



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

Pedro José Leal Mesa

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá D.C., Colombia
2026

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

Pedro José Leal Mesa

Tesis presentada como requisito parcial para optar al título de:

Magíster en Ciencias - Estadística

Director:
Ph.D. Mario E. Arrieta-Prieto

Línea de Investigación:
Análisis de Series de Tiempo y Predicción Conformal

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá D.C., Colombia
2026

Dedicatoria

A mi papá Antonio Leal, a mis compañeros y a mi director por su guía y apoyo constante en este proceso de formación.

“Essentially, all models are wrong, but some are useful.”

— George E. P. Box

Agradecimientos

Al Ph.D. Mario E. Arrieta-Prieto, director de esta tesis, por su invaluable orientación y paciencia. Quiero agradecerle especialmente por haberme incentivado a salir de mi zona de confort y presentar los resultados preliminares de este trabajo en el escenario internacional.

Un agradecimiento profundo y especial por la oportunidad de participar en el **45th International Symposium on Forecasting** en China. Esta experiencia marcó un hito en mi carrera profesional y fue posible gracias a un esfuerzo colectivo que jamás olvidaré: a la Universidad Nacional de Colombia por brindarme los fondos institucionales, y muy especialmente a mis compañeros, amigos, conocidos y familiares, quienes con un cariño inmenso participaron en una rifa para completar el dinero necesario para este viaje. Su solidaridad fue el motor que me llevó al otro lado del mundo.

A mis compañeros de la Maestría en Estadística, con quienes compartí este camino académico, por las discusiones técnicas y el apoyo moral en los momentos de mayor reto.

A mi padre, Antonio Leal, por ser mi apoyo incondicional y por creer en este proyecto desde el primer día. Todo este esfuerzo es también suyo.

Con gratitud,
Pedro José Leal Mesa

Contenido

Agradecimientos	v
1. Introducción y Objetivos	1
1.1. Introducción	1
1.2. Planteamiento del problema	2
1.3. Justificación	3
1.4. Objetivos	4
1.4.1. Objetivo General	4
1.4.2. Objetivos Específicos	4
1.5. Estructura de la tesis	4
2. Sistemas de Predicción Conformal	6
2.1. Pronóstico Probabilístico	6
2.1.1. Definición y Objetivos	6
2.1.2. Ventajas del Pronóstico Probabilístico	8
2.2. Métricas para Evaluación de Pronósticos Probabilísticos	9
2.2.1. Reglas de Puntuación Propias	9
2.2.2. Continuous Ranked Probability Score (CRPS)	10
2.2.3. Expected Continuous Ranked Probability Score (ECRPS)	11
2.2.4. Z-Score Normalizado	12
2.2.5. Transformación Integral de Probabilidad (PIT)	13
2.2.6. Curvas de Confiabilidad (Reliability Diagrams)	14
2.3. Test de Diebold-Mariano	14
2.3.1. Formulación del Test	15
2.3.2. Distribución Asintótica y Estimación de la Varianza	15
2.3.3. Modificaciones para Muestras Pequeñas	17
2.3.4. Enfoque de la Teoría Asintótica de Suavizado Fijo	18

<i>Contenido</i>	vii
2.4. Predicción Conformal por Intervalos: El Enfoque IID	19
2.4.1. El Concepto de No-Conformidad	19
2.4.2. Protocolo de Construcción de Intervalos	20
2.5. Robustez ante la No-Intercambiabilidad: Aproximación de Barber	21
2.5.1. La brecha (gap) de Cobertura y Variación Total	21
2.5.2. Cuantiles Ponderados y Decaimiento Temporal	22
2.6. Sistemas de Predicción Conformal (CPS): De Intervalos a Densidades	22
2.6.1. Formalización de la RPD y el Suavizado (τ)	22
2.6.2. La Máquina de Predicción de Mínimos Cuadrados (LSPM)	23
2.7. Sistemas de Predicción Conformal de Mondrian (MCPS)	25
2.7.1. Origen y Motivación: Validez Marginal vs. Condicional	26
2.7.2. La Taxonomía de Mondrian (κ)	27
2.7.3. Integración del Algoritmo MCPS	27
2.8. Vacío Metodológico: De Intervalos a Distribuciones en Series Temporales	28
2.8.1. Estado Actual y Dimensiones del Vacío	28
2.8.2. Contribución y Relevancia	29
2.9. Análisis de la Consistencia Universal de Vovk	30
2.9.1. Definición de Consistencia Universal	30
2.9.2. Definición de Consistencia Universal	30
2.9.3. Mecanismo de la Demostración: El Enfoque de Histograma	31
2.9.4. La Distancia de Lévy y la Convergencia Débil	31
2.9.5. Implicaciones para el LSPM y la Eficiencia	32
2.10. Hacia la Consistencia Universal en Series de Tiempo Ergódicas	32
3. Metodología de Simulaciones	34
3.1. Introducción	34
3.2. Diseño de la Simulación	35
3.2.1. Selección de Escenarios de Evaluación	35
3.2.2. Estructura del Diseño de Simulación base	36
3.2.3. Protocolo de Simulación y Partición de Datos	38
3.3. Procesos Generadores de Datos	39
3.3.1. Procesos ARMA: Escenario Lineal Estacionario	39
3.3.2. Procesos ARIMA: Escenario Lineal No Estacionario	42
3.3.3. Procesos SETAR: Escenario No Lineal Estacionario	44

3.4.	Modelos predictivos	48
3.4.1.	Modelos tomados de la literatura	48
3.4.2.	Modelos adaptados de la literatura	64
3.4.3.	Contribuciones de este trabajo	90
3.5.	Metodo de comparación	90
3.6.	Simulaciones Complementarias	90
3.6.1.	Simulación 1: Impacto de la Diferenciación en ARIMA ($d = 1$)	91
3.6.2.	Simulación 2: Límites de Integración y Persistencia (Multi-D)	92
3.6.3.	Simulación 3: Efectos del Tamaño Muestral Absoluto	93
3.6.4.	Simulación 4: Proporciones de Calibración con Tamaño Fijo	94
3.6.5.	Simulación 5: Predicción Multi-paso (Horizonte h)	95
3.6.6.	Resumen de Evaluaciones Complementarias	98
4.	Resultados y Análisis de Simulaciones	100
4.1.	Resultados del Diseño Principal	100
4.1.1.	Análisis Agregado de Desempeño	101
4.1.2.	Análisis de Robustez y Significancia Estadística	108
4.2.	Simulación 1: Impacto de la Diferenciación en Procesos ARIMA	111
4.2.1.	Motivación Teórica	111
4.2.2.	Resultados Agregados	112
4.2.3.	Caso Excepcional: Sieve Bootstrap	113
4.2.4.	Análisis de Significancia Estadística	115
4.2.5.	Heterogeneidad por Configuración	115
4.2.6.	Implicaciones Metodológicas	118
4.3.	Simulación 2: Límites de Integración y Persistencia Extrema	119
4.3.1.	Motivación: Persistencia Acumulativa	119
4.3.2.	Diseño Experimental	120
4.3.3.	Resultados: Degradación Sistemática por Orden de Integración	120
4.3.4.	Análisis de Sensibilidad: Media vs Variabilidad	122
4.3.5.	Significancia Estadística del Efecto de Diferenciación	123
4.3.6.	Implicaciones para la Práctica	125
4.4.	Simulación 3: Efectos del Tamaño Muestral Absoluto	126
4.4.1.	Convergencia Asintótica: Análisis por Z-scores	126
4.4.2.	Mejora Relativa: Rentabilidad Marginal del Tamaño Muestral	130
4.4.3.	Análisis de Significancia Estadística del Efecto del Tamaño Muestral	134

4.5.	Simulación 4: Proporciones de Calibración con Tamaño Fijo	138
4.5.1.	Motivación: Trade-off entre Ajuste y Calibración	138
4.5.2.	Resultados Agregados: Patrones de Desempeño por Proporción . .	139
4.5.3.	Análisis de Optimalidad: Z-scores por Proporción	142
4.5.4.	Análisis de Significancia Estadística del Efecto de Proporción . .	144
4.5.5.	Implicaciones Metodológicas	148
4.6.	Simulación 5: Degradación en Predicción Multi-paso	149
4.6.1.	Resultados Agregados: Patrones de Degradación por Horizonte . .	150
4.6.2.	Heterogeneidad por Familia de Procesos	151
4.6.3.	Desempeño Comparativo por Escenario	154
4.6.4.	Interacción Configuración \times Horizonte: Heterogeneidad de Degradación	156
4.6.5.	Implicaciones Metodológicas para Aplicaciones de Largo Plazo . .	161
5.	Aplicaciones a Series de Tiempo Reales	163
5.1.	Metodología de Análisis Exploratorio de Datos	164
5.1.1.	Estructura del Protocolo de Análisis	164
5.1.2.	Transformación Box-Cox	165
5.1.3.	Eliminación de Tendencia mediante LOWESS	166
5.1.4.	Análisis de Estacionariedad	166
5.1.5.	Tests de Estacionariedad y Linealidad	167
5.1.6.	Diagnóstico de Residuos	169
5.1.7.	Análisis Espectral	170
5.1.8.	Síntesis de Hallazgos y Configuración de Modelado	171
5.1.9.	Aplicación del Protocolo	172
5.2.	Serie de Consumo Eléctrico: Dataset Electricity	172
5.2.1.	Descripción del Problema y Contexto	172
5.2.2.	Resultados del Análisis Exploratorio	172
5.2.3.	Configuración Experimental	178
5.2.4.	Resultados	178
5.2.5.	Síntesis del Estudio del Dataset Electricity	184
5.3.	Serie de Tráfico Vial: Dataset Traffic	185
5.3.1.	Descripción del Problema y Contexto	185
5.3.2.	Resultados del Análisis Exploratorio	185
5.3.3.	Configuración Experimental	188
5.3.4.	Resultados	188

5.3.5. Síntesis del Estudio del Dataset Traffic	193
5.4. Serie de Tipo de Cambio: Dataset Exchange Rate	196
5.4.1. Descripción del Problema y Contexto	196
5.4.2. Resultados del Análisis Exploratorio	196
5.4.3. Configuración Experimental	200
5.4.4. Resultados	200
5.4.5. Síntesis del Estudio del Dataset Exchange Rate	208
5.5. Conclusiones Generales de las Aplicaciones	210
5.5.1. Síntesis Comparativa de Resultados	211
5.5.2. Lecciones Metodológicas Fundamentales	213
5.5.3. Protocolo de Análisis Exploratorio como Fundamento	216
5.5.4. Implicaciones para Aplicaciones Prácticas	217
5.5.5. Limitaciones y Direcciones Futuras	219
5.5.6. Síntesis Final	221
6. Conclusiones	223
6.1. Conclusiones del Estudio	223
6.2. Limitaciones del Estudio	224
6.3. Investigación Futura	225
A. Marco Teórico para Consistencia Universal en Series de Tiempo Ergódicas	226
A.1. Redefinición del Objetivo de Consistencia	226
A.2. Supuestos Fundamentales del Marco Propuesto	226
A.3. Mecanismo Propuesto: Transductor Conformal por Kernel	227
A.4. Ruta hacia la Demostración de Consistencia	228
A.4.1. Convergencia del Sesgo	228
A.4.2. Convergencia de la Varianza	228
A.5. Discusión y Perspectivas Futuras	228
Bibliografía	230

Lista de Figuras

2-1. Tipos de predicciones y su relación con la incertidumbre.	7
2-2. Piet Mondrian (1872-1944). Composición con rojo, amarillo y azul. Dominio público.	26
4-1. ECRPS promedio por escenario.	101
4-2. Z-scores de ECRPS por configuración.	102
4-3. Z-scores de ECRPS por configuración según familia de procesos.	103
4-4. Z-scores de ECRPS por distribución.	104
4-5. Z-scores de ECRPS por distribución del error según familia de procesos. . .	105
4-6. ECRPS promedio en función de la varianza.	106
4-7. ECRPS en función de la varianza del error.	107
4-8. Coeficiente de variación del ECRPS por modelo.	108
4-9. Test de Diebold-Mariano modificado con corrección de Bonferroni.	109
4-10. ECRPS promedio por método según modalidad de procesamiento.	113
4-11. Mejora porcentual en ECRPS al diferenciar series ARIMA.	114
4-12. Mejora porcentual por configuración ARIMA.	116
4-13. Mejora porcentual por distribución del error.	117
4-14. Mejora porcentual en ECRPS por orden de integración d . Los colores representan la intensidad de la mejora: verde oscuro indica reducciones superiores al 95%; amarillo pálido indica mejoras marginales ($< 20\%$). Sieve Bootstrap mantiene invarianza aproximada para $d \leq 4$ pero requiere diferenciación para $d \geq 5$	121
4-15. Izquierda: Sensibilidad media del ECRPS al incremento de d	123
4-16. P -valores del test de Diebold-Mariano para cada método y orden de integración.	124
4-17. Z-scores de ECRPS por tamaño muestral total.	127
4-18. Z-scores de ECRPS por tamaño muestral según familia de procesos. . . .	129
4-19. Mejora relativa en ECRPS respecto a $N = 120$ para todos los escenarios. .	131

4-20. Mejora relativa en ECRPS respecto a $N = 120$ según familia de procesos.	133
4-21. Evolución del ECRPS por proporción de calibración (todos los escenarios).	139
4-22. Evolución del ECRPS por proporción de calibración según familia de procesos.	141
4-23. Z-scores de ECRPS por proporción de calibración (todos los escenarios).	143
4-24. Z-scores de ECRPS por proporción según familia de procesos.	145
4-25. Evolución del ECRPS por horizonte de pronóstico (todos los escenarios).	150
4-26. Evolución del ECRPS por horizonte según familia de procesos.	152
4-27. ECRPS promedio por escenario en predicción multi-paso.	155
4-28. Interacción configuración \times horizonte (todos los escenarios). Cada línea representa una configuración paramétrica específica.	156
4-29. Interacción configuración \times horizonte (procesos ARMA).	158
4-30. Interacción configuración \times horizonte (procesos ARIMA).	159
4-31. Interacción configuración \times horizonte (procesos SETAR).	160
5-1. Proceso de transformación de la serie de consumo eléctrico.	174
5-2. Distribución de valores CRPS por modelo en el conjunto de prueba del dataset Electricity. La caja representa el rango intercuartílico (IQR), la línea central indica la mediana, y los puntos individuales muestran valores atípicos.	180
5-3. Histogramas de transformación PIT para todos los modelos en el dataset Electricity.	182
5-4. Curvas de confiabilidad para los modelos evaluados en el dataset Electricity.	183
5-5. Proceso de transformación de la serie de tráfico vehicular.	187
5-6. Distribución de valores CRPS por modelo en el conjunto de prueba del dataset Traffic.	189
5-7. Histogramas de transformación PIT para todos los modelos en el dataset Traffic.	192
5-8. Curvas de confiabilidad para los modelos evaluados en el dataset Traffic.	194
5-9. Proceso de transformación de la serie de tipo de cambio.	198
5-10. Distribución de valores CRPS por modelo en el conjunto de prueba del dataset Exchange Rate.	201
5-11. Histogramas de transformación PIT para todos los modelos en el dataset Exchange Rate.	205
5-12. Curvas de confiabilidad para los modelos evaluados en el dataset Exchange Rate.	207

Lista de Tablas

3-1. Configuraciones paramétricas para procesos ARMA	41
3-2. Configuraciones paramétricas para procesos ARIMA	44
3-3. Configuraciones paramétricas para procesos SETAR	48
3-4. Comparativa: Teoría vs. Implementación del CBB	50
3-5. Comparativa: Teoría vs. Implementación del Sieve Bootstrap	54
3-6. Comparativa: Teoría vs. Implementación del LSPM	58
3-7. Comparativa: Teoría vs. Implementación de DeepAR	63
3-8. Comparativa: Teoría vs. Implementación del LSPMW	67
3-9. Comparativa: Teoría vs. Implementación del MCPS	72
3-10. Comparativa: MCPS vs. AV-MCPS	78
3-11. Comparativa Conceptual:AREPD vs Modelos Relacionados	82
3-12. Comparativa: Teoría vs. Implementación de EnCQR-LSTM	88
3-13. Tamaños muestrales evaluados con proporción fija	93
3-14. Proporciones de calibración evaluadas (N=240 fijo)	94
3-15. Resumen de la carga experimental de simulaciones complementarias	98
4-1. Test de Diebold-Mariano: Sin Diferenciación vs Con Diferenciación en ARIMA	116
4-2. Significancia estadística de diferencias por tamaño muestral: resumen por método	135
4-3. Comparaciones significativas (Bonferroni) por escenario	136
4-4. Significancia estadística de diferencias por proporción: resumen por método	146
4-5. Comparaciones significativas (Bonferroni) por escenario: efecto de proporción	147
4-6. Proporción óptima por método y escenario	148
5-1. Resumen de características identificadas en el análisis exploratorio del dataset Electricity y sus implicaciones metodológicas.	177
5-2. Ranking de modelos según desempeño en dataset Electricity.	179

5-3. Ranking de modelos según desempeño en dataset Traffic.	188
5-4. Matriz de valores p del test de Diebold-Mariano modificado para el dataset Traffic.	191
5-5. Ranking de modelos según desempeño en dataset Exchange Rate.	201
5-6. Matriz de valores p del test de Diebold-Mariano modificado para el dataset Exchange Rate.	203
5-7. Comparación sistemática de características estructurales y desempeño de modelos a través de los tres datasets evaluados.	211

1. Introducción y Objetivos

En este capítulo se describe el contexto general de la investigación, enfocándose en la importancia de la cuantificación de la incertidumbre en el análisis de series temporales. Se presenta el planteamiento del problema, los retos de extender la predicción conformal desde el contexto i.i.d. hacia series temporales, y la justificación académica de este trabajo. Finalmente, se definen los objetivos general y específicos que guiaron el desarrollo de la tesis y se presenta la estructura general del documento.

1.1. Introducción

La predicción es una tarea fundamental en diversas disciplinas, siendo de particular interés en el análisis de series de tiempo. Tradicionalmente, la literatura y la práctica se han centrado en estimaciones puntuales, como la media o el valor esperado de una variable futura. Sin embargo, este enfoque omite información crucial sobre la incertidumbre asociada a la predicción. Problemas clásicos, como la optimización de inventarios o la gestión de riesgos financieros, 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 un conjunto de cuantiles sobre los valores futuros. Para datos idéntica e independientemente distribuidos (i.i.d.), uno de los marcos teóricos más robustos es la *Predicción Conformal* (Conformal Prediction - CP) (Vovk, Gamerman, and Shafer 2005). La CP ofrece garantías exactas de cobertura bajo supuestos de intercambiabilidad y es aplicable a diversos modelos predictivos subyacentes. Su principal fortaleza radica en que es libre de distribución: no requiere supuestos paramétricos sobre la forma de las distribuciones y proporciona regiones de predicción con cobertura controlada en muestra finita.

Sin embargo, la CP presenta dos limitaciones fundamentales que motivan esta investigación. Primero, fue desarrollada originalmente para el contexto i.i.d., donde las observaciones no presentan correlación temporal. Segundo, su formulación clásica se enfoca en la construcción de intervalos o regiones de predicción, sin proporcionar distribuciones predictivas completas. Esta tesis aborda precisamente estas dos extensiones: Adaptar la CP desde datos i.i.d. hacia series temporales con dependencia temporal, y expandir su aplicación desde intervalos de predicción hacia la generación de distribuciones predictivas integrales.

