{14}$$

donde $\alpha_i(y)$ es la puntuación de no conformidad del i-ésimo ejemplo en la secuencia aumentada $z_1, \ldots, z_{n-1}, (x_n, y)$, y $\alpha_n(y)$ es la puntuación de no conformidad del ejemplo de prueba (x_n, y) en esa misma secuencia aumentada. Nótese que aquí se usa el $<$ en la definición del p-valor, lo cual es típico cuando se trabaja directamente con medidas de *conformidad* (donde valores más altos son mejores) o cuando se define la CPD como una CDF. Si A es una medida de *no conformidad* (valores más altos son peores), la definición de los p-valores se ajusta (e.g., usando $\geq$ y $\alpha_j > \alpha_k$).

La propiedad fundamental es que, si los ejemplos $z_1, \ldots, z_n$ son intercambiables y τ se muestrea uniformemente de $[0,1]$, entonces el valor $p_n(y_n, \tau)$ (donde y_n es la verdadera etiqueta del objeto x_n) sigue una distribución uniforme en $[0,1]$ (Vovk, Gammerman y Shafer 2005, Teor. 11.1). Esto implica que la CPD está **probabilísticamente calibrada**. A partir de una CPD, se pueden derivar conjuntos de predicción para cualquier nivel ϵ como $\Gamma^\epsilon_n = \{y \in \mathcal{Y} : \Pi_n(y, \tau) > \epsilon\}$.

Máquina de Predicción de Mínimos Cuadrados (LSPM)

Un ejemplo importante de un CPS es la **Máquina de Predicción de Mínimos Cuadrados (Least Squares Prediction Machine - LSPM)**. Esta se aplica a problemas de regresión ($Y = \mathbb{R}$) y utiliza los residuos de una regresión por mínimos cuadrados como base para la medida de no conformidad. Se consideran tres versiones principales (Vovk, Gammerman y Shafer 2005, Cap. 7.3):

1. **LSPM Ordinaria:** La medida de no conformidad es el residuo ordinario. Para una etiqueta postulada y para x_n, se calcula $\hat{y}_n$ usando la regresión por mínimos cuadrados sobre $z_1, \ldots, z_{n-1}, (x_n, y)$. La puntuación de no conformidad para (x_n, y) es $A(z_1, \ldots, (x_n, y)) = y - \hat{y}_n$.

2. **LSPM con Residuos Eliminados (Deleted LSPM):** La puntuación de no conformidad para (x_n, y) es $y - \hat{y}_{(n)}$, donde $\hat{y}_{(n)}$ es la predicción para y obtenida de una regresión por mínimos cuadrados sobre $z_1, \ldots, z_{n-1}$ (es decir, sin incluir (x_n, y) en el ajuste para predecir y_n).

3. **LSPM Studentizada:** Esta es a menudo la versión más útil y robusta. La puntuación de no conformidad es el residuo studentizado:

$$A(z_1, \ldots, (x_n, y)) = \frac{y - \hat{y}_n}{\sqrt{1 - h_{nn}}} \tag{15}$$

donde $\hat{y}_n$ es la predicción de mínimos cuadrados ordinarios para y (ajustada con (x_n, y) incluido) y h_{nn} es el n-ésimo elemento diagonal de la matriz sombrero $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, donde $\mathbf{X}$ es la matriz de diseño que incluye x_n. El término $\sqrt{1 - h_{nn}}$ ajusta por la varianza del residuo.

Para la LSPM studentizada, la función $\Pi_n(y, \tau)$ (la CPD) es monotónicamente creciente en y siempre que $\max_i h_{ii} < 1$, lo cual usualmente se cumple (Vovk, Gammerman y Shafer 2005, Prop. 7.8). Esto asegura que sea una función de distribución válida. Una representación explícita de la CPD para la LSPM studentizada puede derivarse. Si C_i son los valores de y que igualan las puntuaciones de no conformidad del ejemplo de prueba postulado con las de los ejemplos de calibración (ajustados por sus respectivos h_{jj}), y estos se ordenan como $C_{(1)} \leq \cdots \leq C_{(n-1)}$, la CPD $\Pi_n(y)$ es una función escalonada con saltos en estos puntos $C_{(j)}$. Específicamente, si $y \in (C_{(j)}, C_{(j+1)})$, entonces $\Pi_n(y, \tau)$ toma valores en $[j/n, (j+1)/n]$ (dependiendo de τ).

Consistencia Universal de los Sistemas Predictivos Conformales

Un resultado teórico importante es la existencia de **sistemas predictivos conformales universalmente consistentes**. Un sistema predictivo aleatorizado Q es consistente para una medida de probabilidad P en $\mathcal{Z}$ si, para cualquier función continua acotada $f : \mathbb{R} \to \mathbb{R}$:

$$\int f dQ_n - E_P(f|x_{n+1}) \to 0 \quad (n \to \infty) \tag{16}$$

en probabilidad, donde Q_n es la CPD generada por el sistema para y_{n+1} dado x_{n+1} y los datos de entrenamiento $z_1, \ldots, z_n$. Un sistema es universalmente consistente si es consistente para cualquier P. Vovk, Gammerman y Shafer (2005, Teor. 31, Cap. 4 del libro original, pero aquí se refiere a Cap. 7 del libro de 2022 para CPDs) demuestra que si el espacio de objetos $\mathcal{X}$ es Borel estándar, existe un sistema predictivo conformal universal. Esto implica que la distancia de Lévy entre la CPD y la verdadera distribución condicional de la etiqueta del objeto de prueba converge a cero. Un ejemplo de tal sistema es el *histogram Mondrian predictive system* (Vovk, Gammerman y Shafer 2005, Sec. 3.2 y 4.2). Este resultado es significativo porque establece que las CPDs pueden aproximar la verdadera distribución condicional subyacente de manera no paramétrica y con garantías de validez.