La aplicación de CP a series de tiempo presenta desafíos críticos. La dependencia temporal y los posibles desplazamientos distribucionales invalidan el supuesto de intercambiabilidad, base teórica de las garantías de cobertura de CP. Además, mientras que métodos establecidos como el *Bootstrapping* (Lahiri 2003) han sido adaptados para capturar estructuras de dependencia temporal mediante remuestreo no paramétrico, y técnicas modernas de aprendizaje profundo como *DeepAR* (Salinas et al. 2020) modelan dependencias complejas generando distribuciones predictivas automáticamente, la extensión sistemática de CP a este dominio sigue siendo un área poco explorada.

Este trabajo propone adaptar la CP mediante esquemas de ponderación temporal para manejar la no-intercambiabilidad, extenderla hacia la predicción distribucional completa, y evaluar su desempeño comparativo frente a estos métodos de referencia utilizando métricas probabilísticas rigurosas. Con ello, se busca cerrar una brecha relevante en la literatura: la falta de un marco metodológico y de una evaluación sistemática de CP aplicada a series temporales para la obtención de distribuciones predictivas.

1.2. Planteamiento del problema

La predicción en series de tiempo enfrenta retos fundamentales. La naturaleza secuencial de los datos introduce dependencia temporal que viola el supuesto de independencia requerido por la CP en su formulación original. Esta correlación entre observaciones consecutivas afecta tanto la validez de las garantías teóricas como la eficiencia de los métodos de cuantificación de incertidumbre. Adicionalmente, las características de los datos pueden cambiar con el tiempo a través de tendencias, estacionalidad o cambios estructurales, manifestando no estacionariedad que desafía directamente el supuesto de intercambiabilidad de la predicción conformal.

Los métodos actuales presentan compensaciones importantes. Los modelos autorregresivos clásicos dependen de supuestos paramétricos sobre la estructura del proceso y la distribución del ruido. El *Sieve Bootstrapping* (Lahiri 2003), aunque flexible para capturar dependencia temporal, puede tener dificultades cerca de límites de no estacionariedad. Los modelos de aprendizaje profundo como *DeepAR* (Salinas et al. 2020) ofrecen alto poder predictivo y generan distribuciones predictivas completas, pero operan como “cajas negras” y carecen de garantías teóricas formales de cobertura.

El problema central de este trabajo se resume en la siguiente pregunta: ¿cómo adaptar la teoría de predicción conformal desde el contexto i.i.d. hacia series temporales con dependencia temporal, extendiendo su aplicación desde intervalos hacia distribuciones predictivas completas, y cuál es su desempeño comparativo frente a métodos establecidos como Bootstrapping y DeepAR?

1.3. Justificación

La cuantificación precisa de la incertidumbre es esencial para la toma de decisiones informada en áreas como economía, meteorología y planificación de recursos. La Predicción Conformal ofrece garantías de cobertura en muestra finita y aplicabilidad general (libre de distribución) originalmente desarrolladas para el caso i.i.d. Extender estas propiedades al dominio temporal representa un avance metodológico importante.

Existe un vacío notable en la literatura respecto a la aplicación sistemática de CP en series temporales con no-intercambiabilidad. La brecha metodológica que este trabajo busca cerrar es doble. Primero, la mayoría de trabajos en CP se han centrado en intervalos de predicción bajo intercambiabilidad, relegando el desarrollo de métodos para predicción distribucional completa en contextos temporales. Segundo, la comparación rigurosa de CP adaptada mediante ponderación temporal frente a métodos establecidos utilizando métricas probabilísticas apropiadas ha sido escasamente explorada.

La evaluación mediante el promedio del *Continuous Ranked Probability Score* (ECRPS) metrica que se desarrollará en la sección 2.2.3, además pruebas de significancia estadística como el test de Diebold-Mariano permitirán establecer comparaciones robustas. Esta comparación sistemática con métodos estándar (*Sieve Bootstrapping*) y estado del arte (*DeepAR*) proporcionará una perspectiva clara sobre las fortalezas y debilidades del enfoque conformal en escenarios reales y simulados, facilitando la selección informada de

métodos de pronóstico probabilístico en aplicaciones prácticas.

1.4. **Objetivos**

1.4.1. **Objetivo General**

Formular una metodología fundamentada 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.

1.4.2. **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 Sieve 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.

1.5. **Estructura de la tesis**

El presente documento se organiza de la siguiente manera:

Capítulo 2: Sistemas de Predicción Conformal. Se presenta una introducción al pronóstico probabilístico, se explican las principales métricas para evaluar el desempeño de los modelos (incluyendo CRPS, calibración y nitidez), se desarrollan los fundamentos de los sistemas de predicción conformal, y se incluye una discusión sobre la extensión de

la demostración de consistencia universal en escenarios estacionarios y ergódicos.

Capítulo 3: Metodología de Simulaciones. Se detalla el diseño del entorno de simulación, incluyendo la definición de los procesos generadores de datos, la formulación de los modelos predictivos, los cambios necesarios para su aplicación en este trabajo, y las extensiones de las simulaciones para diferentes escenarios.

Capítulo 4: Resultados y Análisis de Simulaciones. Se exponen y analizan los hallazgos derivados del marco experimental descrito en el capítulo anterior. El análisis incluye la comparación del desempeño mediante la métrica ECRPS, el examen del comportamiento de modelos específicos ante diversos escenarios y la validación de los resultados a través de pruebas de significancia estadística empleando el test de Diebold-Mariano.

Capítulo 5: Aplicaciones a Series de Tiempo Reales. Se implementan los métodos desarrollados en casos de estudio con datos reales, se presentan los resultados obtenidos y se discute su relevancia práctica.

Capítulo 6: Conclusiones. Se sintetizan los hallazgos principales del trabajo, se discuten las limitaciones del estudio, y se proponen direcciones para investigación futura.

2. Sistemas de Predicción Conformal

En este capítulo se presentan los fundamentos teóricos que sustentan el desarrollo de esta investigación. Se inicia con una introducción al pronóstico probabilístico y su importancia en el análisis de series temporales, seguido de una discusión detallada sobre las métricas utilizadas para evaluar el desempeño predictivo. Posteriormente, se desarrollan los conceptos fundamentales de la predicción conformal y sus adaptaciones para series de tiempo.

2.1. Pronóstico Probabilístico

El pronóstico probabilístico representa un cambio de paradigma fundamental en la predicción estadística, pasando de estimaciones puntuales a distribuciones de probabilidad completas sobre cantidades futuras de interés (Gneiting and Katzfuss 2014). A diferencia de las predicciones puntuales tradicionales, que proporcionan únicamente un valor esperado o una estimación central, el pronóstico probabilístico cuantifica la incertidumbre asociada a la predicción mediante la especificación de una distribución predictiva completa (Gneiting, Balabdaoui, and Raftery 2007).

2.1.1. Definición y Objetivos

Formalmente, sea Y_{t+h} una variable aleatoria que representa el valor de una serie temporal en el tiempo $t + h$, donde $h > 0$ denota el horizonte de predicción. Un pronóstico probabilístico es una distribución de probabilidad $F_{t+h|t}$ que caracteriza la incertidumbre sobre Y_{t+h} dado el conjunto de información disponible hasta el tiempo t , denotado por \mathcal{F}_t (Gneiting and Katzfuss 2014).

Gneiting y Raftery explican que el objetivo fundamental del pronóstico probabilístico es maximizar la nitidez de las distribuciones predictivas sujeto a calibración. Estos dos

conceptos son fundamentales para entender la calidad de un pronóstico probabilístico (Gneiting and Raftery 2007; Gneiting, Balabdaoui, and Raftery 2007):

- **Calibración:** Se refiere a la concordancia estadística entre las distribuciones predictivas y las observaciones. Una predicción está calibrada si las realizaciones son estadísticamente indistinguibles de muestras aleatorias de las distribuciones predictivas (Thorarinsdottir and Schuhen 2017).
- **Nitidez (*sharpness*):** Se refiere a la concentración de las distribuciones predictivas y es una propiedad exclusiva de los pronósticos. Cuanto más concentradas sean las distribuciones predictivas, mejor, siempre que se mantenga la calibración (Gneiting and Raftery 2007).

La Figura 2-1 ilustra la diferencia entre una predicción puntual, un intervalo de predicción y una distribución predictiva completa.

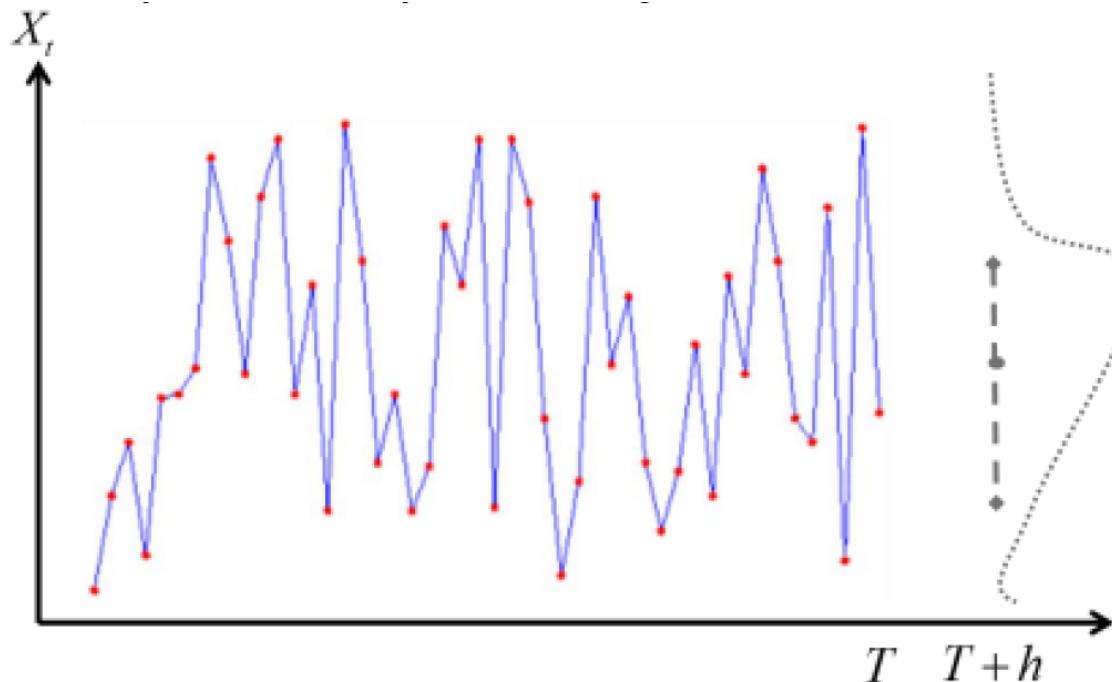


Figura 2-1.: Tipos de predicciones y su relación con la incertidumbre.

2.1.2. Ventajas del Pronóstico Probabilístico

El pronóstico probabilístico ofrece múltiples ventajas sobre las predicciones puntuales tradicionales que justifican su adopción creciente en diversas aplicaciones (Gneiting and Katzfuss 2014). A diferencia de las predicciones puntuales que proporcionan únicamente un valor representativo (típicamente el valor esperado, aunque también puede ser alguna otra estadística de centralidad o de interés para el tomador de decisiones), el pronóstico probabilístico caracteriza la incertidumbre de manera exhaustiva mediante distribuciones de probabilidad completas (Gneiting and Raftery 2007).

Una ventaja fundamental del pronóstico probabilístico sobre los intervalos de predicción es que proporciona la distribución completa en lugar de solo límites de confianza. Mientras que un intervalo de predicción al 95% indica únicamente un rango donde se espera que caiga el valor futuro, la distribución completa permite evaluar tanto la magnitud esperada de un evento como su variabilidad asociada en cualquier región de interés. Esto facilita la optimización de funciones de utilidad esperada en contextos como gestión de inventarios, planificación energética o asignación de capital, donde las decisiones deben considerar explícitamente el trade-off entre riesgo y recompensa.

Además, mientras las predicciones puntuales e intervalos de predicción son inherentemente limitados para caracterizar eventos raros, las distribuciones predictivas permiten estimar probabilidades de eventos extremos, información crucial para la gestión de riesgos financieros y planificación de infraestructura (Thorarinsdottir and Schuhen 2017). Finalmente, las distribuciones predictivas ofrecen flexibilidad comunicativa adaptable a diferentes audiencias mediante intervalos de predicción con diversos niveles de confianza o probabilidades de excedencia de umbrales críticos (Gneiting and Katzfuss 2014).

Estas ventajas han motivado la transición hacia pronósticos probabilísticos en campos tan diversos como meteorología, finanzas, energía, epidemiología y gestión de cadenas de suministro (Gneiting and Katzfuss 2014; Salinas et al. 2020).

2.2. Métricas para Evaluación de Pronósticos Probabilísticos

La evaluación rigurosa del desempeño predictivo es fundamental para comparar metodologías de pronóstico y guiar mejoras en los modelos. En el contexto de pronósticos probabilísticos, las métricas de evaluación deben considerar tanto la calibración como la nitidez de las distribuciones predictivas (Gneiting, Balabdaoui, and Raftery 2007; Thorarinsdottir and Schuhen 2017).

2.2.1. Reglas de Puntuación Propias

Una *regla de puntuación* (*scoring rule*) es una función $S : \mathcal{F} \times \mathbb{R} \rightarrow \mathbb{R} \cup \{-\infty, \infty\}$ que asigna una penalización numérica $S(F, y)$ a cada par formado por una distribución predictiva $F \in \mathcal{F}$ y una observación realizada y (Gneiting and Raftery 2007). En nuestra notación, valores más bajos de la puntuación indican mejor desempeño predictivo.

Propiedad (Propriety) y Propiedad estricta (Strict Propriety)

La propiedad (*propriety*) es una característica fundamental que debe satisfacer toda métrica de evaluación de pronósticos probabilísticos para garantizar que incentive predicciones honestas y bien calibradas (Gneiting and Raftery 2007).

Definición (Regla de Puntuación Propia (Proper)): Una regla de puntuación S es propia relativamente a una clase \mathcal{F} de distribuciones de probabilidad si

$$\mathbb{E}_G[S(G, Y)] \leq \mathbb{E}_G[S(F, Y)] \quad (2-1)$$

para todas las distribuciones $F, G \in \mathcal{F}$, donde $Y \sim G$ (Gneiting and Raftery 2007; Thorarinsdottir and Schuhen 2017).

Intuitivamente, la condición (2-1) establece que una regla de puntuación propia alcanza su valor mínimo esperado cuando la distribución pronosticada coincide con la distribución verdadera de los datos. En otras palabras, si los datos provienen de G , entonces el pronóstico que minimiza la puntuación esperada es precisamente G mismo.

Definición (Regla de Puntuación estrictamente propia (Strictly Proper)): La

regla de puntuación S es estrictamente propia si la desigualdad en (2-1) se cumple con igualdad únicamente cuando $F = G$ (Gneiting and Raftery 2007).

La importancia de la propiedad estricta radica en que establece un principio de alineación de incentivos: si un pronosticador desea minimizar su puntuación esperada, su mejor estrategia es reportar sinceramente su verdadera distribución predictiva (Gneiting and Raftery 2007).

2.2.2. Continuous Ranked Probability Score (CRPS)

El *Continuous Ranked Probability Score* (CRPS) es una de las reglas de puntuación estrictamente propias más utilizadas para evaluar pronósticos probabilísticos de variables continuas (Gneiting and Katzfuss 2014). Su popularidad se debe a su sólida fundamentación teórica y su capacidad para evaluar simultáneamente calibración y nitidez (Gneiting and Raftery 2007; Thorarinsdottir and Schuhen 2017).

Definiciones y Representaciones

El CRPS admite varias representaciones matemáticas equivalentes, cada una con sus propias ventajas conceptuales y computacionales.

Representación integral: La definición original del CRPS está dada por (Gneiting and Raftery 2007):

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - \mathbb{1}\{y \leq x\})^2 dx \quad (2-2)$$

donde F es la función de distribución acumulada (FDA) de la distribución predictiva y y es la observación realizada. Esta representación muestra que el CRPS es proporcional a el área entre la FDA predictiva y la FDA de la observación (Thorarinsdottir and Schuhen 2017).

Representación basada en esperanzas: Una forma alternativa, más conveniente para cálculos, está dada por (Gneiting and Raftery 2007):

$$\text{CRPS}(F, y) = \mathbb{E}_F|X - y| - \frac{1}{2}\mathbb{E}_F|X - X'| \quad (2-3)$$

donde X y X' son variables aleatorias independientes con distribución F . Esta repre-

sentación revela una interpretación intuitiva del CRPS: el primer término mide la distancia esperada entre la distribución predictiva, representada por la variable aleatoria X , y la observación; mientras que el segundo término penaliza la dispersión de la distribución predictiva.

Propiedades del CRPS

El CRPS posee varias propiedades deseables que explican su amplia adopción en la literatura (Gneiting and Raftery 2007; Thorarinsdottir and Schuhen 2017):

1. **Estrictamente propia:** El CRPS es *estrictamente propia* relativo a la clase de todas las distribuciones de probabilidad en \mathbb{R} con primer momento finito (Gneiting and Raftery 2007).
2. **Unidades consistentes:** El CRPS se expresa en las mismas unidades que la variable pronosticada (Gneiting and Katzfuss 2014).
3. **Reducción al error absoluto:** Cuando F es una distribución degenerada (predicción puntual), el CRPS se reduce al error absoluto $|x - y|$, permitiendo un marco de evaluación unificado (Gneiting and Raftery 2007).
4. **Sensibilidad dual:** El CRPS evalúa simultáneamente la calibración y la nitidez (Thorarinsdottir and Schuhen 2017).

2.2.3. Expected Continuous Ranked Probability Score (ECRPS)

El *Expected Continuous Ranked Probability Score* (ECRPS) extiende el CRPS al caso donde se desea cuantificar la discrepancia entre dos distribuciones probabilísticas complejas: una distribución predictiva F y una distribución de referencia G .

Definición del ECRPS

Representación integral: Siguiendo la forma integral del CRPS, el ECRPS se define como:

$$\text{ECRPS}(F, G) = \int_{-\infty}^{\infty} (F(x) - G(x))^2 dx \quad (2-4)$$

donde F y G son las funciones de distribución acumulada de las distribuciones predictiva y de referencia, respectivamente. Esta representación mide el área cuadrática entre ambas FDAs.

Representación basada en esperanzas: De forma análoga a la ecuación (2-3), el ECRPS puede expresarse como:

$$\text{ECRPS}(F, G) = \mathbb{E}_{F,G}|X - Y| - \frac{1}{2}\mathbb{E}_F|X - X'| - \frac{1}{2}\mathbb{E}_G|Y - Y'| \quad (2-5)$$

donde $X, X' \sim F$ son variables aleatorias independientes con distribución F , y $Y, Y' \sim G$ son variables aleatorias independientes con distribución G . Esta forma revela la estructura del ECRPS: el primer término mide la distancia esperada entre ambas distribuciones, mientras que los dos últimos términos penalizan la dispersión de cada distribución.

Estimación empírica

En la práctica, cuando se dispone de muestras $\vec{x} = \{x_1, \dots, x_n\}$ de F y $\vec{y} = \{y_1, \dots, y_m\}$ de G , el ECRPS puede estimarse mediante:

$$\widehat{\text{ECRPS}}(F, G) = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m |x_i - y_j| - \frac{1}{2n^2} \sum_{i=1}^n \sum_{k=1}^n |x_i - x_k| - \frac{1}{2m^2} \sum_{j=1}^m \sum_{\ell=1}^m |y_j - y_\ell| \quad (2-6)$$

El ECRPS hereda las propiedades deseables del CRPS, siendo estrictamente propio y sensible tanto a sesgos sistemáticos como a diferencias en la dispersión entre las distribuciones comparadas.