La construcción de un sistema predictivo conformal universal (como el histograma Mondrian) implica el uso de una medida de no conformidad y una taxonomía (partición del espacio de características) que se refina a medida que aumenta el tamaño de la muestra. Para la medida de no conformidad trivial $A(z_1, \ldots, z_n, (x_{n+1}, y_{n+1})) = y_{n+1}$, y una taxonomía basada en dividir el espacio de objetos en celdas de un histograma cada vez más finas (donde los p-valores se calculan solo con ejemplos dentro de la misma celda que x_{n+1}), se puede demostrar la consistencia universal (Vovk, Gammerman y Shafer 2005, Sec. 3.3 y 4.3).

Predicción Conformal más allá de la Intercambiabilidad: Enfoque Ponderado

Si bien la predicción conformal tradicional (tanto completa/transductiva como inductiva) ofrece garantías de cobertura bajo el supuesto de intercambiabilidad de los datos, este supuesto a menudo se viola en la práctica, especialmente en series de tiempo donde la distribución de los datos puede derivar con el tiempo (distribution drift) o donde existen otras formas de dependencia temporal (Barber et al. 2023). Cuando la intercambiabilidad no se cumple, los métodos conformales estándar pueden perder su garantía de cobertura.

Para abordar estas limitaciones, Barber et al. (2023) proponen generalizaciones de la predicción conformal que son robustas a violaciones de la intercambiabilidad, como la deriva de distribución. Un enfoque clave es el uso de **cuantiles ponderados** en la construcción de los conjuntos de predicción. La idea es asignar pesos a las puntuaciones de no conformidad de las observaciones de calibración, dando potencialmente más relevancia a las observaciones recientes o a aquellas que se consideran más pertinentes para el punto de prueba actual.

Consideremos el contexto de un predictor conformal inductivo (ICP). Sea $\{z_1, \ldots, z_m\}$ el conjunto de entrenamiento propio y $\{z_{m+1}, \ldots, z_l\}$ el conjunto de calibración. Para un nuevo objeto x_{l+1}, y una etiqueta postulada y, se calcula la puntuación de no conformidad $\alpha_{l+1}^y = A(\{z_1, \ldots, z_m\}, (x_{l+1}, y))$. Las puntuaciones de no conformidad para el conjunto de calibración son $\alpha_j = A(\{z_1, \ldots, z_m\}, z_j)$ para $j = m+1, \ldots, l$.

En lugar de calcular el cuantil estándar de las puntuaciones $\{\alpha_j\}$, se introduce un conjunto de pesos $w_{m+1}, \ldots, w_l$, donde $w_j \in [0, 1]$ es el peso asignado a la j-ésima observación del conjunto de calibración. Estos pesos son preespecificados y no dependen de los datos de calibración en sí mismos, aunque pueden depender de características como el índice temporal de la observación. El conjunto de predicción modificado $\widehat{C}_n(x_{n+1})$ (usando $n = l + 1$ para la notación del paper de Barber et al., y asumiendo que las puntuaciones de no conformidad son los residuos $R_i = |Y_i - \hat{\mu}(X_i)|$) se construye como:

$$\widehat{C}_n(x_{n+1}) = \left[\hat{\mu}(x_{n+1}) \pm \widehat{Q}_{1-\alpha}\left(\{\alpha_j\}_{j=m+1}^{l}, \{w_j\}_{j=m+1}^{l}\right)\right] \tag{17}$$

donde $\widehat{Q}_{1-\alpha}(\cdot, \cdot)$ es el $(1-\alpha)$-cuantil ponderado de las puntuaciones de no conformidad de calibración, utilizando los pesos w_j. Este cuantil ponderado q_w se define tal que $\sum_{j:\alpha_j \leq q_w} \tilde{w}_j \geq$

$1-\alpha$ y $\sum_{j:\alpha_j<q_w} \tilde{w}_j < 1-\alpha$, donde $\tilde{w}_j$ son los pesos normalizados:

$$\tilde{w}_j = \frac{w_j}{\sum_{k=m+1}^{l} w_k + 1} \quad \text{para } j = m+1, \ldots, l, \quad \text{y} \quad \tilde{w}_{l+1} = \frac{1}{\sum_{k=m+1}^{l} w_k + 1} \tag{18}$$

El término $+1$ en el denominador corresponde al peso (implícitamente 1) del punto de prueba. Para el predictor conformal completo no intercambiable, el conjunto de predicción para una etiqueta y del objeto x_n se define como:

$$\widehat{C}_n(x_n) = \left\{ y \in \mathcal{Y} : \alpha_n^y \leq \widehat{Q}_{1-\alpha}\left(\{\alpha_i^y\}_{i=1}^{n-1} \cup \{\alpha_n^y\}, \{\tilde{w}_i\}_{i=1}^{n} \right) \right\} \tag{19}$$

donde α_i^y son las puntuaciones de no conformidad de la secuencia aumentada $z_1, \ldots, z_{n-1}, (x_n, y)$ y $\tilde{w}_i$ son los pesos normalizados (incluyendo el peso del punto de prueba).

La garantía de cobertura para estos métodos ponderados es entonces:

$$P(Y_{n+1} \in \widehat{C}_n(X_{n+1})) \geq 1 - \epsilon - \sum_{i=1}^{n} \tilde{w}_i \cdot d_{TV}(Z, Z^{(i)}) \tag{20}$$

donde $Z = (Z_1, \ldots, Z_{n+1})$ es la secuencia completa de datos (entrenamiento y prueba), $Z^{(i)}$ es la secuencia con el punto de prueba (X_{n+1}, Y_{n+1}) intercambiado con el i-ésimo punto de entrenamiento (X_i, Y_i), y d_{TV} es la distancia de variación total entre las distribuciones conjuntas de estas secuencias. Este resultado cuantifica la pérdida de cobertura en función de la desviación de la intercambiabilidad, medida por la suma ponderada de las distancias de variación total. Si los datos son, de hecho, intercambiables, entonces $d_{TV}(Z, Z^{(i)}) = 0$ para todo i, y se recupera la cobertura $1-\alpha$. Si la deriva es gradual, se pueden elegir pesos w_i (e.g., $w_i = \rho^{n+1-i}$ para $\rho < 1$) que asignen mayor importancia a los datos recientes, de modo que la suma ponderada de las distancias d_{TV} pueda mantenerse pequeña, preservando así la cobertura cerca del nivel deseado $1-\alpha$ (Barber et al. 2023).

Este enfoque generaliza la predicción conformal para ser más robusta en entornos no estacionarios, lo cual es de particular relevancia para el análisis de series de tiempo.

Comparación de Supuestos: Predicción Conformal Clásica vs. Enfoques para No Intercambiabilidad

La robustez y las garantías teóricas de la predicción conformal (CP) la han convertido en una herramienta atractiva. Sin embargo, es crucial entender los supuestos subyacentes a sus diferentes formulaciones. A continuación, se comparan los supuestos clave de la CP clásica, tal como se presenta en Vovk, Gammerman y Shafer (2005), con las generalizaciones propuestas por Barber et al. (2023) para abordar escenarios que van más allá de la intercambiabilidad.

Predicción Conformal Clásica (Vovk et al.)

La validez de los predictores conformales (tanto completos/transductivos como inductivos) en su formulación original y sus extensiones directas (e.g., CPS, LSPM) se fundamenta principalmente en el siguiente supuesto:

1. **Intercambiabilidad de los Datos:** Se asume que la secuencia completa de observaciones $z_1, z_2, \ldots, z_N$ (incluyendo el ejemplo de prueba z_N cuando se postula su etiqueta) es intercambiable. Esto significa que la distribución de probabilidad conjunta de la secuencia es invariante bajo cualquier permutación de sus índices:

$$P(Z_1 = z_1, \ldots, Z_N = z_N) = P(Z_{\pi(1)} = z_1, \ldots, Z_{\pi(N)} = z_N)$$

   para cualquier permutación π de $\{1, \ldots, N\}$ (Vovk, Gammerman y Shafer 2005, Cap. 2). Una condición más fuerte, pero suficiente para la intercambiabilidad, es que las observaciones sean independientes e idénticamente distribuidas (i.i.d.).

2. **Simetría del Algoritmo (para predictores conformales completos/transductivos):** Aunque no es un supuesto sobre los datos per se, la construcción de los p-valores en la CP completa (Ecuación 6) implica que la medida de no conformidad $A(\mathcal{Z}, z_i)$ trata simétricamente a todos los ejemplos en la bolsa aumentada $\mathcal{Z}$ al calcular la puntuación de z_i. Esto asegura que, bajo intercambiabilidad de los datos, los rangos de las puntuaciones de no conformidad sean uniformes. Para los ICPs, este supuesto se traduce en que el modelo subyacente se entrena en un conjunto separado y luego se aplica de manera uniforme a los datos de calibración y prueba.

Bajo el supuesto de intercambiabilidad, los predictores conformales suavizados garantizan una cobertura marginal exacta $P(Y_{n+1} \in \Gamma^{\epsilon}_{n+1}) = 1 - \epsilon$, y los no suavizados garantizan una cobertura conservativa $P(Y_{n+1} \in \Gamma^{\epsilon}_{n+1}) \geq 1 - \epsilon$. Estas garantías son válidas para cualquier distribución de datos subyacente que satisfaga la intercambiabilidad y para cualquier medida de no conformidad (o modelo subyacente).

Predicción Conformal más allá de la Intercambiabilidad (Barber et al.)