2.2.4. Z-Score Normalizado

Para facilitar la comparación del desempeño entre diferentes escenarios de simulación con escalas y niveles de dificultad heterogéneos, se emplea la normalización mediante Z-score del ECRPS. Para cada método de pronóstico m evaluado en un conjunto de escenarios \mathcal{S} , el Z-score se define como:

$$Z_m = \frac{\text{ECRPS}_m - \mu_{\mathcal{S}}}{\sigma_{\mathcal{S}}} \quad (2-7)$$

donde $\mu_{\mathcal{S}}$ y $\sigma_{\mathcal{S}}$ representan la media y desviación estándar del ECRPS calculados sobre todos los métodos en el conjunto de escenarios \mathcal{S} .

Esta normalización permite identificar de manera clara en qué escenarios particulares un método presenta dificultades (valores positivos altos) o se desempeña excepcionalmente bien (valores negativos). Por ejemplo, un método conformal con ponderación temporal podría mostrar Z-scores bajos en escenarios estacionarios pero Z-scores elevados ante cambios estructurales abruptos, revelando limitaciones específicas del esquema de ponderación empleado.

2.2.5. Transformación Integral de Probabilidad (PIT)

La Transformación Integral de Probabilidad (PIT, por sus siglas en inglés) constituye una herramienta fundamental para evaluar la calibración probabilística de pronósticos (Dawid 1984; Diebold, Gunther, and Tay 1998). Para una distribución predictiva F_t y una observación realizada y_t , el valor PIT se define como:

$$p_t = F_t(y_t) \quad (2-8)$$

Si los pronósticos están perfectamente calibrados y F_t es continua, entonces $p_t \sim \text{Uniform}(0, 1)$ (Gneiting, Balabdaoui, and Raftery 2007). La evaluación de calibración se realiza mediante histogramas de los valores PIT acumulados sobre el horizonte de evaluación. La forma del histograma revela tipos específicos de descalibración (Thorarinsdottir and Schuhen 2017):

- **Histograma en forma de U:** Indica distribuciones predictivas subdispersas (intervalos de predicción demasiado estrechos).
- **Histograma en forma de ∩:** Indica distribuciones predictivas sobredispersas (intervalos de predicción demasiado amplios).
- **Histograma triangular:** Sugiere sesgo sistemático en las predicciones.
- **Histograma uniforme:** Evidencia calibración probabilística adecuada.

Es importante notar que la uniformidad del PIT es una condición necesaria pero no suficiente para pronósticos ideales, como lo demuestran los ejemplos contraintuitivos de Hamill (Hamill 2001) y la discusión en Gneiting, Balabdaoui, and Raftery (2007). Por esta razón, el análisis PIT debe complementarse con evaluaciones de nitidez.

2.2.6. Curvas de Confiabilidad (Reliability Diagrams)

Las curvas de confiabilidad proporcionan una evaluación visual de la calibración marginal al comparar las probabilidades pronosticadas con las frecuencias observadas (Gneiting, Balabdaoui, and Raftery 2007). Para un conjunto de umbrales $\{u_1, \dots, u_K\}$, se define:

- **Probabilidad pronosticada:** $\hat{p}_k = \frac{1}{T} \sum_{t=1}^T F_t(u_k)$, la probabilidad media de que la observación sea menor o igual al umbral u_k .
- **Frecuencia observada:** $\bar{o}_k = \frac{1}{T} \sum_{t=1}^T \mathbb{1}\{y_t \leq u_k\}$, la proporción de observaciones que efectivamente no excedieron u_k .

La curva de confiabilidad grafica \bar{o}_k versus \hat{p}_k . Para pronósticos perfectamente calibrados, los puntos deberían alinearse sobre la diagonal de 45 grados. Desviaciones sistemáticas indican problemas de calibración:

- Puntos por debajo de la diagonal: el método subestima las probabilidades de no-excedencia (sobreestima eventos extremos).
- Puntos por encima de la diagonal: el método sobreestima las probabilidades de no-excedencia (subestima eventos extremos).

Esta herramienta es particularmente útil para identificar si la descalibración es uniforme a lo largo de toda la distribución predictiva o si se concentra en regiones específicas como las colas (Thorarinsdottir and Schuhen 2017).

2.3. Test de Diebold-Mariano

La evaluación comparativa de distintas metodologías de pronóstico requiere herramientas estadísticas que permitan determinar si las diferencias observadas en el desempeño predictivo son estadísticamente significativas o simplemente producto del azar. El test de Diebold-Mariano (Diebold and Mariano 1995) constituye uno de los procedimientos más ampliamente utilizados para este propósito, ofreciendo un marco general y flexible para contrastar la hipótesis nula de igual precisión predictiva entre dos métodos de pronóstico competidores.

2.3.1. Formulación del Test

Sean $\hat{y}_{t+h}^{(1)}$ y $\hat{y}_{t+h}^{(2)}$ dos pronósticos h pasos adelante para una variable y_{t+h} , producidos por dos metodologías diferentes. Los errores de pronóstico correspondientes son:

$$e_{t+h}^{(i)} = y_{t+h} - \hat{y}_{t+h}^{(i)}, \quad i = 1, 2 \quad (2-9)$$

El test de Diebold-Mariano se basa en una función de pérdida $L(\cdot)$ que cuantifica el costo asociado con cada error de pronóstico. Para un horizonte temporal de evaluación que abarca n observaciones, se define el diferencial de pérdida en el tiempo t como:

$$d_t = L(e_t^{(1)}) - L(e_t^{(2)}), \quad t = 1, \dots, n \quad (2-10)$$

Tradicionalmente, el test de Diebold-Mariano se ha aplicado utilizando la pérdida cuadrática para evaluar estimaciones puntuales. Sin embargo, este marco es suficientemente general para acomodar cualquier función de pérdida (Diebold and Mariano 1995). En el presente trabajo, se utilizará principalmente el CRPS o el ECRPS, según corresponda como métrica de pérdida fundamental. Esto permite extender la comparación de Diebold-Mariano.

La hipótesis nula de igual precisión predictiva se formula como:

$$H_0 : \mathbb{E}[d_t] = 0 \quad (2-11)$$

Esta hipótesis establece que la pérdida esperada es idéntica para ambos métodos de pronóstico. El estadístico de prueba se construye a partir de la media muestral del diferencial de pérdida:

$$\bar{d} = \frac{1}{n} \sum_{t=1}^n d_t \quad (2-12)$$

2.3.2. Distribución Asintótica y Estimación de la Varianza

Bajo condiciones de regularidad de procesos temporales que incluyen la estacionariedad débil y la existencia de momentos de orden finito, Diebold y Mariano demuestran que:

$$\sqrt{n} \bar{d} \xrightarrow{d} N(0, 2\pi f_d(0)) \quad (2-13)$$

donde $f_d(0)$ denota la densidad espectral de la serie d_t evaluada en frecuencia cero, la cual equivale a la varianza de largo plazo:

$$\sigma^2 = \text{Var}(\sqrt{n} \bar{d}) = \gamma_0 + 2 \sum_{k=1}^{\infty} \gamma_k \quad (2-14)$$

siendo $\gamma_k = \text{Cov}(d_t, d_{t-k})$ la autocovarianza de orden k .

Un aspecto fundamental del test de Diebold-Mariano es que permite explícitamente la presencia de autocorrelación en el diferencial de pérdida d_t . Esta característica es especialmente relevante en el contexto de pronósticos a múltiples pasos adelante ($h > 1$), donde los errores de pronóstico exhiben típicamente estructura de autocorrelación hasta el orden $(h - 1)$ (Diebold and Mariano 1995). Esta estructura surge porque pronósticos óptimos h pasos adelante generan errores que siguen un proceso de media móvil MA($h - 1$).

En la práctica, la varianza de largo plazo σ^2 debe ser estimada. Diebold y Mariano proponen utilizar un estimador basado en autocovarianzas ponderadas por kernel:

$$\hat{\sigma}^2 = \hat{\gamma}_0 + 2 \sum_{k=1}^M k \left(\frac{k}{M} \right) \hat{\gamma}_k \quad (2-15)$$

donde $\hat{\gamma}_k = n^{-1} \sum_{t=k+1}^n (d_t - \bar{d})(d_{t-k} - \bar{d})$ son las autocovarianzas muestrales, $k(\cdot)$ es una función kernel (por ejemplo, kernel de Bartlett o Parzen), y M es el parámetro de ancho de banda o truncamiento que controla el número de autocovarianzas incluidas en la estimación.

Para el caso específico donde se conoce que el diferencial de pérdida sigue un proceso MA($h - 1$), Diebold y Mariano sugieren simplificar el estimador utilizando $M = h - 1$ con kernel rectangular:

$$\hat{\sigma}_{DM}^2 = \hat{\gamma}_0 + 2 \sum_{k=1}^{h-1} \hat{\gamma}_k \quad (2-16)$$

El estadístico de prueba resultante es:

$$DM = \frac{\sqrt{n} \bar{d}}{\hat{\sigma}} \quad (2-17)$$

Bajo la hipótesis nula, este estadístico converge en distribución a una normal estándar: $DM \xrightarrow{d} N(0, 1)$. Para un test bilateral al nivel de significancia α , se rechaza H_0 si $|DM| >$

$z_{\alpha/2}$, donde $z_{\alpha/2}$ denota el cuantil $(1 - \alpha/2)$ de la distribución normal estándar.

2.3.3. Modificaciones para Muestras Pequeñas

A pesar de la solidez teórica del test de Diebold-Mariano bajo la teoría asintótica estándar, diversos estudios han documentado distorsiones de tamaño en muestras finitas, particularmente cuando el número de observaciones de pronóstico es limitado. Harvey et al. (Harvey, Leybourne, and Newbold 1997) demostraron mediante simulaciones Monte Carlo que el test original tiende a rechazar la hipótesis nula (es decir, presenta un tamaño empírico superior al nominal), especialmente para horizontes de pronóstico largos y muestras pequeñas.

Para abordar estas limitaciones, Harvey et al. proponen una corrección del estadístico que mejora sustancialmente el desempeño en muestras finitas. La modificación se fundamenta en el uso de un estimador aproximadamente insesgado de la varianza de \bar{d} . Partiendo de la expresión exacta:

$$\text{Var}(\bar{d}) = n^{-1} \left[\gamma_0 + 2n^{-1} \sum_{k=1}^{h-1} (n-k)\gamma_k \right] \quad (2-18)$$

y calculando el valor esperado del estimador empleado en (2-16), se obtiene que:

$$\mathbb{E}[\hat{\sigma}_{DM}^2] \approx \left[\frac{n+1-2h+n^{-1}h(h-1)}{n} \right] \text{Var}(\bar{d}) \quad (2-19)$$

Esta relación sugiere el estadístico modificado:

$$DM^* = \left[\frac{n+1-2h+n^{-1}h(h-1)}{n} \right]^{1/2} DM \quad (2-20)$$

Adicionalmente, Harvey et al. recomiendan comparar DM^* con valores críticos de la distribución t de Student con $(n-1)$ grados de libertad, en lugar de la distribución normal estándar. Esta segunda modificación reconoce implícitamente la incertidumbre adicional asociada con la estimación de la varianza en muestras finitas.

Los resultados de simulación reportados por Harvey et al. (Harvey, Leybourne, and Newbold 1997) indican que el test modificado presenta un tamaño empírico considerablemente más cercano al nominal, especialmente para $n \leq 50$ y horizontes de pronóstico $h \geq 2$. Aunque el test modificado exhibe una ligera pérdida de potencia en comparación con el

test original cuando ambos están correctamente calibrados, esta reducción es marginal y ampliamente compensada por la ganancia en confiabilidad inferencial.

2.3.4. Enfoque de la Teoría Asintótica de Suavizado Fijo

Una alternativa más reciente para abordar las distorsiones de tamaño del test de Diebold-Mariano en muestras pequeñas es el enfoque de *asintótica de suavizado fijo* (fixed-smoothing asymptotics), desarrollado por Coroneo e Iacone (Coroneo and Iacone 2020). Este marco teórico reconoce que en aplicaciones prácticas de evaluación de pronósticos, el tamaño muestral n es frecuentemente limitado, haciendo que la aproximación asintótica estándar (que requiere $M/n \rightarrow 0$) sea inadecuada.

La idea fundamental es mantener constante la razón entre el parámetro de ancho de banda y el tamaño muestral conforme n aumenta. Formalmente, bajo la *asintótica fixed-b*, se asume que $M/n \rightarrow b$ para algún $b \in (0, 1]$ fijo. Bajo este régimen asintótico alternativo, el estimador de varianza (2-15) ya no es consistente para σ^2 . Sin embargo, Kiefer y Vogelsang (2005) (Kiefer and Vogelsang 2005) demostraron que el estadístico resultante converge a una distribución no estándar que depende de b y del kernel empleado.

Para el kernel de Bartlett, la distribución límite puede caracterizarse explícitamente, y sus cuantiles pueden aproximarse mediante fórmulas polinomiales. Específicamente, para un test bilateral al 5% de significancia, el valor crítico $c_\alpha(b)$ satisface:

$$c_\alpha(b) \approx \alpha_0 + \alpha_1 b + \alpha_2 b^2 + \alpha_3 b^3 \quad (2-21)$$

donde los coeficientes $\{\alpha_i\}$ han sido tabulados por Kiefer y Vogelsang.

Coroneo e Iacone (Coroneo and Iacone 2020) extienden este marco al contexto específico de evaluación de pronósticos, demostrando mediante simulaciones Monte Carlo que los tests basados en asintótica de suavizado fijo exhiben un tamaño empírico notablemente más preciso que el test de Diebold-Mariano estándar, incluso para muestras tan pequeñas como $n = 40$. Los autores proponen utilizar anchos de banda $M = \lfloor n^{1/2} \rfloor$ para el estimador con kernel de Bartlett, encontrando que esta elección ofrece un equilibrio favorable entre tamaño y potencia del test.

Una segunda variante dentro del paradigma de suavizado fijo es la *asintótica fixed-m*, que emplea un estimador de varianza basado en el periodograma suavizado con kernel de

Daniell:

$$\hat{\sigma}_{DAN}^2 = \frac{2\pi}{m} \sum_{j=1}^m I(\lambda_j) \quad (2-22)$$

donde $I(\lambda_j)$ denota el periodograma de d_t evaluado en la frecuencia de Fourier $\lambda_j = 2\pi j/n$, y m es un parámetro de truncamiento mantenido fijo conforme $n \rightarrow \infty$. Bajo condiciones de regularidad, el estadístico resultante converge a una distribución t con $2m$ grados de libertad (Coroneo and Iacone 2020).

2.4. Predicción Conformal por Intervalos: El Enfoque IID

La predicción conformal clásica, introducida por Vovk et al. Vovk, Gammerman, and Shafer 2005, se fundamenta en la capacidad de generar conjuntos de predicción Γ^ϵ que garantizan una cobertura de confianza exacta para cualquier nivel de significancia $\epsilon \in (0, 1)$. A diferencia de los métodos estadísticos tradicionales que dependen de la teoría asintótica (grandes muestras), la predicción conformal es válida para muestras finitas, siempre que se cumpla el supuesto de intercambiabilidad de los datos.

2.4.1. El Concepto de No-Conformidad

El núcleo de esta metodología es la *medida de no-conformidad* (NCM, por sus siglas en inglés). Una NCM es una función $A(B, z)$ que cuantifica el grado de “extrañeza” de un ejemplo z en relación con un multiconjunto (o *bag*) de ejemplos B . En el contexto de regresión, donde $z = (x, y)$, la medida de no-conformidad más común es el error absoluto de predicción, definido como:

$$\alpha_i = |y_i - \hat{y}_i| \quad (2-23)$$

donde \hat{y}_i es la estimación producida por un algoritmo de aprendizaje subyacente (denominado *underlying algorithm*). Es importante subrayar que la predicción conformal es agnóstica al modelo: puede envolver desde una regresión lineal simple hasta redes neuronales profundas, transformando sus predicciones puntuales en intervalos con validez estadística.

2.4.2. Protocolo de Construcción de Intervalos

Para construir un intervalo de predicción para un nuevo objeto x_n , se define $z_i = (x_i, y_i)$ que permite escribir el un conjunto de entrenamiento como z_1, \dots, z_{n-1} , sobre el cual se realiza un proceso de prueba de hipótesis inversa. Para cada valor potencial $y \in \mathbb{R}$:

1. **Conjunto aumentado:** Se asume hipotéticamente que la verdadera etiqueta de x_n es y , formando el conjunto aumentado $z_1, z_2, \dots, z_{n-1}, z_n$, donde $z_n = (x_n, y)$.
2. **Cálculo de Puntajes:** Se calculan los puntajes de no-conformidad $\alpha_1, \dots, \alpha_n$ para todos los elementos, incluyendo el ejemplo hipotético.
3. **Derivación del p-valor:** Se calcula la proporción de ejemplos que son “al menos tan extraños” como el nuevo ejemplo z_n :

$$p(y) = \frac{|\{i = 1, \dots, n : \alpha_i \geq \alpha_n\}|}{n} \quad (2-24)$$

4. **Inversión de la Región de Aceptación:** El intervalo de predicción Γ^ϵ se define como el conjunto de todos los valores y que no pueden ser rechazados al nivel de significancia ϵ :

$$\Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) = \{y \in \mathbb{R} : p(y) > \epsilon\} \quad (2-25)$$

Este procedimiento garantiza que $P(y_n \notin \Gamma^\epsilon) \leq \epsilon$. Si los puntajes α_i tienen una distribución continua (sin empates), la probabilidad de error es exactamente ϵ (Vovk, Gammerman, and Shafer 2005).

La validez de cobertura mencionadas dependen fundamentalmente del supuesto de *intercambiabilidad*. Esta propiedad es la que permite afirmar que, bajo la hipótesis nula de que y es la etiqueta verdadera, el par (x_n, y) es estadísticamente indistinguible de los datos de entrenamiento z_1, \dots, z_{n-1} . En términos prácticos, la intercambiabilidad asegura que el puntaje de no-conformidad del nuevo ejemplo, α_n , tiene la misma probabilidad de ocupar cualquier posición en el ranking respecto a los puntajes previos $\alpha_1, \dots, \alpha_{n-1}$. Es precisamente esta distribución uniforme de los rangos de los puntajes lo que transforma el cálculo del p-valor en la ecuación (2-24) en una prueba de hipótesis válida, permitiendo que la probabilidad de error esté acotada por ϵ de manera exacta, independientemente de la distribución subyacente de los datos (Vovk, Gammerman, and Shafer 2022).

2.5. Robustez ante la No-Intercambiabilidad: Aproximación de Barber

Uno de los desafíos críticos en el análisis de series temporales es que el supuesto de intercambiabilidad rara vez se sostiene. Fenómenos como la autocorrelación, la heterocedastidad y la deriva de parámetros (drift) invalidan la asunción de que el pasado y el futuro son estadísticamente idénticos. Barber et al. proponen una extensión fundamental para estos escenarios.