El trabajo de Barber et al. (2023) busca extender la CP a escenarios donde la intercambiabilidad no se sostiene, como es común en series de tiempo con deriva de distribución o en datos con estructuras de dependencia más complejas. Los supuestos y el enfoque cambian de la siguiente manera:

1. **Relajación del Supuesto de Intercambiabilidad:** No se requiere que la secuencia completa de datos $Z_1, \ldots, Z_N$ sea intercambiable. Los datos pueden ser, por ejemplo, independientes pero no idénticamente distribuidos (i.n.i.d.), o incluso pueden presentar alguna forma de dependencia.

2. **Inexistencia de Supuestos Distribucionales (más allá de la estructura de dependencia):** Al igual que la CP clásica, este enfoque generalizado no impone supuestos paramétricos sobre la forma de las distribuciones individuales de los datos.

3. **Introducción de Pesos (para robustez a la no intercambiabilidad):** Se introduce un esquema de ponderación $w_1, \ldots, w_n$ para las n observaciones de calibración (o para

todas las n observaciones en la CP completa ponderada). Estos pesos son predefinidos y tienen como objetivo dar más (o menos) importancia a ciertas observaciones al calcular el cuantil de las puntuaciones de no conformidad. La elección de los pesos depende del conocimiento a priori sobre la naturaleza de la no intercambiabilidad (e.g., pesos decrecientes para datos recientes en presencia de deriva).

4. **Mecanismo de Aleatorización Modificado (para algoritmos no simétricos, no detallado aquí pero mencionado en el paper):** El paper también introduce una nueva técnica de aleatorización para permitir el uso de algoritmos de ajuste de modelos que no tratan los puntos de datos simétricamente (e.g., un modelo que da más peso a las observaciones recientes durante el entrenamiento). Este aspecto no se ha desarrollado en detalle en el presente marco teórico.

Bajo este marco generalizado, la garantía de cobertura se modifica. En lugar de una cobertura exacta $1-\epsilon$, la cobertura alcanzada se relaciona con la desviación de la intercambiabilidad, como se muestra en la Ecuación 20:

$$P(Y_N \in \widehat{C}_N(X_N)) \geq 1 - \alpha - \sum_{i=1}^{N-1} \tilde{w}_i \cdot d_{TV}(Z, Z^{(i)})$$

(adaptando la notación para N puntos en total, donde X_N, Y_N es el punto de prueba). El término $\sum \tilde{w}_i \cdot d_{TV}(Z, Z^{(i)})$ representa el déficit de cobertura debido a la no intercambiabilidad. Si los datos son intercambiables ($d_{TV} = 0$), se recupera la garantía $1-\alpha$. Si la desviación de la intercambiabilidad es pequeña o si los pesos w_i se eligen adecuadamente para mitigar el efecto de esta desviación, la cobertura puede mantenerse cercana al nivel nominal.

Implicaciones del Contraste

La principal diferencia radica en la robustez frente al supuesto de intercambiabilidad. La CP clásica ofrece garantías fuertes (cobertura exacta o conservativa) pero solo si se cumple la intercambiabilidad. Su violación puede llevar a una pérdida significativa de cobertura. El enfoque de Barber et al. (2023), a través de la ponderación, busca mantener una cobertura cercana a la nominal incluso cuando la intercambiabilidad se viola, cuantificando la posible pérdida de cobertura. Esto es especialmente relevante para series de tiempo, donde la estacionariedad (una forma de intercambiabilidad) es un supuesto fuerte que a menudo no se cumple debido a tendencias, estacionalidades cambiantes o derivas de distribución. La introducción de pesos permite adaptar la predicción conformal a estos escenarios dinámicos, tratando las observaciones más recientes o relevantes con mayor importancia. La elección de los pesos se convierte en un nuevo elemento de modelado, que debe basarse en el conocimiento del dominio o en la naturaleza esperada de la no intercambiabilidad.

Evaluación de Pronósticos Probabilísticos: Calibración, Nitidez y Reglas de Puntuación

La evaluación del desempeño de los pronósticos probabilísticos es una tarea fundamental para comparar diferentes modelos y para entender las fortalezas y debilidades de un método de pronóstico particular. Un buen pronóstico probabilístico no solo debe ser preciso, sino también confiable en la cuantificación de la incertidumbre. La evaluación se basa típicamente en el paradigma de maximizar la **nitidez (sharpness)** de las distribuciones predictivas sujeto a que estén **calibradas (calibration)** (Gneiting y Raftery 2007).

Modos de Calibración

La calibración se refiere a la consistencia estadística entre los pronósticos distribucionales y las observaciones que se materializan. Es una propiedad conjunta de las predicciones y los eventos observados. Gneiting y Raftery (2007) distinguen varios modos de calibración:

- **Calibración Probabilística (Probabilistic Calibration):** Es la forma más fuerte y deseable de calibración. Se dice que una secuencia de pronósticos (funciones de distribución acumulada predictivas F_t) está probabilísticamente calibrada con respecto a la secuencia de verdaderos procesos generadores de datos (G_t) si los valores transformados por la integral de probabilidad (PIT, Probability Integral Transform), $p_t = F_t(y_t)$, son indistinguibles de una muestra i.i.d. de una distribución Uniforme(0,1). Formalmente, si $y_t \sim G_t$, la calibración probabilística se cumple si, asintóticamente:

$$\frac{1}{T}\sum_{t=1}^{T} G_t(F_t^{-1}(p)) \to p \quad \text{para todo } p \in (0,1) \tag{21}$$

  Empíricamente, esto se evalúa mediante la uniformidad del histograma de los valores PIT. Un histograma PIT uniforme indica una buena calibración probabilística. Desviaciones sistemáticas, como forma de U (underdispersive, intervalos de predicción demasiado estrechos) o forma de campana (overdispersive, intervalos demasiado anchos), indican problemas de calibración.

- **Calibración de Excedencia (Exceedance Calibration):** Se enfoca en la consistencia de los umbrales. Un pronóstico está calibrado en excedencia si la frecuencia observada con la que un umbral x es excedido coincide con la probabilidad predicha para ese evento, promediado sobre el tiempo. Formalmente:

$$\frac{1}{T}\sum_{t=1}^{T} G_t^{-1}(F_t(x)) \to x \quad \text{para todo } x \in \mathbb{R} \tag{22}$$

- **Calibración Marginal (Marginal Calibration):** Se refiere a la consistencia entre la distribución predictiva promedio y la distribución climatológica observada. Si $\bar{F}(x) =$

$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} F_t(x)$ es la CDF predictiva promedio y $\bar{G}(x) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} G_t(x)$ es la verdadera CDF climatológica, entonces hay calibración marginal si $\bar{F}(x) = \bar{G}(x)$ para todo x. Empíricamente, se compara la CDF empírica de las observaciones con la CDF promedio de las distribuciones predictivas.

- **Calibración Fuerte (Strong Calibration):** Un sistema de pronóstico está fuertemente calibrado si es probabilística, de excedencia y marginalmente calibrado.

La calibración probabilística es la más fundamental y a menudo implica las otras formas bajo ciertas condiciones (Gneiting y Raftery 2007; Thorarinsdottir y Schuhen 2017).

Nitidez (Sharpness)

La nitidez se refiere a la concentración de las distribuciones predictivas y es una propiedad exclusiva de los pronósticos, independiente de las observaciones. Pronósticos más nítidos (es decir, distribuciones más concentradas o intervalos de predicción más estrechos) son preferibles, siempre y cuando estén calibrados (Gneiting y Raftery 2007). La evaluación de la nitidez se puede realizar mediante el análisis de la anchura de los intervalos de predicción o mediante diagramas de nitidez que muestren, por ejemplo, la distribución de las longitudes de los intervalos de predicción centrales al 50 % y 90 %.

Reglas de Puntuación (Scoring Rules)

Para evaluar y comparar de manera integral el desempeño de los pronósticos probabilísticos, se utilizan **reglas de puntuación (scoring rules)**. Una regla de puntuación $S(F, y)$ asigna una recompensa (o penalización) numérica basada en la distribución predictiva F emitida y el valor y que se materializa (Gneiting y Raftery 2007).

- Una regla de puntuación es **propia (proper)** si un pronosticador maximiza su puntuación esperada si y solo si emite su verdadera creencia sobre la distribución de probabilidad del evento futuro.
- Una regla de puntuación es **estrictamente propia (strictly proper)** si el máximo de la puntuación esperada se alcanza únicamente cuando la distribución pronosticada coincide con la verdadera distribución generadora de datos.

Las reglas de puntuación estrictamente propias son deseables porque incentivan la honestidad y la precisión del pronosticador. Evalúan simultáneamente tanto la calibración como la nitidez.

Una regla de puntuación ampliamente utilizada para pronósticos de densidad de variables continuas es el **Continuous Ranked Probability Score (CRPS)**. Para una CDF

predictiva F y una observación y, el CRPS se define como (Gneiting y Raftery 2007):

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - \mathbf{1}\{x \geq y\})^2 dx \tag{23}$$

donde $\mathbf{1}\{\cdot\}$ es la función indicadora. El CRPS generaliza el error absoluto medio (MAE) para pronósticos puntuales y se expresa en la misma unidad que la variable observada. Un valor de CRPS más bajo indica un mejor pronóstico. Se puede interpretar como una medida integrada de las diferencias cuadráticas entre la CDF pronosticada y la CDF empírica de la observación (una función escalón en y). Una representación alternativa y a menudo más conveniente para el cálculo es:

$$\text{CRPS}(F, y) = E_F|X - y| - \frac{1}{2}E_F|X - X'| \tag{24}$$

donde X y X' son variables aleatorias independientes con CDF F, y $E_F[\cdot]$ denota la esperanza bajo F.

Para comparar el desempeño global de dos distribuciones predictivas F y G (donde G podría ser una distribución de referencia o la verdadera distribución empírica de un conjunto de datos de prueba), se puede utilizar el **Expected Continuous Ranked Probability Score (ECRPS)**. El ECRPS extiende el CRPS tomando la esperanza sobre una distribución de referencia G (Gneiting y Katzfuss 2014):

$$\text{ECRPS}(F, G) = E_{Y\sim G}[\text{CRPS}(F, Y)] = \int_{-\infty}^{\infty} \text{CRPS}(F, y) dG(y) \tag{25}$$

Esta formulación permite una comparación integral de la distribución predictiva F contra un conjunto de realizaciones $Y \sim G$, siendo particularmente útil para evaluar modelos que generan pronósticos probabilísticos completos.