2.5.1. La brecha (gap) de Cobertura y Variación Total

Barber et al. formalizan la degradación de la validez conformal mediante el uso de la *Distancia de Variación Total* (d_{TV}). Si la distribución de los datos cambia en el tiempo, existe una brecha de cobertura (*coverage gap*). El teorema principal de Barber establece que la pérdida de cobertura está acotada por la suma ponderada de las distancias entre la distribución de los datos de entrenamiento y la distribución del dato de prueba:

$$\text{Coverage gap} \leq \frac{\sum_{i=1}^n w_i \cdot d_{TV}(Z, Z^i)}{1 + \sum_{i=1}^n w_i} \quad (2-26)$$

donde:

- $Z = (Z_1, \dots, Z_{n+1})$ representa la secuencia completa de datos (entrenamiento y prueba)
- Z^i denota la secuencia después de intercambiar el punto de prueba (X_{n+1}, Y_{n+1}) con el i -ésimo punto de entrenamiento (X_i, Y_i)
- $w_i \in [0, 1]$ son pesos pre-especificados que reflejan la confianza en cada punto de datos
- $d_{TV}(Z, Z^i)$ es la distancia de variación total entre las distribuciones de Z y Z^i
- ϵ representa el nivel de error nominal (típicamente α en predicción conformal estándar)

2.5.2. Cuantiles Ponderados y Decaimiento Temporal

Para contrarrestar este efecto en series de tiempo, Barber et al. introducen los *Weighted Conformal Predictors*. En lugar de asignar un peso uniforme de $1/n$ a cada residuo histórico, se asignan pesos w_i que reflejan la relevancia del dato. En series no estacionarias, los datos más recientes son mejores predictores del futuro.

Se define comúnmente un decaimiento geométrico para los pesos:

$$w_i = \rho^{n-i}, \quad \rho \in (0, 1) \quad (2-27)$$

donde un ρ cercano a 1 asume una estabilidad lenta, mientras que un ρ menor reacciona rápidamente a cambios estructurales. El p-valor pesado se calcula como una suma ponderada de funciones indicadoras:

$$p^y = \frac{\sum_{i=1}^{n-1} w_i \mathbb{1}_{\alpha_i \geq \alpha_n} + w_n}{\sum_{j=1}^n w_j} \quad (2-28)$$

Este enfoque permite que la predicción conformal sea “adaptativa”, manteniendo la cobertura cercana al nivel nominal incluso cuando la serie temporal experimenta cambios súbitos en su media o varianza (Barber et al. 2023).

2.6. Sistemas de Predicción Conformal (CPS): De Intervalos a Densidades

El Capítulo 7 de la obra de Vovk, Gammerman, and Shafer 2005 marca la transición de la predicción de intervalos a la predicción de distribuciones completas. Un *Sistema de Predicción Conformal* (CPS) no entrega un rango, sino una *Distribución Predictiva Aleatorizada* (RPD), denotada como $\Pi_n(y, \tau)$, que representa la función de distribución acumulada de la variable de interés.

2.6.1. Formalización de la RPD y el Suavizado (τ)

Para asegurar que la distribución resultante sea continua y cumpla con las propiedades de una FDA (Función de Distribución Acumulada), se introduce una variable de suavizado

$\tau \sim U(0, 1)$. La función Π se define como:

$$\Pi_n(y, \tau) := \frac{|\{i : \alpha_i < \alpha_n^y\}| + \tau |\{i : \alpha_i = \alpha_n^y\}|}{n} \quad (2-29)$$

Es fundamental precisar que la validez de la RPD radica en que, bajo el supuesto de intercambiabilidad, la función Π_n evaluada en la variable aleatoria subyacente genera una distribución uniforme. Específicamente, al evaluar la RPD en las etiquetas verdaderas y_n junto a sus respectivas variables de suavizado τ_n , los valores resultantes $P_n = \Pi_n(y_n, \tau_n)$ constituyen una muestra aleatoria de la distribución $U(0, 1)$, propiedad conocida como *calibración fuerte en probabilidad* Vovk, Gammerman, and Shafer 2022. En este proceso, el rol de $\tau \sim U(0, 1)$ es resolver la falta de continuidad producida por los posibles empates en los puntajes de conformidad ($\alpha_i = \alpha_n^y$), permitiendo que una función de saltos se transforme en una medida aleatorizada continua y asegurando así una cuantificación exacta de la incertidumbre.

2.6.2. La Máquina de Predicción de Mínimos Cuadrados (LSPM)

LSPM (*Least Squares Prediction Machine*) es la aplicación primordial de los CPS al ámbito de la regresión. La LSPM utiliza la estructura de la regresión lineal para optimizar la eficiencia de la distribución predictiva, aprovechando las propiedades geométricas del método de mínimos cuadrados.

Consideramos el problema de regresión con p atributos. El espacio de objetos es $\mathbf{X} := \mathbb{R}^p$ y el espacio de ejemplos es $\mathbf{Z} := \mathbb{R}^{p+1} = \mathbb{R}^p \times \mathbb{R}$. La matriz de datos \bar{X} es de dimensión $n \times p$, y la matriz sombrero (hat) se define como:

$$\bar{H} = \bar{X}(\bar{X}^\top \bar{X})^{-1} \bar{X}^\top \quad (2-30)$$

Los elementos de esta matriz se denotan como $\bar{h}_{i,j}$, siendo $\bar{h}_{i,i}$ el apalancamiento (*leverage*) del i -ésimo punto de datos.

Variantes de la LSPM

Vovk distingue tres formas de calcular las medidas de no conformidad dentro de una LSPM, cada una con diferentes propiedades teóricas y prácticas:

1. **LSPM Ordinaria:** Define la medida de no conformidad como el residuo ordinario:

$$A([z_1, \dots, z_n], z_n) := y_n - \hat{y}_n \quad (2-31)$$

donde \hat{y}_n es la predicción para y_n usando regresión de mínimos cuadrados sobre los objetos x_n con z_1, \dots, z_{n-1} como conjunto de entrenamiento. Este enfoque tiende a ser demasiado optimista (sobreajuste), ya que el modelo ya ha “visto” los datos de entrenamiento.

2. **LSPM Eliminada (*Deleted*):** Utiliza un esquema de validación cruzada interna (*leave-one-out*):

$$A([z_1, \dots, z_n], z_n) := y_n - \hat{y}_{(n)} \quad (2-32)$$

donde $\hat{y}_{(n)}$ es la predicción para y_n usando mínimos cuadrados sobre x_n con z_1, \dots, z_{n-1} como conjunto de entrenamiento (excluyendo z_n del test). Esto asegura que el residuo sea una medida honesta de la capacidad de generalización.

3. **LSPM Estudiantizada:** Es la variante más robusta y matemáticamente rigurosa. Ajusta cada residuo por su apalancamiento (*leverage*), \bar{h}_n :

$$A([z_1, \dots, z_n], z_n) := \frac{y_n - \hat{y}_n}{\sqrt{1 - \bar{h}_n}} \quad (2-33)$$

Forma Explícita de la Distribución Predictiva

Para la LSPM estudiantizada, la distribución predictiva toma una forma particularmente elegante. Dado un conjunto de entrenamiento y un objeto de prueba x_n , los residuos estudiantizados $\alpha_i^y = \phi_i^y$ se calculan como:

$$\alpha_i^y - \phi_i^y = B_i y - A_i, \quad i = 1, \dots, n-1 \quad (2-34)$$

donde, en la notación de Vovk, y es la etiqueta del n -ésimo objeto x_n y:

$$B_i := \sqrt{1 - \bar{h}_n + \frac{\bar{h}_{i,n}}{\sqrt{1 - \bar{h}_i}}} \quad (2-35)$$

$$A_i := \frac{\sum_{j=1}^{n-1} \bar{h}_{j,n} x_j}{\sqrt{1 - \bar{h}_n}} + \frac{y_i - \sum_{j=1}^{n-1} \bar{h}_{i,j} y_j}{\sqrt{1 - \bar{h}_i}} \quad (2-36)$$

Ordenando la secuencia de conformidad C_1, \dots, C_{n-1} en orden ascendente como $C_{(1)} \leq \dots \leq C_{(n-1)}$, y definiendo $C_{(0)} := -\infty$ y $C_{(n)} := \infty$, la distribución predictiva se expresa como:

$$\bar{\Pi}_n(y) := \begin{cases} \left[\frac{i'}{n}, \frac{i'+1}{n} \right] & \text{si } y \in (C_{(i)}, C_{(i+1)}) \text{ para } i \in \{0, 1, \dots, n-1\} \\ \left[\frac{i'}{n}, \frac{i''}{n} \right] & \text{si } y = C_{(i)} \text{ para } i \in \{1, \dots, n-1\} \end{cases} \quad (2-37)$$

donde $i' := \min\{j : C_{(j)} = C_{(i)}\}$ y $i'' := \max\{j : C_{(j)} = C_{(i)}\}$.

Esta representación explícita revela que la distribución predictiva de la LSPM estudiantizada es una función escalonada con a lo sumo $n - 1$ puntos de discontinuidad, cada uno correspondiente a un valor único de los puntajes de conformidad ordenados. El grosor de cada intervalo es igual a $1/n$, excepto en los puntos de empate donde varios puntajes de conformidad coinciden.

2.7. Sistemas de Predicción Conformal de Mondrian (MCPS)

A pesar de las sólidas garantías de validez marginal que ofrecen los Sistemas de Predicción Conformal (CPS) descritos en la sección 2.6, estos presentan una limitación teórica y práctica fundamental: la garantía de error es un promedio sobre todo el espacio de datos. Esto implica que el sistema puede ser extremadamente preciso en ciertas regiones del espacio de características y, simultáneamente, cometer errores sistemáticos en otras, siempre que el error global no supere el nivel ϵ . Los *Sistemas de Predicción Conformal de Mondrian* (MCPS, por sus siglas en inglés) introducen el concepto de *validez condicional por categorías*, permitiendo que la calibración se mantenga exacta dentro de subconjuntos específicos de los datos (Vovk, Gammerman, and Shafer 2022).

2.7.1. Origen y Motivación: Validez Marginal vs. Condicional

El apelativo “Mondrian” deriva del estilo geométrico del pintor neerlandés Piet Mondrian, cuya estética se fundamenta en la compartmentación del lienzo en rectángulos de colores puros delimitados por una cuadrícula, tal como se ilustra en la Figura ???. Bajo esta analogía, un sistema de predicción conformal Mondriano partitiona el espacio de ejemplos \mathcal{Z} en categorías mutuamente excluyentes o taxonomías. Este enfoque permite que las garantías de cobertura sean válidas no solo de forma agregada, sino específicamente dentro de cada subgrupo definido, abordando así el problema de la validez condicional.

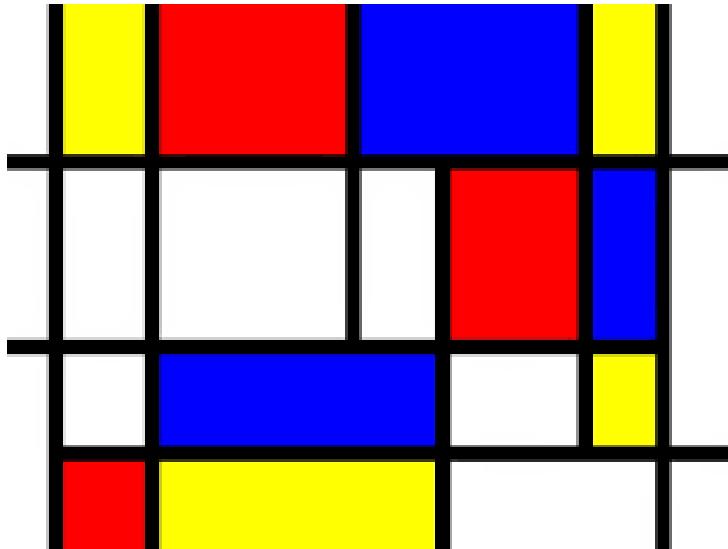


Figura 2-2.: Piet Mondrian (1872-1944). Composición con rojo, amarillo y azul. Dominio público.

La necesidad de este enfoque surge cuando existen grupos de datos con dificultades predictivas heterogéneas. Por ejemplo, en una serie temporal de demanda eléctrica, predecir el consumo en un día festivo es intrínsecamente más difícil que en un día laboral. Un CPS global podría subestimar masivamente la incertidumbre en los días festivos, compensándola con una sobreestimación en los días laborales. El enfoque de Mondrian garantiza que la probabilidad de error sea exactamente ϵ tanto para los días laborales como para los festivos, de forma independiente (Vovk, Nouretdinov, et al. 2017).

2.7.2. La Taxonomía de Mondrian (κ)

La base matemática de un MCPS es la *taxonomía*. Una taxonomía es una función medible $\kappa : \mathbb{N} \times (\mathbf{X} \times \mathbf{Y}) \rightarrow K$, donde K es un conjunto numerable de categorías. Para cada par de ejemplo (x_i, y_i) y su posición en la secuencia i , la taxonomía asigna una categoría κ_i .

Existen tres tipos principales de taxonomías aplicables a series temporales:

1. **Taxonomías de Objetos:** dependen solo de las características x_i (ej. agrupar por niveles de volatilidad observada).
2. **Taxonomías de Etiquetas:** dependen de la respuesta y_i . Esto da lugar a los *Label-Conditional Conformal Predictors*, vitales cuando el impacto de un error depende de la magnitud del valor (ej. errores en valores extremos son más costosos).
3. **Taxonomías Temporales:** dependen del índice i . Este es el puente con el trabajo de Barber et al. Barber et al. 2023, donde la categoría de Mondrian puede ser una “ventana deslizante” de los datos más recientes para adaptarse a la no-intercambiabilidad.

2.7.3. Integración del Algoritmo MCPS

La integración de la lógica de Mondrian en un Sistema de Predicción Conformal se realiza modificando el cálculo del p-valor o de la RPD (Distribución Predictiva Aleatorizada). En lugar de comparar el puntaje del nuevo ejemplo α_n con todos los puntajes históricos, solo se compara con aquellos que pertenecen a su misma categoría.

Sea $\sigma = \{z_1, \dots, z_{n-1}\}$ el conjunto de entrenamiento y $z_n = (x_n, y)$ el ejemplo de prueba con etiqueta hipotética y . El proceso para generar la RPD de Mondrian Π_M es el siguiente:

1. Se identifica la categoría del nuevo ejemplo: $k = \kappa(n, (x_n, y))$.
2. Se filtran los índices de los ejemplos de entrenamiento que pertenecen a dicha categoría:

$$S_k = \{i \in \{1, \dots, n-1\} : \kappa(i, z_i) = k\} \quad (2-38)$$

3. Se calculan los puntajes de conformidad α_i solo para $i \in S_k \cup \{n\}$.

4. La RPD de Mondrian se define como:

$$\Pi_M(y, \tau) := \frac{|\{i \in S_k : \alpha_i < \alpha_n^y\}| + \tau |\{i \in S_k \cup \{n\} : \alpha_i = \alpha_n^y\}|}{|S_k| + 1} \quad (2-39)$$

El denominador $|S_k| + 1$ es clave: representa el tamaño de la “muestra local”. Si una categoría tiene pocos ejemplos, la distribución predictiva será naturalmente más dispersa (reflejando mayor incertidumbre), mientras que categorías ricas en datos producirán densidades más nítidas Vovk, Gammerman, and Shafer 2022.

2.8. Vacío Metodológico: De Intervalos a Distribuciones en Series Temporales

A pesar de los avances en predicción conformal y sus extensiones para dependencia temporal, existe un vacío metodológico fundamental: la mayoría de adaptaciones de CP para series temporales se concentran exclusivamente en intervalos de predicción, dejando prácticamente inexplorada la generación de distribuciones predictivas completas en este contexto.

2.8.1. Estado Actual y Dimensiones del Vacío

La predicción conformal clásica (Vovk, Gammerman, and Shafer 2005) fue desarrollada para el contexto i.i.d. Las extensiones para series temporales, como el enfoque de ponderación de Barber et al. (Barber et al. 2023), se enfocan en mantener garantías de cobertura para intervalos mediante esquemas adaptativos. Los Sistemas de Predicción Conformal (CPS) de la Sección 2.6 y las taxonomías de Mondrian (Sección 2.7) representan el marco teórico más completo para distribuciones conformales, pero su aplicación se restringe casi exclusivamente a datos intercambiables.

Actualmente no existe una metodología consolidada que combine: (i) generación de distribuciones predictivas completas, (ii) adaptación a dependencia temporal, y (iii) garantías teóricas sobre calibración. Los métodos existentes presentan compensaciones: el *Sieve Bootstrapping* (Lahiri 2003) ofrece flexibilidad pero carece de garantías formales de calibración; *DeepAR* (Salinas et al. 2020) muestra excelente desempeño empírico pero opera

como caja negra sin fundamento teórico sobre cobertura.

El vacío identificado se manifiesta en tres dimensiones críticas. Primero, los intervalos de predicción proporcionan información limitada: en problemas de decisión bajo incertidumbre como optimización de inventarios o gestión de riesgos financieros, conocer únicamente los límites de un intervalo resulta insuficiente. La distribución predictiva completa permite evaluar probabilidades de eventos específicos, calcular valores esperados de funciones de pérdida asimétricas, y cuantificar riesgos extremos. Segundo, existe una desconexión entre fundamento teórico y práctica: mientras la teoría conformal proporciona garantías exactas bajo intercambiabilidad, las adaptaciones para dependencia temporal carecen de un marco unificado que caracterice rigurosamente el comportamiento bajo correlación temporal. Tercero, hay una carencia de evaluación comparativa sistemática: prácticamente no existen estudios que evalúen distribuciones conformales completas frente a distribuciones generadas por métodos establecidos utilizando métricas probabilísticas apropiadas.

2.8.2. Contribución y Relevancia

Esta tesis aborda directamente el vacío metodológico mediante tres contribuciones principales. Primero, desarrolla una extensión sistemática de los Sistemas de Predicción Conformal desde el contexto i.i.d. hacia series temporales, incorporando esquemas de ponderación temporal que manejan la no-intercambiabilidad mientras generan distribuciones predictivas completas. Segundo, establece un marco de evaluación comparativa riguroso basado en métricas probabilísticas como el *Continuous Ranked Probability Score* (CRPS) y pruebas de significancia estadística, permitiendo comparar distribuciones conformales con métodos establecidos bajo criterios objetivos. Tercero, proporciona evidencia empírica mediante simulaciones controladas y aplicaciones a datos reales sobre el desempeño relativo de CP frente a alternativas como *Sieve Bootstrapping* y *DeepAR*.

La relevancia práctica de cerrar esta brecha es significativa. En pronóstico de demanda, la distribución completa permite optimizar niveles de stock considerando costos asimétricos. En gestión de riesgos energéticos, conocer la distribución predictiva de generación renovable facilita estrategias de cobertura. En meteorología operacional, distribuciones calibradas son fundamentales para sistemas de alerta temprana. Al ofrecer métodos conformales que generan distribuciones completas con propiedades de calibración controlables, este trabajo proporciona a los practicantes una alternativa fundamentada que combina garantías teóricas con aplicabilidad en problemas reales de pronóstico temporal.

2.9. Análisis de la Consistencia Universal de Vovk

Para consolidar el marco teórico de esta investigación, es imperativo discutir el sustento matemático que garantiza que los Sistemas de Predicción Conformal (CPS) no solo son válidos en muestras finitas, sino también óptimos a medida que el volumen de datos aumenta. Este respaldo proviene de la demostración de la *consistencia universal* de Vovk (Vovk 2019), formalizada en el Teorema 31 de su obra reciente.