La evaluación de pronósticos probabilísticos mediante el análisis conjunto de calibración, nitidez y reglas de puntuación propias ofrece una visión completa del desempeño predictivo, permitiendo una selección de modelos más informada y una mejor comprensión de sus características (Gneiting y Raftery 2007; Thorarinsdottir y Schuhen 2017).

Metodología Propuesta

La presente investigación se desarrollará en varias fases metodológicas interconectadas, diseñadas para formular, implementar y evaluar rigurosamente un nuevo enfoque de pronóstico probabilístico para series de tiempo basado en la predicción conformal, con especial atención a la robustez frente a la no intercambiabilidad inherente a los datos temporales.

Fase 1: Formulación del Modelo Conformal para Series de Tiempo

El núcleo de esta fase será el desarrollo de un modelo predictivo que genere distribuciones de pronóstico completas y calibradas, fundamentado en los principios de los Sistemas Predictivos Conformales (CPS) y adaptado para el dominio de series de tiempo.

1. **Definición de la Medida de No Conformidad (Nonconformity Measure - NCM):** Se investigarán y seleccionarán medidas de no conformidad adecuadas para series de tiempo. Esto podría implicar el uso de residuos de modelos de series de tiempo subyacentes (e.g., ARIMA, modelos de suavizado exponencial, o regresiones con variables rezagadas). Se considerará la posibilidad de que la NCM dependa de un modelo predictivo puntual que capture la estructura temporal de la serie.

2. **Construcción del Sistema Predictivo Conformal (CPS):** Basándose en la NCM seleccionada, se formulará un CPS siguiendo los principios descritos por Vovk, Gammerman y Shafer (2005) (particularmente el Capítulo 7 sobre regresión probabilística y la LSPM como ejemplo). El objetivo será generar una distribución predictiva acumulada (CDF) $F_{n+1}(y|z_1, \ldots, z_n, x_{n+1})$ para la observación futura y_{n+1} dado el nuevo objeto x_{n+1} y el historial $z_1, \ldots, z_n$. Esto se logrará calculando los p-valores conformales para un continuo de etiquetas posibles y.

3. **Adaptación para No Intercambiabilidad (Inspirado en Barber et al.):** Para abordar la violación del supuesto de intercambiabilidad, común en series de tiempo (e.g., por deriva de distribución), se incorporarán mecanismos de ponderación en el cálculo de los cuantiles de las puntuaciones de no conformidad o en la construcción de la CPD. Siguiendo las ideas de Barber et al. (2023), se asignarán pesos w_i a las puntuaciones de no conformidad históricas α_i, donde estos pesos podrían decrecer con la antigüedad de la observación para dar más relevancia a los datos recientes. La formulación de la CPD se ajustará para reflejar esta ponderación, buscando mantener una cobertura cercana a la nominal incluso bajo deriva. Se explorarán diferentes esquemas de ponderación (e.g., exponencial, lineal decreciente) y su impacto en la validez y eficiencia del pronóstico.

4. **Incorporación de Descomposición de Series de Tiempo (Opcional/Exploratorio):** Se considerará la integración de técnicas de descomposición de series de tiempo (tendencia, estacionalidad, residuo) como un preprocesamiento. El modelo conformal se aplicaría entonces a la serie de residuos, y los pronósticos probabilísticos para los residuos se recombinarían con los pronósticos puntuales de la tendencia y la estacionalidad para obtener la distribución predictiva final de la serie original.

Fase 2: Diseño del Entorno de Simulación y Generación de Datos

Para evaluar el desempeño del modelo propuesto, se diseñará un entorno de simulación exhaustivo que permita generar las verdaderas distribuciones predictivas contra las cuales se compararán los modelos.

1. **Generación de Series de Tiempo Sintéticas:** Se simularán múltiples escenarios de series de tiempo utilizando procesos ARMA(p,q) con diferentes órdenes (p,q) y parámetros. Se introducirán diversas distribuciones para el término de error (e_t) del proceso ARMA (e.g., Normal, Uniforme, Exponencial, t-Student, mezclas de Normales)

para evaluar la robustez de los modelos a la no-Gaussianidad del ruido. Se generarán series de tiempo de una longitud adecuada (e.g., 250 observaciones, descartando las primeras 50 para mitigar efectos de inicialización, y usando 180 para entrenamiento y 1 para prueba, como en Arrieta Prieto (2017)).

2. **Construcción de las Distribuciones Predictivas Teóricas (Ground Truth):** Para cada serie simulada X_t y cada horizonte de pronóstico (principalmente a un paso adelante, y_{n+1}), se construirán dos representaciones de la verdadera distribución predictiva:

   - **Distribución Teórica Analítica:** Dado que el proceso subyacente es un ARMA ($X_t = \sum \phi_i X_{t-i} + \sum \theta_j e_{t-j} + e_t$) y la distribución de e_t es conocida (por diseño de la simulación), la distribución condicional de X_{n+1} dados los valores pasados $X_n, X_{n-1}, \ldots$ y los errores pasados $e_n, e_{n-1}, \ldots$ (o sus estimaciones) puede, en muchos casos, derivarse analíticamente. Esta será la $f(y_{n+1}|\mathcal{F}_n, \text{parámetros verdaderos})$. Por ejemplo, si $e_t \sim N(0, \sigma^2)$, entonces $X_{n+1}|\mathcal{F}_n \sim N(\sum \phi_i X_{n+i} + \sum \theta_j e_{n-j}, \sigma^2)$. Se derivarán estas densidades para cada una de las distribuciones de error consideradas.
   - **Distribución Sieve Bootstrap como Verdad No Paramétrica:** Para cada serie generada, se aplicará el Sieve Bootstrap (ajustando un AR(p) a la serie y remuestreando sus residuos) para generar un gran número de trayectorias futuras. La distribución empírica de estas trayectorias servirá como una aproximación no paramétrica de la verdadera distribución predictiva, representando un escenario donde no se conoce la forma paramétrica exacta del proceso de error, solo su estructura de dependencia capturada por el AR(p) del Sieve. Esto permitirá evaluar qué tan bien los modelos se desempeñan cuando la verdad es menos estructurada o no coincide con sus supuestos paramétricos.

3. **Introducción de Deriva de Distribución (Opcional/Exploratorio):** Para algunos escenarios, se podría simular deriva en los parámetros del proceso ARMA o en la distribución del error para evaluar específicamente la efectividad de los mecanismos de ponderación del modelo CP propuesto.

Fase 3: Cálculo de los Pronósticos de los Modelos

Una vez generado el conjunto de datos sintéticos y definidas las distribuciones de referencia, se procederá a generar los pronósticos probabilísticos utilizando los siguientes modelos:

1. **Modelo Conformal Propuesto (MCP):** El modelo desarrollado en la Fase 1, que integra CP con ponderación para no intercambiabilidad, se aplicará a cada serie simulada para generar su distribución predictiva para y_{n+1}.

2. **DeepAR:** Se entrenará y aplicará un modelo DeepAR (Salinas et al. 2020) a las series simuladas para obtener sus pronósticos probabilísticos.

3. **Bootstrapping para Pronóstico:** Se implementará un método de bootstrapping (e.g., Sieve Bootstrap o bootstrapping de residuos de un modelo base, según se defina más específicamente) para generar distribuciones predictivas para cada serie simulada. Este servirá como un benchmark no paramétrico o semi-paramétrico.

Para cada modelo, el resultado será una distribución predictiva (o una colección de muestras de la misma) para el valor futuro de interés en cada serie simulada.

Fase 4: Análisis de Resultados y Evaluación del Desempeño

El desempeño de los tres modelos (MCP, DeepAR, Bootstrapping) se evaluará comparando sus distribuciones predictivas con las distribuciones verdaderas generadas en la Fase 2 (tanto la analítica como la del Sieve Bootstrap).

1. **Métricas de Evaluación:**
   - **ECRPS (Expected Continuous Ranked Probability Score):** Será la métrica principal. Se calculará el ECRPS de cada modelo con respecto a la distribución teórica analítica y también con respecto a la distribución generada por Sieve Bootstrap (considerada como la verdad no paramétrica). Esto permitirá evaluar la precisión y calibración en diferentes contextos de verdad.
   - **Cobertura de Intervalos de Predicción:** Se construirán intervalos de predicción a partir de las distribuciones de cada modelo y se verificará su cobertura empírica con los valores verdaderos simulados.
   - **Estabilidad y Robustez:** Se analizará la media y la desviación estándar del ECRPS a través de los 120 escenarios simulados para cada modelo, para entender su comportamiento general y su sensibilidad a las diferentes condiciones de los datos.

2. **Pruebas de Significación Estadística:** Se aplicarán pruebas como Wilcoxon y Nemenyi para determinar si las diferencias observadas en el ECRPS entre el MCP y los benchmarks son estadísticamente significativas en los diferentes conjuntos de escenarios (e.g., agrupados por tipo de distribución de error o presencia de deriva).

3. **Análisis Comparativo y Conclusiones:** Se analizarán los resultados para:
   - Identificar las fortalezas y debilidades del modelo conformal propuesto.
   - Determinar bajo qué condiciones (distribución del error, deriva) el MCP ofrece un mejor desempeño o mayor robustez que DeepAR y Bootstrapping.
   - Evaluar el impacto de la estrategia de ponderación en el MCP.

- Extraer conclusiones sobre la viabilidad y el potencial de los métodos conformales adaptados para el pronóstico probabilístico en series de tiempo.

Este diseño metodológico permitirá una evaluación exhaustiva y comparativa del modelo propuesto en un amplio espectro de condiciones, facilitando una comprensión profunda de su desempeño y aplicabilidad.

Cronograma

El siguiente cronograma detalla las fases y actividades propuestas para la ejecución de este trabajo de grado, distribuidas en un periodo de 12 meses correspondientes al presente año 2025.