2.9.1. Definición de Consistencia Universal

2.9.2. Definición de Consistencia Universal

Antes de proceder con la definición formal, es crucial aclarar la distinción entre dos propiedades fundamentales de los CPS: *validez* y *consistencia*. En el marco de predicción conformal, el término **validez** se refiere específicamente a la propiedad R2 establecida por Vovk, Gammerman, and Shafer 2005, que garantiza calibración probabilística automática en muestras finitas. Formalmente, un sistema predictivo aleatorizado cumple la propiedad de validez si, para cualquier nivel $\alpha \in [0, 1]$ y cualquier medida de probabilidad P , la distribución de Q evaluada en datos aleatorios $(z_1, \dots, z_{n+1}) \sim P$ y $\tau \sim U[0, 1]$ es uniforme:

$$\forall \alpha \in [0, 1] : P(Q(z_1, \dots, z_n, z_{n+1}, \tau) \leq \alpha) = \alpha \quad (2-40)$$

Esta propiedad es más fuerte que la simple calibración marginal, pues no solo garantiza que $\mathbb{E}[Q(\cdot, y)] = P(y_{n+1} \leq y)$, sino que asegura que los p-valores conformales son uniformes e independientes para diferentes n , permitiendo aplicar la ley de los grandes números (Vovk 2019). A lo largo de este trabajo, usaremos “validez” exclusivamente en este sentido técnico.

Sin embargo, la validez por sí sola no garantiza que la distribución predictiva Π_n sea una buena aproximación de la verdadera distribución condicional $P(y|x)$. Aquí es donde entra el concepto de *consistencia universal*. Vovk define un sistema predictivo como *universalmente consistente* si, para cualquier medida de probabilidad P bajo el modelo IID y para cualquier función continua acotada f , se cumple que:

$$\int f d\Pi_n - \mathbb{E}_P(f | x_{n+1}) \rightarrow 0 \quad \text{en probabilidad cuando } n \rightarrow \infty \quad (2-41)$$

Esta propiedad implica que, asintóticamente, el CPS aproxima tan precisamente como se quiera la verdadera distribución generadora de los datos, eliminando la incertidumbre epistémica conforme el tamaño de la muestra n tiende al infinito (Vovk 2019).

2.9.3. Mecanismo de la Demostración: El Enfoque de Histograma

La prueba de Vovk sobre la existencia de un CPS consistente universalmente se apoya en la construcción de un *Histogram Conformal Predictive System*. El argumento se divide en dos pilares fundamentales que vinculan la teoría de martingalas con la ley de los grandes números:

1. **Teorema de Convergencia de Martingalas de Lévy:** Vovk utiliza particiones anidadas del espacio de objetos X (celdas de histograma que se encogen conforme n crece). Según el teorema de Lévy, la esperanza condicional de la función sobre una celda que se reduce tiende al valor puntual de la esperanza condicional en el objeto de prueba x_{n+1} (Vovk 2019).
2. **Ley de los Grandes Números (LGN):** Mientras las celdas se encogen para ganar resolución, el número de ejemplos dentro de cada celda debe tender a infinito ($nh_n \rightarrow \infty$, donde h_n es el ancho de la celda). Esto permite que la frecuencia empírica de las etiquetas dentro de la categoría de Mondrian converja a la esperanza real en esa región del espacio (Vovk 2019).

2.9.4. La Distancia de Lévy y la Convergencia Débil

Un punto crítico de la demostración es el uso de la noción de Belyaev sobre secuencias de distribuciones que se aproximan débilmente. Vovk demuestra que bajo un CPS universal, la *Distancia de Lévy* entre la distribución predictiva conformal y la verdadera distribución condicional converge a cero en probabilidad Vovk 2019.

Este resultado es el que otorga rigor a la aplicación de CPS en problemas de alta criticidad, como el pronóstico de carga eléctrica o la gestión de riesgos financieros. Indica que el analista no tiene que elegir entre un modelo “seguro” (conformal) y un modelo “preciso” (bayesiando/paramétrico); el CPS universal ofrece ambas ventajas simultáneamente:

- **A corto plazo:** Garantiza cobertura exacta mediante calibración fuerte.

- **A largo plazo:** Garantiza convergencia a la distribución real de los datos sin requerir asunciones paramétricas.

2.9.5. Implicaciones para el LSPM y la Eficiencia

Aunque el modelo de mínimos cuadrados (LSPM) estudiado en la sección 2.6 es eficiente bajo ruido gaussiano, Vovk advierte que no es universalmente consistente si la relación real entre X y Y no es lineal (Vovk, Gammerman, and Shafer 2005). Por ello, el desarrollo de CPS basados en kernels o en métodos de vecinos cercanos (como se discute en el capítulo 4 de su obra) es lo que permite alcanzar la consistencia universal en espacios de características complejos. Esta conclusión justifica el uso de arquitecturas no lineales conformizadas en la presente tesis, ya que heredan la solidez de la prueba de consistencia de Vovk.

2.10. Hacia la Consistencia Universal en Series de Tiempo Ergódicas

Si bien el trabajo de Vovk (Vovk 2019) establece una base sólida para la consistencia universal bajo el modelo IID, las aplicaciones en entornos reales, como las series de tiempo, exigen una transición hacia modelos que capturen la dependencia temporal. Durante el desarrollo de esta investigación, se intentó abordar el problema teórico de probar las condiciones formales bajo las cuales un Sistema de Predicción Conformal (CPS) podría alcanzar consistencia universal en el contexto de procesos estocásticos dependientes, donde la suposición de intercambiabilidad no se cumple.

Sin embargo, por cuestiones de alcance y limitaciones temporales, este esfuerzo resultó en la identificación de una ruta teórica que plantea los supuestos fundamentales, la arquitectura propuesta y las direcciones necesarias para una demostración futura. Los detalles completos de este marco conceptual —incluyendo la redefinición del objetivo de consistencia, los supuestos de estacionariedad y α -mixing, y el mecanismo propuesto de transductor conformal por kernel— se presentan en el Apéndice A como una guía para investigaciones posteriores que busquen extender los resultados de Vovk al dominio de series de tiempo ergódicas.

A pesar de no contar con una prueba formal completa, la identificación de estas condiciones ideales (ergodicidad, mezcla fuerte con decaimiento algebraico, y regularidad de Lipschitz) constituye un aporte conceptual que orienta el diseño empírico de los métodos implementados en los capítulos subsecuentes de esta tesis, particularmente en la selección de ventanas temporales móviles y esquemas de ponderación espacial por kernel.

3. Metodología de Simulaciones

Este capítulo describe el diseño experimental desarrollado para evaluar el desempeño de los métodos de pronóstico probabilístico en series temporales. Se presenta la justificación de los escenarios de evaluación, la metodología de simulación empleada y las características específicas de los procesos generadores de datos utilizados.

3.1. Introducción

La evaluación rigurosa de metodologías de pronóstico probabilístico requiere un marco experimental controlado que permita comparar el desempeño de diferentes técnicas bajo condiciones conocidas. A diferencia de los estudios con datos reales, donde la distribución verdadera es desconocida y la evaluación se limita a métricas indirectas, los estudios de simulación ofrecen la ventaja fundamental de conocer exactamente el proceso generador de datos (DGP, por sus siglas en inglés) (Rob J Hyndman and Athanasopoulos [2021](#)).

Este conocimiento del DGP permite evaluar directamente la calidad de las distribuciones predictivas mediante su comparación con la verdadera distribución teórica. En particular, el uso del ECRPS (Expected Continuous Ranked Probability Score) como métrica principal de evaluación se justifica porque permite cuantificar simultáneamente la calibración y la nitidez de los pronósticos probabilísticos, comparando las muestras generadas por cada método con muestras de la distribución teórica verdadera (Gneiting and Katzfuss [2014](#)).

El diseño experimental desarrollado considera tres dimensiones fundamentales de variación: (1) la estructura temporal del proceso (estacionariedad y linealidad), (2) la distribución del término de error, y (3) la magnitud de la varianza del ruido. Esta combinación genera un espacio de escenarios suficientemente amplio para evaluar la robustez y adaptabilidad de los métodos bajo diferentes condiciones operativas.

3.2. Diseño de la Simulación

3.2.1. Selección de Escenarios de Evaluación

El presente estudio considera tres escenarios fundamentales que caracterizan diferentes clases de comportamiento en series temporales. La selección de estos escenarios se fundamenta en la clasificación teórica de procesos estocásticos y en consideraciones de relevancia práctica.

Escenario 1: Lineal Estacionario (ARMA)

El primer escenario considera procesos autorregresivos de media móvil (ARMA), que representan la clase fundamental de modelos lineales estacionarios. Un proceso ARMA(p, q) se caracteriza por su capacidad de capturar tanto la persistencia temporal (componente AR) como la dependencia de shocks pasados (componente MA), manteniendo propiedades estadísticas constantes en el tiempo (Arrieta Prieto 2017).

La estacionariedad de estos procesos garantiza que la media, varianza y estructura de autocorrelación permanezcan invariantes bajo traslaciones temporales, lo que facilita la modelación y el pronóstico (Rob J Hyndman and Athanasopoulos 2021). Este escenario permite evaluar el desempeño de los métodos en condiciones ideales, donde los supuestos fundamentales de muchas técnicas estadísticas se cumplen.

Escenario 2: Lineal No Estacionario (ARIMA)

El segundo escenario aborda procesos autorregresivos integrados de media móvil (ARIMA), que extienden la clase ARMA para series con tendencias estocásticas. La presencia de raíces unitarias en el polinomio autorregresivo genera comportamientos de paseo aleatorio que son comunes en series económicas y financieras (Rob J Hyndman and Athanasopoulos 2021).

La no estacionariedad introduce desafíos adicionales para el pronóstico probabilístico, ya que la incertidumbre crece sin límite conforme aumenta el horizonte de predicción. Este escenario permite evaluar la capacidad de los métodos para adaptarse a estructuras no estacionarias mediante diferenciación o técnicas adaptativas.

Escenario 3: No Lineal Estacionario (SETAR)

El tercer escenario considera modelos autorregresivos de umbral auto-excitados (SETAR), que permiten cambios estructurales endógenos en la dinámica del proceso. Estos modelos capturan no linealidades mediante el cambio de régimen determinado por valores pasados de la propia serie (P. Chen and Semmler 2023).

La estacionariedad global de un proceso SETAR requiere condiciones específicas sobre los parámetros autorregresivos en cada régimen y la frecuencia de transición entre regímenes. Estas condiciones se discuten en detalle en la Sección 3.3.3. Este escenario es particularmente relevante para evaluar la capacidad de los métodos conformales de capturar dinámicas asimétricas y dependientes del estado del sistema.

3.2.2. Estructura del Diseño de Simulación base

El diseño experimental implementa un esquema completo que combina sistemáticamente tres dimensiones de variación para cada uno de los tres escenarios considerados. Esta estructura genera un total de 420 configuraciones únicas de simulación, distribuidas equitativamente entre los escenarios.

Dimensión 1: Configuraciones Paramétricas del Proceso

Para cada clase de modelo (ARMA, ARIMA, SETAR), se consideran 7 configuraciones paramétricas distintas que representan diferentes grados de complejidad y características dinámicas. Las especificaciones detalladas de estas configuraciones se presentan en la Sección 3.3. Esta diversidad paramétrica permite evaluar la sensibilidad de los métodos a diferentes estructuras de dependencia temporal.

Dimensión 2: Distribuciones del Término de Error

Se consideran cinco familias de distribuciones para el término de innovación ε_t , seleccionadas para representar diferentes características de forma, simetría y comportamiento en las colas:

1. **Normal:** $\varepsilon_t \sim N(0, \sigma^2)$. Representa el caso base con colas ligeras y simetría perfecta.

2. **T-Student:** $\varepsilon_t \sim \sigma \cdot \frac{t_{18}}{\sqrt{18/16}}$, donde t_{18} denota una distribución t de Student con 18 grados de libertad. Esta parametrización garantiza varianza unitaria y genera colas más pesadas que la normal, capturando eventos extremos más frecuentes.
3. **Exponencial:** $\varepsilon_t \sim \sigma(Y - 1)$, donde $Y \sim \text{Exp}(1)$. Produce asimetría positiva y es relevante para series que modelan variables intrínsecamente positivas o con shocks unidireccionales.
4. **Uniforme:** $\varepsilon_t \sim U(-\sqrt{3}\sigma, \sqrt{3}\sigma)$. Genera soporte acotado y ausencia de colas, representando un caso extremo de curtosis negativa.
5. **Mixtura de Normales:** $\varepsilon_t \sim 0.75 \cdot N(-\sigma/4, \sigma^2/16) + 0.25 \cdot N(3\sigma/4, \sigma^2/16)$. Produce bimodalidad y permite evaluar el desempeño bajo distribuciones predictivas complejas con múltiples modas.

Esta selección permite evaluar la robustez de los métodos ante desviaciones del supuesto de normalidad que frecuentemente se asume en la literatura de pronóstico (Arrieta Prieto 2017).

Dimensión 3: Niveles de Varianza del Error

Se consideran cuatro niveles de varianza $\sigma^2 \in \{0.2, 0.5, 1.0, 3.0\}$ que representan diferentes razones señal-ruido. El nivel base $\sigma^2 = 1.0$ corresponde a la parametrización estándar, mientras que $\sigma^2 = 0.2$ representa un escenario de alta predictibilidad y $\sigma^2 = 3.0$ captura situaciones de alta volatilidad donde la incertidumbre inherente domina la dinámica del sistema.

Combinatoria Total

La combinación de estas tres dimensiones genera:

$$N_{\text{config}} = 7 \text{ modelos} \times 5 \text{ distribuciones} \times 4 \text{ varianzas} = 140 \text{ configuraciones por escenario} \quad (3-1)$$

Con tres escenarios (ARMA, ARIMA, SETAR), el espacio experimental completo comprende:

$$N_{\text{total}} = 140 \times 3 = 420 \text{ configuraciones únicas} \quad (3-2)$$

Adicionalmente, considerando que cada configuración se evalúa en un horizonte de predicción de 12 pasos usando ventana rodante para que se realice predicción a un paso adelante, el número total de combinaciones configuración-horizonte es de $420 \times 12 = 5040$.

3.2.3. Protocolo de Simulación y Partición de Datos

Para cada una de las 420 configuraciones, se implementa el siguiente protocolo de simulación:

1. **Generación de la Serie:** Se simulan $n_{\text{total}} = 302$ observaciones del proceso especificado, precedidas por un período de burn-in de 50 observaciones que se descartan para eliminar el efecto de las condiciones iniciales. Esto resulta en una serie efectiva de longitud $n = 252$.
2. **Partición Tripartita:** La serie se divide en tres conjuntos disjuntos:
 - **Conjunto de Entrenamiento:** $n_{\text{train}} = 200$ observaciones iniciales utilizadas para la estimación inicial de parámetros.
 - **Conjunto de Calibración:** $n_{\text{cal}} = 40$ observaciones subsecuentes utilizadas para la calibración de los hiperparametros y la construcción de distribuciones conformales.
 - **Conjunto de Prueba:** $n_{\text{test}} = 12$ observaciones finales utilizadas para la evaluación del desempeño predictivo.
3. **Esquema de Ventana Rodante:** La evaluación se realiza mediante una ventana rodante (rolling window) donde:
 - Para el primer paso de predicción, se utilizan las primeras 200 observaciones para entrenamiento y las siguientes 40 para calibración.
 - Para cada paso $h = 1, \dots, 12$, la ventana de entrenamiento se extiende para incluir las observaciones anteriores, manteniendo fijo el conjunto de calibración de tamaño 40 inmediatamente anterior al punto de predicción.
 - Este esquema emula una situación operativa donde el analista actualiza periódicamente los modelos conforme nueva información se hace disponible.
4. **Generación de Distribuciones Predictivas:** Para cada método y cada paso de

predicción h , se generan muestras de la distribución predictiva. Estas muestras se comparan con muestras de la distribución teórica verdadera del proceso (conocida por construcción del DGP) mediante el cálculo del ECRPS para ese paso específico.

3.3. Procesos Generadores de Datos

Esta sección describe formalmente los modelos utilizados como procesos generadores de datos en cada escenario, junto con las configuraciones paramétricas específicas consideradas. Para cada clase de modelo, se presentan las ecuaciones fundamentales, las condiciones de estacionariedad (cuando corresponda) y las parametrizaciones concretas evaluadas.

3.3.1. Procesos ARMA: Escenario Lineal Estacionario

Definición y Representación

Un proceso autorregresivo de media móvil de órdenes p y q , denotado ARMA(p, q), se define mediante la ecuación en diferencias estocástica:

$$Y_t = c + \sum_{i=1}^p \phi_i Y_{t-i} + \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j} \quad (3-3)$$

donde c es un término constante, $\{\phi_i\}_{i=1}^p$ son los coeficientes autorregresivos, $\{\theta_j\}_{j=1}^q$ son los coeficientes de media móvil, y $\{\varepsilon_t\}$ es un proceso de ruido blanco con media cero y varianza σ^2 .

Utilizando el operador de rezagos L definido por $L^k Y_t = Y_{t-k}$, el proceso puede expresarse en forma compacta:

$$\Phi(L)Y_t = c + \Theta(L)\varepsilon_t \quad (3-4)$$

donde $\Phi(L) = 1 - \sum_{i=1}^p \phi_i L^i$ es el polinomio autorregresivo y $\Theta(L) = 1 + \sum_{j=1}^q \theta_j L^j$ es el polinomio de media móvil.

Condiciones de Estacionariedad e Invertibilidad

La estacionariedad y la invertibilidad de un proceso ARMA están determinadas por las raíces de sus polinomios característicos (Rob J Hyndman and Athanasopoulos 2021):

- **Estacionariedad:** El proceso es estacionario en covarianza si y solo si todas las raíces del polinomio autorregresivo $\Phi(z) = 0$ se encuentran estrictamente fuera del círculo unitario complejo. Equivalentemente, las raíces del polinomio $\Phi(L)$ deben satisfacer $|z_i| > 1$ para todo i .
- **Invertibilidad:** El proceso es invertible si y solo si todas las raíces del polinomio de media móvil $\Theta(z) = 0$ se encuentran estrictamente fuera del círculo unitario complejo.

Estas condiciones garantizan que el proceso admite representaciones de Wold ($MA(\infty)$) y autorregresiva ($AR(\infty)$) convergentes, lo que es fundamental para la teoría de pronóstico (Arrieta Prieto 2017).

Distribución Predictiva Verdadera

Para un proceso ARMA estacionario e invertible, la distribución del siguiente valor Y_{n+1} condicionada a la historia observada $\mathcal{F}_n = \{Y_1, \dots, Y_n, \varepsilon_1, \dots, \varepsilon_n\}$ tiene una forma analítica explícita. Dado que el modelo es lineal, la distribución condicional está completamente caracterizada por su media y varianza condicionales.

La media condicional se obtiene de la ecuación estructural del modelo:

$$\mathbb{E}[Y_{n+1} | \mathcal{F}_n] = c + \sum_{i=1}^p \phi_i Y_{n+1-i} + \sum_{j=1}^q \theta_j \varepsilon_{n+1-j} \quad (3-5)$$

donde todos los términos del lado derecho son conocidos. La varianza condicional es constante e igual a la varianza del ruido:

$$\text{Var}[Y_{n+1} | \mathcal{F}_n] = \sigma^2 \quad (3-6)$$

Por lo tanto, si el ruido ε_t sigue una distribución F con media cero y varianza σ^2 , la

distribución predictiva verdadera es:

$$Y_{n+1} \mid \mathcal{F}_n \sim F(\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n], \sigma^2) \quad (3-7)$$

Esta distribución puede evaluarse numéricamente generando una muestra grande de errores futuros $\varepsilon_{n+1}^{(b)} \sim F(0, \sigma^2)$ y calculando:

$$Y_{n+1}^{(b)} = c + \sum_{i=1}^p \phi_i Y_{n+1-i} + \sum_{j=1}^q \theta_j \varepsilon_{n+1-j} + \varepsilon_{n+1}^{(b)}, \quad b = 1, \dots, B \quad (3-8)$$

donde B es un número suficientemente grande (en esta investigación, $B = 1000$). Esta muestra empírica aproxima la distribución predictiva verdadera y sirve como referencia para el cálculo del ECRPS.

Configuraciones Paramétricas Evaluadas

La Tabla 3-1 presenta las siete configuraciones ARMA consideradas en este estudio. La selección incluye modelos puramente autorregresivos [AR(1), AR(2)], puramente de media móvil [MA(1), MA(2)], y mixtos [ARMA(1,1), ARMA(2,2), ARMA(2,1)], con diferentes grados de persistencia temporal y complejidad estructural.

Nombre	p	q	ϕ	θ
AR(1)	1	0	[0.9]	[]
AR(2)	2	0	[0.5, -0.3]	[]
MA(1)	0	1	[]	[0.7]
MA(2)	0	2	[]	[0.4, 0.2]
ARMA(1,1)	1	1	[0.6]	[0.3]
ARMA(2,2)	2	2	[0.4, -0.2]	[0.5, 0.1]
ARMA(2,1)	2	1	[0.7, 0.2]	[0.5]

Tabla 3-1.: Configuraciones paramétricas para procesos ARMA.

Todas las configuraciones fueron verificadas para satisfacer las condiciones de estacionariedad e invertibilidad mediante el cálculo numérico de las raíces de los polinomios característicos correspondientes.

3.3.2. Procesos ARIMA: Escenario Lineal No Estacionario

Definición y Operador de Diferenciación

Un proceso autorregresivo integrado de media móvil de órdenes (p, d, q) , denotado ARIMA(p, d, q), se construye aplicando el operador de diferenciación $\Delta = 1 - L$ un total de d veces a una serie Y_t y modelando la serie diferenciada resultante $W_t = \Delta^d Y_t$ mediante un proceso ARMA(p, q) estacionario:

$$\Phi(L)W_t = c + \Theta(L)\varepsilon_t \quad (3-9)$$

donde $W_t = (1 - L)^d Y_t$.

Equivalentemente, en términos de la serie original:

$$\Phi(L)(1 - L)^d Y_t = c + \Theta(L)\varepsilon_t \quad (3-10)$$

El orden de integración d representa el número de raíces unitarias en el polinomio autorregresivo ampliado. En la gran mayoría de aplicaciones prácticas, $d \in \{0, 1, 2\}$, siendo $d = 1$ el caso más frecuente (Rob J Hyndman and Athanasopoulos 2021).

Propiedades de Estacionariedad

Un proceso ARIMA(p, d, q) es no estacionario por construcción cuando $d > 0$, debido a la presencia de raíces unitarias. Sin embargo, la serie diferenciada $W_t = \Delta^d Y_t$ es estacionaria si el componente ARMA(p, q) subyacente satisface las condiciones de estacionariedad e invertibilidad descritas en la Sección 3.3.1.

Esta propiedad de *estacionariedad en diferencias* es fundamental para el pronóstico, ya que permite aplicar toda la teoría desarrollada para procesos estacionarios a la serie transformada W_t , recuperando posteriormente los pronósticos en la escala original mediante integración sucesiva (Rob J Hyndman and Athanasopoulos 2021).

Distribución Predictiva Verdadera

Para un proceso ARIMA(p, d, q), la distribución del siguiente valor Y_{n+1} condicionada a la historia observada se obtiene mediante un procedimiento de dos etapas que explota la

estructura de diferenciación del modelo.

Primero, se predice el siguiente valor de la serie diferenciada $W_{n+1} = \Delta^d Y_{n+1}$ usando la distribución ARMA subyacente. Para el caso más común $d = 1$, la serie diferenciada es:

$$W_t = Y_t - Y_{t-1} \quad (3-11)$$

y su predicción un paso adelante, condicionada a la historia $\mathcal{F}_n = \{Y_1, \dots, Y_n, \varepsilon_1, \dots, \varepsilon_n\}$, sigue la distribución ARMA:

$$W_{n+1} | \mathcal{F}_n \sim F(\mathbb{E}[W_{n+1} | \mathcal{F}_n], \sigma^2) \quad (3-12)$$

donde:

$$\mathbb{E}[W_{n+1} | \mathcal{F}_n] = c + \sum_{i=1}^p \phi_i W_{n+1-i} + \sum_{j=1}^q \theta_j \varepsilon_{n+1-j} \quad (3-13)$$

Segundo, se recupera la predicción en la escala original mediante la relación de integración:

$$Y_{n+1} = Y_n + W_{n+1} \quad (3-14)$$

Por lo tanto, la distribución predictiva verdadera para Y_{n+1} es:

$$Y_{n+1} | \mathcal{F}_n \sim F(Y_n + \mathbb{E}[W_{n+1} | \mathcal{F}_n], \sigma^2) \quad (3-15)$$

Esta distribución puede evaluarse numéricamente generando muestras del incremento futuro:

$$W_{n+1}^{(b)} = c + \sum_{i=1}^p \phi_i W_{n+1-i} + \sum_{j=1}^q \theta_j \varepsilon_{n+1-j} + \varepsilon_{n+1}^{(b)} \quad (3-16)$$

y aplicando la transformación:

$$Y_{n+1}^{(b)} = Y_n + W_{n+1}^{(b)}, \quad b = 1, \dots, B \quad (3-17)$$

donde $\varepsilon_{n+1}^{(b)} \sim F(0, \sigma^2)$ son errores futuros independientes. Esta muestra empírica representa la distribución predictiva verdadera que sirve como referencia para el ECRPS.

Configuraciones Paramétricas Evaluadas

La Tabla 3-2 presenta las siete configuraciones ARIMA($p, 1, q$) consideradas en este estudio. Todas las configuraciones utilizan $d = 1$, reflejando el caso más común en aplicaciones económicas y financieras. La selección incluye desde el paseo aleatorio puro [ARIMA(0,1,0)] hasta modelos con estructura autorregresiva y de media móvil en la serie diferenciada.

Nombre	p	d	q	ϕ	θ
ARIMA(0,1,0)	0	1	0	[]	[]
ARIMA(1,1,0)	1	1	0	[0.6]	[]
ARIMA(2,1,0)	2	1	0	[0.5, -0.2]	[]
ARIMA(0,1,1)	0	1	1	[]	[0.5]
ARIMA(0,1,2)	0	1	2	[]	[0.4, 0.25]
ARIMA(1,1,1)	1	1	1	[0.7]	[-0.3]
ARIMA(2,1,2)	2	1	2	[0.6, 0.2]	[0.4, -0.1]

Tabla 3-2.: Configuraciones paramétricas para procesos ARIMA.

3.3.3. Procesos SETAR: Escenario No Lineal Estacionario

Definición y Mecanismo de Cambio de Régimen

Un modelo autorregresivo de umbral auto-excitado con dos régímenes, denotado SETAR(2; p_1, p_2), se define mediante una estructura de cambio de régimen determinado por valores pasados de la propia serie (P. Chen and Semmler 2023):

$$Y_t = \begin{cases} \phi_0^{(1)} + \sum_{i=1}^{p_1} \phi_i^{(1)} Y_{t-i} + \varepsilon_t^{(1)} & \text{si } Y_{t-d} \leq r \\ \phi_0^{(2)} + \sum_{i=1}^{p_2} \phi_i^{(2)} Y_{t-i} + \varepsilon_t^{(2)} & \text{si } Y_{t-d} > r \end{cases} \quad (3-18)$$

donde:

- r es el *valor umbral* (threshold value) que determina el cambio de régimen
- d es el *rezago de umbral* (threshold delay) que especifica qué valor pasado de la serie se utiliza para determinar el régimen activo

- $\phi_0^{(j)}$ y $\{\phi_i^{(j)}\}_{i=1}^{p_j}$ son los parámetros específicos del régimen j
- $\varepsilon_t^{(j)} \sim WN(0, \sigma_j^2)$ son procesos de ruido blanco que pueden tener varianzas diferentes en cada régimen

La notación SETAR($2; d, p$) denota un modelo de dos regímenes con rezago de umbral d y orden autorregresivo común p en ambos regímenes (aunque en general p_1 y p_2 pueden diferir).

Estacionariedad en Procesos SETAR

La estacionariedad de procesos SETAR es sustancialmente más compleja que en modelos lineales, ya que la dinámica cambia endógenamente según el estado del sistema. Las condiciones suficientes para la estacionariedad han sido objeto de extensa investigación (P. Chen and Semmler 2023).

Caso SETAR(2; 1, 1): Para el caso más simple de dos regímenes con orden autorregresivo 1, Petruccelli and Woolford 1984 demostraron que el proceso es ergódico si y solo si:

$$|\phi_1^{(1)}| < 1, \quad |\phi_1^{(2)}| < 1, \quad \text{y} \quad |\phi_1^{(1)}\phi_1^{(2)}| < 1 \quad (3-19)$$

Esta condición requiere que cada régimen sea individualmente estable y que el producto de los coeficientes autorregresivos sea menor que uno en valor absoluto. Esta última condición captura el efecto de la interacción entre regímenes.

Caso General SETAR($2; p_1, p_2$): Para órdenes autorregresivos mayores, Chan and Tong 1985 proporcionaron una condición suficiente basada en el radio espectral de las matrices compañeras:

$$\max_j \sum_{i=1}^{p_j} |\phi_i^{(j)}| < 1 \quad (3-20)$$

Sin embargo, esta condición es bastante conservadora. Un criterio más general y menos restrictivo se basa en el concepto de *radio espectral conjunto* (joint spectral radius) de las matrices compañeras de ambos regímenes (P. Chen and Semmler 2023). Sea $\Phi^{(j)}$ la matriz compañera del régimen j :

$$\Phi^{(j)} = \begin{pmatrix} \phi_1^{(j)} & \phi_2^{(j)} & \cdots & \phi_{p-1}^{(j)} & \phi_p^{(j)} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \quad (3-21)$$

El radio espectral conjunto se define como:

$$\rho(\{\Phi^{(1)}, \Phi^{(2)}\}) = \lim_{k \rightarrow \infty} \max \|\Phi^{(i_1)} \dots \Phi^{(i_k)}\|^{1/k} \quad (3-22)$$

donde el máximo se toma sobre todas las secuencias posibles de k matrices.

El proceso SETAR es estacionario si $\rho(\{\Phi^{(1)}, \Phi^{(2)}\}) < 1$. Este criterio es menos restrictivo que (3-20) y permite que algunos regímenes individuales sean incluso explosivos, siempre que la dinámica global del sistema sea estabilizadora (P. Chen and Semmler 2023).

Distribución Predictiva Verdadera

La distribución del siguiente valor Y_{n+1} en un proceso SETAR condicionada a la historia observada $\mathcal{F}_n = \{Y_1, \dots, Y_n, \varepsilon_1, \dots, \varepsilon_n\}$ depende críticamente del régimen que será activado en el tiempo $n + 1$. A diferencia de los modelos lineales, la predicción requiere determinar primero qué régimen gobernará la dinámica futura.

El régimen activo en el tiempo $n + 1$ se determina comparando el valor retardado Y_{n+1-d} con el umbral r :

$$\text{Régimen}_{n+1} = \begin{cases} 1 & \text{si } Y_{n+1-d} \leq r \\ 2 & \text{si } Y_{n+1-d} > r \end{cases} \quad (3-23)$$

Dado que Y_{n+1-d} ya es conocido en el tiempo n (pues $n + 1 - d \leq n$ para $d \geq 1$), el régimen futuro es determinístico y no hay incertidumbre sobre cuál dinámica aplicar. Una vez identificado el régimen $j \in \{1, 2\}$, la media condicional se calcula mediante:

$$\mathbb{E}[Y_{n+1} | \mathcal{F}_n, \text{Régimen}_{n+1} = j] = \phi_0^{(j)} + \sum_{i=1}^{p_j} \phi_i^{(j)} Y_{n+1-i} \quad (3-24)$$

donde todos los valores Y_{n+1-i} en el lado derecho son observados. La varianza condicional

es constante dentro de cada régimen:

$$\text{Var}[Y_{n+1} | \mathcal{F}_n, \text{Régimen}_{n+1} = j] = \sigma_j^2 \quad (3-25)$$

Por lo tanto, la distribución predictiva verdadera es:

$$Y_{n+1} | \mathcal{F}_n \sim F_j(\mathbb{E}[Y_{n+1} | \mathcal{F}_n, \text{Régimen}_{n+1} = j], \sigma_j^2) \quad (3-26)$$

donde F_j es la distribución del ruido en el régimen j y el subíndice j se determina mediante (3-23).

Esta distribución puede evaluarse numéricamente generando una muestra grande de errores futuros específicos del régimen activo:

$$Y_{n+1}^{(b)} = \phi_0^{(j)} + \sum_{i=1}^{p_j} \phi_i^{(j)} Y_{n+1-i} + \varepsilon_{n+1}^{(b)}, \quad b = 1, \dots, B \quad (3-27)$$

donde $\varepsilon_{n+1}^{(b)} \sim F_j(0, \sigma_j^2)$ son errores independientes del régimen determinado. A diferencia de los modelos ARMA, aquí no existe incertidumbre sobre el régimen en predicciones un paso adelante, lo que simplifica considerablemente la evaluación de la distribución predictiva verdadera.

Configuraciones Paramétricas Evaluadas

La Tabla 3-3 presenta las siete configuraciones SETAR consideradas en este estudio. Las configuraciones incluyen diferentes órdenes autorregresivos, rezagos de umbral y valores de umbral, representando una amplia gama de comportamientos no lineales.

Finalmente, es importante destacar que todas las configuraciones detalladas en la Tabla 3-3 fueron seleccionadas bajo un estricto criterio de estabilidad. Para garantizar el rigor estadístico de las comparaciones en este escenario, se realizó un análisis de estacionariedad basado en el cálculo numérico del radio espectral conjunto (ρ) para cada par de matrices compañeras. Se verificó que en la totalidad de los casos empleados en la simulación se cumple la condición $\rho < 1$, asegurando que los procesos SETAR generados son globalmente estacionarios.

Nombre	$\phi^{(1)}$	$\phi^{(2)}$	r	d
SETAR-1	[0.6]	[-0.5]	0.0	1
SETAR-2	[0.7]	[-0.7]	0.0	2
SETAR-3	[0.5, -0.2]	[-0.3, 0.1]	0.5	1
SETAR-4	[0.8, -0.15]	[-0.6, 0.2]	1.0	2
SETAR-5	[0.4, -0.1, 0.05]	[-0.3, 0.1, -0.05]	0.0	1
SETAR-6	[0.5, -0.3, 0.1]	[-0.4, 0.2, -0.05]	0.5	2
SETAR-7	[0.3, 0.1]	[-0.2, -0.1]	0.8	3

Tabla 3-3.: Configuraciones paramétricas para procesos SETAR.

3.4. Modelos predictivos

Para evaluar la capacidad de cuantificación de la incertidumbre en diversos entornos estocásticos, esta investigación emplea un conjunto heterogéneo de nueve modelos predictivos. Esta selección abarca desde métodos de remuestreo clásicos y propuestas de predicción conformal, hasta arquitecturas de aprendizaje profundo y modelos híbridos. El uso de esta diversidad de enfoques permite contrastar cómo las garantías teóricas de cada familia de modelos se traducen en un rendimiento práctico bajo la métrica ECRPS, especialmente cuando se enfrentan a la ruptura de los supuestos de intercambiabilidad y linealidad.

3.4.1. Modelos tomados de la literatura

Circular Block Bootstrap (CBB)

Explicación Teórica del Modelo El método *Circular Block Bootstrap* (CBB), introducido por Politis and Romano (1992), representa una evolución metodológica del remuestreo por bloques que aborda una limitación fundamental de los esquemas no circulares como el Moving Block Bootstrap (MBB). Según Lahiri (2003), el problema radica en que las observaciones ubicadas en los extremos de la serie temporal $\{X_1, \dots, X_n\}$ aparecen con menor frecuencia en los bloques remuestreados, generando una infra-representación sistemática de los bordes y sesgo en la estimación de varianza.

Fundamento Teórico El CBB resuelve esta asimetría mediante la *circunscripción* de los datos: la serie temporal se conceptualiza como una estructura circular donde X_n es

seguido inmediatamente por X_1 , permitiendo la continuidad periódica. Formalmente, para una serie de longitud n y bloques de tamaño l , se definen exactamente n bloques posibles:

$$B_i = \{X_i, X_{i+1 \bmod n}, \dots, X_{i+l-1 \bmod n}\}, \quad i = 1, \dots, n \quad (3-28)$$

donde el operador módulo (\bmod) implementa la extensión circular. Esta construcción garantiza que cada observación histórica tiene probabilidad idéntica $1/n$ de ser seleccionada como punto de inicio de un bloque, eliminando el sesgo de borde.

Algoritmo de Remuestreo El procedimiento de generación de muestras bootstrap opera en dos etapas:

1. **Muestreo de puntos de inicio:** Se seleccionan B índices $\{i_1, \dots, i_B\}$ uniformemente de $\{1, \dots, n\}$, donde B es el número de réplicas bootstrap deseadas.
2. **Construcción de bloques circulares:** Cada réplica X_b^* se forma extrayendo el bloque circular iniciado en i_b :

$$X_b^* = X_{(i_b+r) \bmod n}, \quad r \in \{0, 1, \dots, l-1\} \quad (3-29)$$

Para pronóstico un paso adelante, la distribución predictiva se approxima mediante el conjunto de valores finales de cada bloque: $\{\hat{y}_{t+1}^{(1)}, \dots, \hat{y}_{t+1}^{(B)}\}$, donde $\hat{y}_{t+1}^{(b)} = X_b^*$.

Propiedades Estadísticas Lahiri (2003) establece que bajo condiciones de mixing (dependencia que decae con el tiempo), el estimador CBB de la varianza es consistente cuando $l \rightarrow \infty$ y $l/n \rightarrow 0$ conforme $n \rightarrow \infty$. La equiprobabilidad de selección garantiza distribuciones predictivas mejor calibradas en contextos de dependencia temporal, haciendo al CBB apropiado para series financieras y económicas donde la estructura de autocorrelación es relevante.

Clasificación del Modelo El CBB es un método **no paramétrico** puro: no asume ninguna forma funcional para la distribución subyacente de los datos ni estima parámetros poblacionales. La distribución predictiva emerge directamente del remuestreo empírico de la historia observada, preservando las características distribucionales y de dependencia presentes en la muestra sin imponer supuestos estructurales.

De la Teoría a la Práctica La implementación desarrollada introduce tres adaptaciones principales respecto a la formulación teórica estándar:

Adaptación 1: Simplificación del Esquema de Remuestreo La teoría clásica del CBB (Politis and Romano 1992) genera bloques completos de longitud l que luego se concatenan para formar series bootstrap de longitud n . En contraste, la implementación para pronóstico un paso adelante simplifica el proceso: dado que solo se requiere predecir \hat{y}_{t+1} , se muestrea directamente el valor en la posición $(i_b + r) \bmod n$ donde $r = n \bmod l$ representa la posición relativa dentro del último bloque histórico. Esta simplificación reduce la complejidad computacional de $O(Bl)$ a $O(B)$ operaciones de indexación.

Adaptación 2: Selección Automática de l Mientras que la teoría requiere especificación manual del tamaño de bloque basado en análisis del proceso estocástico subyacente, la implementación incorpora la heurística automática de Politis and White (2004): $l \approx 1.5 \times n^{1/3}$, ademas de otros valores de referencia para balancear eficiencia y captura de dependencia.