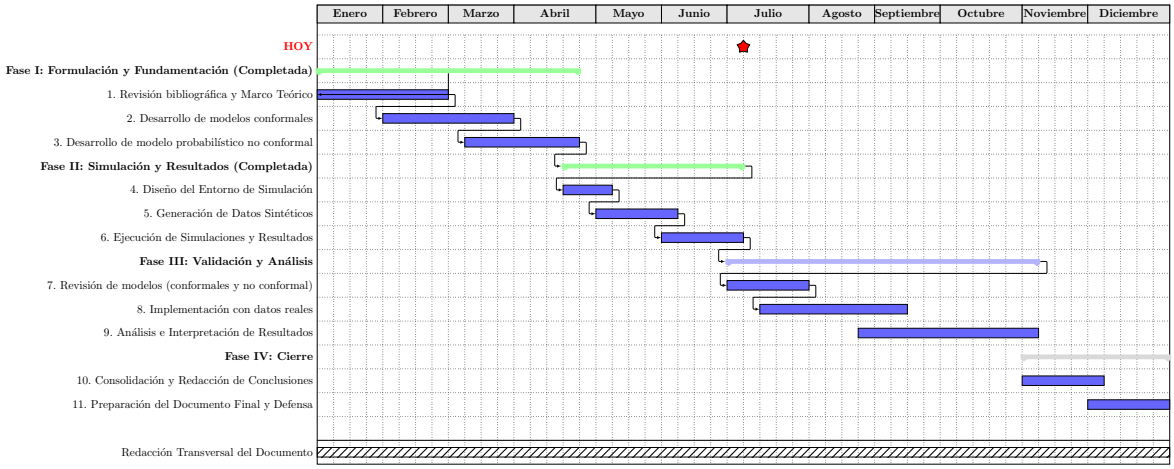

Figura 1: Diagrama de Gantt actualizado mostrando el progreso del trabajo (HOY = punto actual del proyecto).

Referencias

Alonso, A., D. Peña y J. Romo (2002). «Forecasting Time Series With Sieve Bootstrap». En: *Journal of Statistical Planning and Inference* 108.1-2, págs. 185-206. DOI: 10.1016/S0378-3758(02)00269-2.

Arrieta Prieto, Mario Enrique (2017). «Evaluation of the Sieve Bootstrap's performance in comparison with the classic approach for forecasting purposes in time series analysis». En: *XXVII Simposio Internacional de Estadística / 5th International Workshop on Applied Statistics*. Poster Presentation. Medellín, Colombia.

Barber, Rina Foygel et al. (2023). *Conformal Prediction Beyond Exchangeability*. arXiv preprint arXiv:2202.13415v5. Version 5. Accessed on 14 de julio de 2025. arXiv: 2202.13415 [stat.ME].

Bühlmann, Peter (1995). *Sieve Bootstrap for Time Series*. Technical report 431. University of California Berkeley.

— (1997). «Sieve Bootstrap for Time Series». En: *Bernoulli* 3.2, págs. 123-148. DOI: 10.2307/3318610.

Gneiting, Tilmann y Matthias Katzfuss (2014). «Probabilistic forecasting». En: *Annual Review of Statistics and Its Application* 1, págs. 125-151.

Gneiting, Tilmann y Adrian E. Raftery (2007). «Strictly Proper Scoring Rules, Prediction, and Estimation». En: *Journal of the American Statistical Association* 102.477, págs. 359-378. DOI: 10.1198/016214506000001437.

Härdle, Wolfgang K., Joel L. Horowitz y Jens-Peter Kreiss (2003). «Bootstrap Methods for Time Series». En: *International Statistical Review* 71.2, págs. 435-459. DOI: 10.1111/j.1751-5823.2003.tb00485.x.

Hyndman, Rob J y George Athanasopoulos (2021). *Forecasting: Principles and Practice*. 3rd. Melbourne, Australia: OTexts. URL: https://otexts.com/fpp3/.

Künsch, Hans R. (1989). «The Jackknife and the Bootstrap for General Stationary Observations». En: *The Annals of Statistics* 17.3, págs. 1217-1241. DOI: 10.1214/aos/1176347265.

Politis, Dimitris N. y Joseph P. Romano (1994). «The Stationary Bootstrap». En: *Journal of the American Statistical Association* 89.428, págs. 1303-1313. DOI: 10.1080/01621459.1994.10476870.

Salinas, David et al. (2020). «DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks». En: *International Journal of Forecasting* 36.3, págs. 1181-1191. DOI: 10.1016/j.ijforecast.2019.07.001. arXiv: 1704.04110.

Thorarinsdottir, Thordis L. y Nina Schuhen (2017). *Verification: assessment of calibration and accuracy*. Inf. téc. SAMBA/17/17. Norwegian Computing Center.

Vovk, Vladimir, Alexander Gammerman y Glenn Shafer (2005). *Algorithmic Learning in a Random World*. New York: Springer. DOI: 10.1007/b138548.

Vovk, Vladimir, Ilia Nouretdinov et al. (2017). *Conformal predictive distributions with kernels*. Working paper 20. Working Paper 20. Accessed on 14 de julio de 2025. On-line compression modelling project (new series). URL: http://alrw.net/articles/CPDK.pdf.