Adaptación 3: Congelamiento Post-Optimización A diferencia de implementaciones estándar que podrían recalcular l en cada paso, el diseño experimental congela el hiperparámetro tras la fase de validación, este congelamiento es crítico para prevenir *data leakage*: re-optimizar en ventanas rolling introduciría información futura en la selección del modelo, violando la evaluación predictiva rigurosa.

Aspecto	Teoría Clásica	Implementación
Remuestreo	Bloques completos concatenados	Valor directo $(i + n \bmod l) \bmod n$
Longitud de bloque	Manual / dependiente del contexto	Optimización
Actualización de l	No especificada	Congelada post-optimización
Complejidad	$O(Bl)$ para serie completa	$O(B)$ para un pronóstico

Tabla 3-4.: Comparativa: Teoría vs. Implementación del CBB

Optimización, Parámetros e Hiperparámetros

Hiperparámetro Principal: `block_length (l)` Controla la cantidad de dependencia temporal preservada en las réplicas bootstrap. Su configuración óptima se determina mediante una estrategia de optimización que emplea una búsqueda en grilla. La grilla de valores candidatos para l está definida por las siguientes cuatro opciones, que representan diferentes escalas de dependencia temporal en función del tamaño de muestra n :

1. $l = 5$, para modelar estructuras de dependencia de corto plazo.
2. $l = \lfloor 1.5 \cdot n^{1/3} \rfloor$, una aproximación heurística común en métodos de *block bootstrap*.
3. $l = \lfloor \sqrt{n} \rfloor$, que ofrece un balance entre corto y mediano plazo.
4. $l = \lfloor n/5 \rfloor$, diseñada para capturar posibles estructuras de dependencia de largo plazo.

La selección del valor óptimo de l de entre estas cuatro opciones se realiza minimizando el **CRPS promedio** (ECRPS) sobre un conjunto de validación, conforme a la métrica detallada en la subsección 2.2.3.

Protocolo de Congelamiento Una vez identificado el valor óptimo mediante validación, este se congela y se utiliza de manera fija durante la fase de prueba. Esta lógica garantiza que no haya contaminación de información y que el modelo evaluado sea idéntico al seleccionado durante la validación.

Sieve Bootstrap (SB)

Explicación Teórica del Modelo El *Sieve Bootstrap*, introducido por Bühlmann (1997) y analizado en profundidad por Lahiri (2003), representa un enfoque alternativo al remuestreo por bloques para series temporales dependientes. En lugar de preservar la dependencia mediante partición física de la serie, el método emplea una aproximación paramétrica para filtrar la estructura de autocorrelación.

Fundamento Teórico El Sieve Bootstrap se fundamenta en el teorema de Wold, que establece que cualquier proceso estocástico estacionario linealmente regular admite una representación autorregresiva de orden infinito, AR(∞). Formalmente, para una serie

temporal $\{X_t\}$ estacionaria con media μ , existe una representación:

$$X_t - \mu = \sum_{j=1}^{\infty} \phi_j (X_{t-j} - \mu) + \epsilon_t \quad (3-30)$$

donde $\{\epsilon_t\}$ es un proceso de innovaciones i.i.d. con media cero y varianza σ^2 .

En la práctica, esta representación infinita se aproxima mediante un modelo autorregresivo finito AR(p) donde p crece con el tamaño muestral n :

$$X_t = \phi_0 + \sum_{j=1}^p \phi_j X_{t-j} + \epsilon_t \quad (3-31)$$

Algoritmo de Remuestreo El procedimiento del Sieve Bootstrap opera en tres etapas secuenciales:

1. **Ajuste del tamiz autorregresivo:** Se estima el modelo AR(p) mediante mínimos cuadrados ordinarios sobre la serie histórica, obteniendo coeficientes $\hat{\phi} = (\hat{\phi}_0, \hat{\phi}_1, \dots, \hat{\phi}_p)$.
2. **Extracción de residuos:** Se calculan los residuos del modelo ajustado:

$$\hat{\epsilon}_t = X_t - \hat{\phi}_0 - \sum_{j=1}^p \hat{\phi}_j X_{t-j}, \quad t = p+1, \dots, n \quad (3-32)$$

que idealmente deben comportarse como realizaciones i.i.d. Estos residuos se centran: $\tilde{\epsilon}_t = \hat{\epsilon}_t - \bar{\epsilon}$.

3. **Generación de muestras bootstrap:** Para cada réplica $b = 1, \dots, B$:

- Se remuestrean con reemplazo los residuos centrados: $\epsilon_b^* \sim \{\tilde{\epsilon}_{p+1}, \dots, \tilde{\epsilon}_n\}$
- Se genera la predicción un paso adelante:

$$\hat{X}_{n+1}^{(b)} = \hat{\phi}_0 + \sum_{j=1}^p \hat{\phi}_j X_{n+1-j} + \epsilon_b^* \quad (3-33)$$

La distribución predictiva empírica está dada por el conjunto $\{\hat{X}_{n+1}^{(1)}, \dots, \hat{X}_{n+1}^{(B)}\}$.

Propiedades de Consistencia Bühlmann (1997) demuestra que bajo condiciones de regularidad (estacionaridad, ergodicidad, y $p = p_n \rightarrow \infty$ con $p_n^3/n \rightarrow 0$), el Sieve Bootstrap aproxima consistentemente la distribución del estimador de interés. La clave es que el orden p debe crecer suficientemente para capturar la dependencia, pero no tan rápido como para introducir varianza excesiva por sobreparametrización.

Clasificación del Modelo El Sieve Bootstrap es un método **semiparamétrico**: utiliza una estructura paramétrica (el modelo AR) para filtrar la dependencia temporal, pero trata la distribución de los residuos de forma no paramétrica mediante bootstrap empírico. No asume una forma distribucional específica para las innovaciones, solo que sean aproximadamente i.i.d. después del filtrado AR.

De la Teoría a la Práctica La implementación desarrollada incorpora tres adaptaciones clave para el contexto de pronóstico rolling:

Adaptación 1: Selección de Orden Basada en Validación Mientras la teoría asintótica sugiere $p \rightarrow \infty$ con n , la implementación emplea una grilla discreta de órdenes candidatos $p \in \{5, 10, 20\}$ evaluados mediante ECRPS en validación. Esta discretización responde a dos consideraciones prácticas: (i) evitar sobreparametrización en muestras finitas ($n = 200$), y (ii) reducir tiempo computacional frente a búsquedas exhaustivas.

Adaptación 2: Congelamiento de Parámetros AR La implementación introduce un mecanismo de congelamiento crítico: durante la fase de calibración, se ajusta el modelo AR(p^*) con orden óptimo p^* sobre los datos de entrenamiento+calibración combinados, almacenando permanentemente:

- Coeficientes autorregresivos: $\hat{\phi}^* = (\hat{\phi}_0^*, \dots, \hat{\phi}_{p^*}^*)$
- Residuos centrados: $\tilde{\epsilon}^* = \{\tilde{\epsilon}_{p^*+1}, \dots, \tilde{\epsilon}_{n_{\text{calib}}}\}$

En cada ventana rolling subsecuente, se reutilizan $\hat{\phi}^*$ y $\tilde{\epsilon}^*$ sin re-estimación, aplicando solo los últimos p^* valores observados para generar la predicción. Esta estrategia previene data leakage y reduce variabilidad numérica.

Adaptación 3: Predicción Secuencial Eficiente En lugar de generar series bootstrap completas de longitud n (complejidad $O(Bn)$), la implementación genera directamente predicciones un paso adelante (complejidad $O(B)$): dado el vector de historia reciente $\mathbf{X}_{n-p^*:n} = (X_{n-p^*+1}, \dots, X_n)$, cada predicción se calcula como:

$$\hat{X}_{n+1}^{(b)} = \hat{\phi}_0^* + \sum_{j=1}^{p^*} \hat{\phi}_j^* X_{n+1-j} + \epsilon_b^* \quad (3-34)$$

donde ϵ_b^* se muestrea de $\tilde{\epsilon}^*$ con reemplazo.

Aspecto	Teoría Clásica	Implementación
Orden AR	$p \rightarrow \infty$ con n	Grilla discreta $\{5, 10, 20\}$
Selección de p	Criterios asintóticos	Validación cruzada (ECRPS)
Parámetros $\hat{\phi}$	Re-estimados en cada muestra	Congelados post-calibración
Residuos	Recalculados dinámicamente	Pool fijo $\tilde{\epsilon}^*$
Complejidad	$O(Bn)$	$O(B)$

Tabla 3-5.: Comparativa: Teoría vs. Implementación del Sieve Bootstrap

Optimización, Parámetros e Hiperparámetros

Hiperparámetro Principal: `order (p)` Define la profundidad del tamiz autorregresivo, controlando cuánta memoria del proceso se captura. Durante la fase de validación se evalúan tres configuraciones sobre $n_{\text{train}} = 200$:

- $p = 5$ (dependencias de corto plazo, hasta una semana)
- $p = 10$ (memoria intermedia, aproximadamente dos semanas)
- $p = 20$ (dependencias extendidas, un mes)

Métrica de Optimización La selección del orden óptimo p^* se realiza minimizando el **ECRPS** (véase 2.2.3) sobre el conjunto de validación. El orden seleccionado es aquel que produce las distribuciones predictivas más calibradas durante la fase de validación.

Protocolo de Congelamiento Una vez identificado p^* mediante validación, el método ajusta el modelo $\text{AR}(p^*)$ sobre los datos de entrenamiento+calibración y almacena perma-

nentemente los coeficientes $\hat{\phi}^*$ y residuos centrados $\tilde{\epsilon}^*$. Estos parámetros se reutilizan sin re-estimación en toda la fase de prueba rolling, previniendo data leakage y garantizando evaluación rigurosa.

Least Squares Prediction Machine (LSPM)

Explicación Teórica del Modelo El *Least Squares Prediction Machine* (LSPM), introducido por Vovk, Gammerman, and Shafer (2022), representa una evolución de la predicción conformal que trasciende la generación de intervalos de confianza para construir distribuciones predictivas completas. A diferencia de los predictores conformales estándar, el LSPM se define como un Sistema Predictivo Conformal (CPS), cuya salida es una Función de Distribución Predictiva Conformal (CPD).

Fundamento Teórico: Predicción Conformal La predicción conformal se fundamenta en el principio de intercambiabilidad: dada una secuencia de pares $(x_1, y_1), \dots, (x_{n-1}, y_{n-1})$ y un nuevo objeto x_n , se asume que para cualquier etiqueta candidata y , la secuencia aumentada $(x_1, y_1), \dots, (x_{n-1}, y_{n-1}), (x_n, y)$ es intercambiable. Bajo este supuesto, se puede construir una distribución predictiva válida sin asumir una forma paramétrica específica para los errores.

El LSPM utiliza mínimos cuadrados ordinarios (OLS) como algoritmo subyacente. La versión más robusta es el **LSPM Studentizado**, que emplea los elementos diagonales de la matriz de proyección (matriz hat) \bar{H} para normalizar los residuos. Según Vovk, Gammerman, and Shafer (2022), esta normalización garantiza que la distribución predictiva sea monótonamente creciente, incluso cuando el nuevo objeto posee alto apalancamiento (*leverage*).

Construcción de la Distribución Predictiva Para un conjunto de datos aumentado que incluye $n - 1$ ejemplos de entrenamiento $(x_1, y_1), \dots, (x_{n-1}, y_{n-1})$ y un nuevo objeto x_n con etiqueta hipotética y , se construye la matriz de diseño aumentada:

$$\bar{X} = \begin{pmatrix} 1 & x_1^T \\ \vdots & \vdots \\ 1 & x_{n-1}^T \\ 1 & x_n^T \end{pmatrix} \quad (3-35)$$

La matriz hat se define como:

$$\bar{H} = \bar{X}(\bar{X}^T \bar{X})^{-1} \bar{X}^T \quad (3-36)$$

donde $h_{i,j}$ denota el elemento en la fila i y columna j de \bar{H} .

Valores Críticos Studentizados La distribución predictiva se construye mediante valores críticos C_i que actúan como puntos de salto de una función escalonada. Para la versión studentizada, según las ecuaciones 7.15 y 7.16 de Vovk, Gammerman, and Shafer (2022):

$$C_i = \frac{A_i}{B_i}, \quad i = 1, \dots, n - 1 \quad (3-37)$$

donde:

$$B_i = \sqrt{1 - h_{n,n}} + \frac{h_{i,n}}{\sqrt{1 - h_{i,i}}} \quad (3-38)$$

$$A_i = \frac{\sum_{j=1}^{n-1} h_{j,n} y_j}{\sqrt{1 - h_{n,n}}} + \frac{y_i - \sum_{j=1}^{n-1} h_{i,j} y_j}{\sqrt{1 - h_{i,i}}} \quad (3-39)$$

El término A_i combina dos componentes: (i) la predicción OLS estándar sobre el punto nuevo, normalizada por su leverage, y (ii) el residuo studentizado del punto i en un ajuste leave-one-out. El término B_i normaliza esta combinación considerando el leverage tanto del punto nuevo ($h_{n,n}$) como del punto histórico ($h_{i,i}$) y su covarianza ($h_{i,n}$).

Propiedades Estadísticas Vovk, Gammerman, and Shafer (2022) demuestra que bajo intercambiabilidad, la función de distribución construida mediante estos valores críticos es estadísticamente válida: para cualquier nivel de confianza α , el intervalo de predicción conformal tiene cobertura exacta $1 - \alpha$ en expectativa sobre la secuencia intercambiable. La studentización es crucial para mantener esta validez incluso cuando $h_{n,n} \rightarrow 1$ (leverage extremo del punto de prueba).

Clasificación del Modelo El LSPM es un método **no paramétrico** basado en distribución-libre (*distribution-free*). Aunque utiliza OLS como algoritmo subyacente, no asume ninguna forma distribucional para los errores ϵ_i . La validez estadística proviene

únicamente del supuesto de intercambiabilidad, no de supuestos gaussianos o paramétricos sobre los residuos.

De la Teoría a la Práctica La implementación desarrollada adapta el marco teórico de Vovk al contexto específico de pronóstico en series temporales:

Adaptación 1: Construcción Autorregresiva La teoría original asume vectores de características x_i independientes. En series temporales, se construyen objetos dinámicamente mediante retardos: dado p lags, cada observación se transforma en un vector autorregresivo:

$$x_t = (y_{t-1}, y_{t-2}, \dots, y_{t-p})^T \quad (3-40)$$

Esto convierte al LSPM en un predictor conformal autorregresivo AR(p), donde la matriz de diseño se construye dinámicamente en cada ventana rolling usando los últimos n valores disponibles.

Adaptación 2: Estabilidad Numérica Para calcular $\bar{H} = \bar{X}(\bar{X}^T \bar{X})^{-1} \bar{X}^T$, la implementación emplea la pseudoinversa de Moore-Penrose en lugar de la inversa estándar. Esta decisión es crítica: series temporales con alta autocorrelación generan matrices $\bar{X}^T \bar{X}$ casi singulares (número de condición alto), causando inestabilidad numérica en la inversión directa. La pseudoinversa provee una solución regularizada que previene errores computacionales.

Adaptación 3: Filtrado de Singularidades La teoría requiere que $h_{i,i} < 1$ y $h_{n,n} < 1$ para que los denominadores en A_i y B_i sean no nulos. La implementación incorpora dos mecanismos de seguridad:

- Umbral de tolerancia: Se filtran observaciones con $|1 - h_{i,i}| < 10^{-10}$ o $|1 - h_{n,n}| < 10^{-10}$.
- Filtrado de divisores: Se descartan valores críticos donde $|B_i| < 10^{-10}$.

Estos filtros previenen divisiones por cero cuando un punto es tan influyente que el modelo lo ajusta sin residuo.

Aspecto	Teoría Clásica	Implementación
Vectores x_i	Características independientes	Retardos AR: $(y_{t-1}, \dots, y_{t-p})$
Cálculo de \bar{H}	Inversa $(\bar{X}^T \bar{X})^{-1}$	Pseudoinversa (Moore-Penrose)
Singularidades	Supuesto $h_{i,i}, h_{n,n} < 1$	Filtrado explícito (10^{-10})
Valores críticos	Todos los $n - 1$ puntos	Solo puntos con B_i válido

Tabla 3-6.: Comparativa: Teoría vs. Implementación del LSPM

Optimización, Parámetros e Hiperparámetros

Parámetro Principal: `n_lags (p)` Define el orden autorregresivo del modelo, controlando cuántos retardos se usan para construir la matriz de diseño. La implementación emplea la heurística:

$$p = \max(1, \lfloor n^{1/3} \rfloor) \quad (3-41)$$

Esta regla está motivada por consideraciones teóricas de métodos de tamiz (sieve methods) en estadística no paramétrica, donde el número de parámetros p debe crecer con el tamaño muestral n pero a una tasa controlada que preserve consistencia. La tasa $n^{1/3}$ es estándar en la literatura de bootstrap para series temporales (Bühlmann 1997; Politis and White 2004), garantizando que $p \rightarrow \infty$ conforme $n \rightarrow \infty$, pero manteniendo $p^3/n \rightarrow 0$ para evitar sobreajuste. Esta misma tasa aparece en la heurística de longitud de bloque del CBB, reflejando un principio unificado: el número de parámetros debe escalar subcuadráticamente con la muestra. Para $n = 200$, resulta $p \approx 5$ lags, suficiente para capturar autocorrelaciones de corto plazo sin saturar los grados de libertad del modelo OLS subyacente.

Métrica de Optimización Dado que el LSPM es un método conformal teóricamente válido sin hiperparámetros libres (la versión studentizada es la única apropiada según Vovk, Gammerman, and Shafer (2022)), la optimización se limita a seleccionar p mediante la heurística anterior.

Protocolo de Congelamiento El método ejecuta una operación crítica: congela p basado en el tamaño del conjunto de entrenamiento+calibración. Este valor congelado p^* se almacena y se reutiliza en todas las ventanas rolling subsecuentes, garantizando que la dimensionalidad de la matriz de diseño permanezca constante y previniendo data leakage.

DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks

Explicación Teórica del Modelo DeepAR, introducido por Salinas et al. (2020), representa un cambio de paradigma en el pronóstico probabilístico de series temporales al trasladar el enfoque desde el modelado individual de cada serie hacia el **aprendizaje de un modelo global** a partir de múltiples series relacionadas mediante una arquitectura de red neuronal recurrente autorregresiva.

Fundamento Arquitectónico El modelo emplea redes LSTM (Long Short-Term Memory) para procesar secuencias temporales de forma autorregresiva. Para una serie temporal i con valores $z_{i,t}$, DeepAR modela la distribución condicional del futuro dado el pasado:

$$P(z_{i,t_0:T} \mid z_{i,1:t_0-1}, x_{i,1:T}) \quad (3-42)$$

donde t_0 denota el punto de inicio del horizonte de predicción, $z_{i,1:t_0-1}$ representa el rango de condicionamiento, $z_{i,t_0:T}$ los valores futuros, y $x_{i,1:T}$ son covariables conocidas.

La arquitectura factoriza esta distribución mediante el producto de verosimilitudes condicionales:

$$Q_\Theta(z_{i,t_0:T} \mid z_{i,1:t_0-1}, x_{i,1:T}) = \prod_{t=t_0}^T Q_\Theta(z_{i,t} \mid z_{i,1:t-1}, x_{i,1:T}) \quad (3-43)$$

donde cada factor está parametrizado por la salida de la red recurrente:

$$Q_\Theta(z_{i,t} \mid z_{i,1:t-1}, x_{i,1:T}) = \ell(z_{i,t} \mid \theta(h_{i,t}, \Theta)) \quad (3-44)$$

El estado oculto $h_{i,t}$ se actualiza recursivamente mediante:

$$h_{i,t} = h(h_{i,t-1}, z_{i,t-1}, x_{i,t}, \Theta) \quad (3-45)$$

donde $h(\cdot)$ es una función implementada por una red LSTM multicapa con parámetros Θ .

Naturaleza Autorregresiva En cada paso temporal t , la red consume como entrada el valor observado del paso anterior $z_{i,t-1}$ junto con las covariables $x_{i,t}$ y el estado oculto

previo $h_{i,t-1}$. Durante el entrenamiento, todos los valores $z_{i,t}$ en el rango de predicción son conocidos. Durante la predicción, para $t \geq t_0$, los valores futuros se reemplazan por muestras $\tilde{z}_{i,t} \sim \ell(\cdot | \theta(h_{i,t}, \Theta))$ generadas por el propio modelo, que se retroalimentan para calcular el siguiente estado oculto mediante *muestreo ancestral*.

Modelado Probabilístico DeepAR no predice directamente valores futuros, sino los **parámetros de una distribución de probabilidad** $\theta(h_{i,t})$ sobre valores posibles. Para datos de valores reales, se emplea **verosimilitud Gaussiana**:

$$\ell_G(z | \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z - \mu)^2}{2\sigma^2}\right) \quad (3-46)$$

donde la media $\mu(h_{i,t})$ y desviación estándar $\sigma(h_{i,t})$ se obtienen mediante transformaciones de la salida de la red:

$$\mu(h_{i,t}) = w_\mu^T h_{i,t} + b_\mu, \quad \sigma(h_{i,t}) = \log(1 + \exp(w_\sigma^T h_{i,t} + b_\sigma)) \quad (3-47)$$

Los parámetros Θ del modelo se aprenden maximizando la log-verosimilitud:

$$\mathcal{L}(\Theta) = \sum_{i=1}^N \sum_{t=1}^T \log \ell(z_{i,t} | \theta(h_{i,t}, \Theta)) \quad (3-48)$$

Una ventaja fundamental es que el modelo es completamente observable durante el entrenamiento, no requiriendo técnicas de inferencia variacional o métodos de Monte Carlo para aproximar la función objetivo.

Manejo de Escalas Heterogéneas DeepAR introduce un mecanismo de escalamiento que normaliza las entradas y salidas autorregresivas por un factor de escala específico de cada serie ν_i :

$$\tilde{z}_{i,t} = \frac{z_{i,t}}{\nu_i}, \quad \nu_i = 1 + \frac{1}{t_0} \sum_{t=1}^{t_0} z_{i,t} \quad (3-49)$$

Los parámetros de la verosimilitud se ajustan correspondientemente:

$$\mu_{\text{escalado}} = \nu_i \cdot \mu(h_{i,t}), \quad \sigma_{\text{escalado}} = \nu_i \cdot \sigma(h_{i,t}) \quad (3-50)$$

Generación de Pronósticos Probabilísticos La generación de pronósticos se realiza mediante muestreo ancestral, produciendo B trayectorias completas $\{\tilde{z}_{i,t_0:T}^{(b)}\}_{b=1}^B$. El conjunto de trayectorias representa una muestra empírica de la distribución predictiva conjunta, preservando las correlaciones temporales aprendidas por el modelo.

Clasificación del Modelo DeepAR es un **modelo semi-paramétrico**. La componente paramétrica reside en la arquitectura LSTM con parámetros Θ que deben estimarse, y en la elección de la familia de distribuciones de verosimilitud (Gaussiana, Binomial Negativa). La componente no paramétrica emerge del muestreo ancestral, que genera distribuciones predictivas empíricas sin asumir formas funcionales rígidas para la distribución conjunta multi-paso.

De la Teoría a la Práctica La implementación de DeepAR para series temporales univariadas desarrollada en esta investigación traduce el marco teórico autorregresivo a un sistema predictivo concreto mediante adaptaciones específicas al contexto de este estudio.

Adaptaciones Principales (1) **Simplificación de Covariables:** A diferencia del trabajo original de Salinas et al. (2020) que asume múltiples covariables $x_{i,t}$, esta implementación **no utiliza covariables externas**. La arquitectura se reduce a su forma puramente autorregresiva:

$$h_{i,t} = h(h_{i,t-1}, z_{i,t-1}, \Theta) \quad (3-51)$$

Esta simplificación elimina la dependencia de información auxiliar, centrando la comparación en la capacidad de extraer patrones de la historia temporal intrínseca.

(2) **Construcción de Ventanas:** Dado que cada configuración genera una única serie de longitud $n = 252$, se emplea *windowing* deslizante para generar múltiples instancias de entrenamiento. Se construyen pares $(X^{(k)}, y^{(k)})$ mediante:

$$X^{(k)} = [y_k, \dots, y_{k+p-1}], \quad y^{(k)} = y_{k+p} \quad (3-52)$$

para $k = 1, \dots, n_{\text{train}} - p$, generando aproximadamente $n_{\text{train}} - p$ instancias de entrenamiento.

(3) Normalización Z-Score: En lugar del escalamiento por ν_i del original, se emplea normalización Z-score completa:

$$\tilde{z}_t = \frac{z_t - \mu_{\text{train}}}{\sigma_{\text{train}} + \epsilon} \quad (3-53)$$

garantizando que las entradas a la LSTM tengan media cero y varianza unitaria. Las predicciones se des-normalizan mediante la transformación inversa.

(4) Early Stopping: Se reserva el 20% final de las instancias generadas como conjunto de validación interno. El entrenamiento se detiene si la pérdida de validación no mejora durante un número de épocas de paciencia (típicamente 5), reteniendo los parámetros correspondientes a la época con menor pérdida.

(5) Protocolo de Congelamiento: Para evitar *data leakage* en la evaluación rolling window, el modelo se entrena una única vez sobre $n_{\text{train}} = 200$ observaciones. Los parámetros de normalización μ_{frozen} , σ_{frozen} y los pesos de la red Θ_{frozen} se congelan completamente. Durante la fase de evaluación rolling, para cada paso $t = 1, \dots, 12$, se normalizan los últimos p valores observados usando parámetros congelados, se calcula la predicción con Θ_{frozen} sin re-entrenamiento, y se des-normalizan las muestras predictivas. Este protocolo previene contaminación de información futura y garantiza validez estadística.

La Tabla 3-7 resume las diferencias metodológicas entre la implementación original y la adaptación desarrollada.

Optimización, Parámetros e Hiperparámetros

Hiperparámetros Arquitectónicos `hidden_size` (h): Dimensionalidad del estado oculto de cada celda LSTM. Controla la capacidad representacional del modelo.

`n_lags` (p): Número de valores pasados utilizados como entrada autorregresiva. Similar al orden en modelos AR(p).

`num_layers` (L): Número de capas LSTM apiladas. Valores mayores incrementan el número de parámetros: $\Theta \propto L \times h^2$.

`epochs` (E): Número máximo de pasadas completas sobre el conjunto de entrenamiento. El early stopping típicamente detiene antes de alcanzar este máximo.

Aspecto	DeepAR Original	Implementación (Esta Tesis)
Contexto de aplicación	Miles de series relacionadas	Serie temporal única (windowing local)
Covariables	Múltiples features temporales	Sin covariables externas
Verosimilitud	Gaussiana y Binomial Negativa	Gaussiana únicamente
Normalización	Escalamiento por ν_i	Z-score completo
Muestreo de entrenamiento	Ponderado por velocidad	Uniforme sobre ventanas
Regularización	No especificada	Early stopping con validación interna
Protocolo de evaluación	Modelo único para todas las series	Congelamiento total para rolling forecast
Framework	MXNet	PyTorch

Tabla 3-7.: Comparativa: Teoría vs. Implementación de DeepAR

lr (learning rate): Controla el tamaño del paso en la actualización de parámetros mediante el optimizador Adam (Kingma and Ba 2015).

Espacio de Búsqueda La optimización de hiperparámetros explora dos configuraciones estratégicamente diseñadas:

Configuración Ligera:

- `hidden_size=20, n_lags=10, num_layers=1, epochs=25, lr=0.01`

Configuración Profunda:

- `hidden_size=32, n_lags=15, num_layers=2, epochs=30, lr=0.005`

Protocolo de Optimización La selección entre ambas configuraciones se realiza minimizando **ECRPS** sobre un conjunto de calibración temporal de 40 observaciones posteriores al entrenamiento. Formalmente:

$$(h^*, n_{\text{lags}}^*, n_{\text{layers}}^*, \text{epochs}^*, \text{lr}^*) = \arg \min_{(h, n_{\text{lags}}, n_{\text{layers}}, \text{epochs}, \text{lr}) \in \mathcal{H}} \frac{1}{N_{\text{cal}}} \sum_{i=1}^{N_{\text{cal}}} \text{CRPS}(\hat{F}_i, y_i) \quad (3-54)$$

donde \mathcal{H} es el conjunto de configuraciones candidatas y \hat{F}_i es la distribución predictiva

empírica generada por muestreo ancestral.

Congelamiento de Hiperparámetros Una vez optimizada, la configuración seleccionada se congela:

1. Estima los parámetros de normalización: $\mu_{\text{frozen}}, \sigma_{\text{frozen}}$
2. Entrena la red LSTM hasta convergencia (early stopping)
3. Almacena los pesos de la red: Θ_{frozen}
4. Establece el flag

Durante toda la evaluación rolling window posterior, el método verifica este flag: si es `True`, omite completamente el entrenamiento y procede directamente a generar predicciones con el modelo existente. Este protocolo garantiza que no haya fuga de información futura y que las métricas de desempeño reflejen la capacidad predictiva genuina del modelo.

3.4.2. Modelos adaptados de la literatura

Least Squares Prediction Machine with Weighted Residuals (LSPMW)

Explicación Teórica del Modelo El *Least Squares Prediction Machine with Weighted Residuals* (LSPMW) constituye una extensión del LSPM diseñada para contextos donde el supuesto de intercambiabilidad se viola por deriva distributiva (*distribution drift*) o no estacionaridad. Esta variante se fundamenta en los desarrollos de Barber et al. (2023) sobre predicción conformal no intercambiable.

Fundamento Teórico: Cuantiles Ponderados Barber et al. (2023) demuestran que cuando la intercambiabilidad falla, la pérdida de cobertura (*coverage gap*) de un predictor conformal puede acotarse mediante la distancia de variación total entre distribuciones. Para mitigar este problema en presencia de deriva temporal, proponen sustituir la distribución empírica uniforme por una distribución empírica ponderada.

Formalmente, dado un conjunto de valores críticos (residuos conformales) $\{C_1, \dots, C_{n-1}\}$ ordenados cronológicamente, se asignan pesos temporales no uniformes w_i que priorizan

observaciones recientes. La función de distribución empírica ponderada se define como:

$$\hat{F}_n(y) = \sum_{i=1}^{n-1} \tilde{w}_i \cdot \mathbb{1}(C_i \leq y) \quad (3-55)$$

donde $\tilde{w}_i = w_i / \sum_{j=1}^{n-1} w_j$ son los pesos normalizados.

Esquema de Decaimiento Geométrico Para derivar gradual, Barber et al. (2023) recomiendan un esquema de decaimiento geométrico:

$$w_i = \rho^{n-1-i}, \quad \rho \in (0, 1) \quad (3-56)$$

donde i indexa los residuos en orden cronológico (de más antiguo a más reciente). El hiperparámetro ρ controla la tasa de olvido:

- $\rho \rightarrow 1$: Convergencia al LSPM uniforme (memoria larga)
- $\rho \rightarrow 0$: Concentración extrema en el pasado inmediato (memoria corta)

El peso efectivo del i -ésimo residuo decrece exponencialmente conforme retrocedemos en el tiempo, otorgando a la observación más reciente ($i = n - 1$) peso $\rho^0 = 1$, y a la más antigua ($i = 1$) peso ρ^{n-2} .

Cuantiles Ponderados Para un nivel de cobertura α , el cuantil $(1 - \alpha)$ de la distribución ponderada se obtiene mediante:

$$q_{1-\alpha} = \inf \left\{ y : \sum_{i:C_i \leq y} \tilde{w}_i \geq 1 - \alpha \right\} \quad (3-57)$$

Este cuantil ponderado adapta la región de predicción conforme la distribución subyacente cambia en el tiempo.

Propiedades de Cobertura Barber et al. (2023) establecen que bajo deriva Lipschitz-continua con constante L , el error de cobertura del predictor ponderado satisface:

$$\left| \mathbb{P}(Y_{n+1} \in \hat{C}_{1-\alpha}) - (1 - \alpha) \right| \leq O \left(\frac{L}{\rho} + \frac{1}{\sqrt{n_{\text{eff}}}} \right) \quad (3-58)$$

donde $n_{\text{eff}} = (\sum_i \tilde{w}_i^2)^{-1}$ es el tamaño de muestra efectivo. Esta cota revela el trade-off: ρ pequeño reduce sesgo por deriva (L/ρ disminuye) pero aumenta varianza (n_{eff} disminuye).

Clasificación del Modelo El LSPMW es un método **no paramétrico adaptativo**. Mantiene la propiedad distribution-free del LSPM (no asume forma distribucional de errores) pero introduce un mecanismo de ponderación que adapta la distribución predictiva a cambios temporales sin modelar explícitamente la deriva.

De la Teoría a la Práctica La implementación desarrollada traduce la teoría de Barber et al. (2023) mediante un esquema de **muestreo ponderado adaptativo** que recalcula dinámicamente los residuos conformales:

Adaptación 1: Muestreo Ponderado vs. Expansión por Replicación A diferencia de métodos que pre-computan una distribución expandida, la implementación emplea *muestreo estratificado* en tiempo real. Para generar una distribución predictiva de tamaño $N = 1000$, se ejecuta:

$$\tilde{R}_j \sim \text{Categorical}(\{C_1, \dots, C_n\}, \{\tilde{w}_1, \dots, \tilde{w}_n\}), \quad j = 1, \dots, N \quad (3-59)$$

donde cada muestra \tilde{R}_j se extrae de los valores críticos $\{C_i\}$ con probabilidades proporcionales a los pesos normalizados $\tilde{w}_i = w_i / \sum_k w_k$. Este enfoque es matemáticamente equivalente a muestrear de la distribución empírica ponderada $\hat{F}_n(y)$.

Adaptación 2: Recálculo Dinámico de Residuos Contrario a esquemas de conglomeramiento completo, la implementación mantiene únicamente los hiperparámetros (ρ^* , p^*) fijos tras calibración. En cada paso del rolling forecast:

1. Se recalculan valores críticos $\{C_1, \dots, C_m\}$ usando la ventana temporal actual.

2. Se actualizan pesos $w_i = (\rho^*)^{m-1-i}$ para $i = 1, \dots, m$ (ajustados al tamaño m de la ventana).
3. Se normalizan: $\tilde{w}_i = w_i / \sum_j w_j$.
4. Se genera la distribución predictiva mediante muestreo con reemplazo usando estos pesos actualizados.

Este diseño permite que el modelo se adapte continuamente a derivas distributivas, pues los residuos más recientes (calculados con datos actuales) reciben mayor ponderación automáticamente.

Adaptación 3: Protocolo de Congelamiento Parcial El método almacena únicamente:

- `_frozen_rho`: Valor óptimo ρ^* seleccionado por validación
- `_frozen_n_lags`: Orden autorregresivo p^* heredado de LSPM

No se congelan los residuos conformales ni sus pesos asociados, permitiendo que la distribución predictiva refleje el estado más reciente de la serie temporal. Esta filosofía difiere del LSPM estándar, donde congelar la distribución completa es apropiado bajo intercambiabilidad, pero resulta inadecuado bajo deriva continua.

Aspecto	Teoría (Barber et al.)	Implementación
Salida	Cuantiles ponderados específicos	Distribución completa (1000 muestras)
Método de generación	Definición abstracta $\hat{F}_n(y)$	Muestreo ponderado con reemplazo
Actualización	Marco teórico general	Recálculo dinámico de residuos
Residuos	Valores críticos conceptuales	LSPM studentizados recalculados
Optimización	ρ teórico o fijo	Búsqueda por validación (ECRPS)
Congelamiento	No especificado	Solo hiperparámetros (ρ^*, p^*)

Tabla 3-8.: Comparativa: Teoría vs. Implementación del LSPMW

Optimización, Parámetros e Hiperparámetros

Hiperparámetro Principal: rho (ρ) Controla la tasa de decaimiento temporal de los pesos. Durante validación se evalúan valores en el rango [0.90, 0.99]:

- $\rho = 0.90$: Adaptación rápida, memoria efectiva ≈ 10 observaciones

- $\rho = 0.95$: Balance intermedio, memoria efectiva ≈ 20 observaciones
- $\rho = 0.99$: Adaptación lenta, cercano al LSPM uniforme

El tamaño efectivo de muestra bajo decaimiento geométrico es:

$$n_{\text{eff}} = \frac{1}{\sum_{i=1}^{n-1} \tilde{w}_i^2} \approx \frac{1 - \rho}{1 - \rho^n} \quad (3-60)$$

Métrica de Optimización La selección de ρ^* se realiza minimizando el **ECRPS** (véase 2.2.3) sobre el conjunto de validación. Para cada candidato ρ , se genera la distribución ponderada en cada tiempo de validación mediante muestreo estratificado y se evalúa la calibración predictiva.

Protocolo de Congelamiento Parcial El método ejecuta:

1. Congela p^* (número de lags) heredado de la clase base LSPM.
2. Identifica y almacena ρ^* óptimo del proceso de validación.
3. Activa la bandera para uso de hiperparámetros congelados.
4. *No congela* valores críticos ni pesos, preservando adaptabilidad temporal.

Durante la fase de prueba rolling, el método `fit_predict()`:

1. Recalcula residuos conformales $\{C_1, \dots, C_m\}$ usando p^* sobre la ventana actual.
2. Computa pesos actualizados: $w_i = (\rho^*)^{m-1-i}$ para $i = 1, \dots, m$.
3. Normaliza: $\tilde{w}_i = w_i / \sum_j w_j$.
4. Genera 1000 muestras mediante `np.random.choice` con probabilidades $\{\tilde{w}_i\}$.

Mondrian Conformal Predictive System (MCPS)

Explicación Teórica del Modelo El *Mondrian Conformal Predictive System* (MCPS), formalizado por Boström, Johansson, and Löfström (2021), constituye una extensión localmente adaptativa del Sistema Predictivo Conformal estándar (SCPS). A diferencia del SCPS, que asume homogeneidad en la distribución de errores sobre todo el espacio de entrada, el MCPS reconoce que la incertidumbre predictiva varía significativamente según

el régimen operativo del modelo. Este enfoque resulta especialmente relevante en aplicaciones reales donde, por ejemplo, predicciones en rangos bajos del modelo pueden exhibir patrones de error distintos a predicciones en rangos altos.

Fundamento: Particionamiento Mondrian La innovación central radica en la estrategia de **particionamiento Mondrian** (Vovk, Gammerman, and Shafer 2005), que divide el conjunto de calibración \mathcal{D}_c en subconjuntos disjuntos basándose en características compartidas de las predicciones. Sea $h : \mathcal{X} \rightarrow \mathbb{R}$ un modelo de regresión entrenado, y $B \in \mathbb{N}$ el número de bins. Se define una partición:

$$\mathcal{D}_c = \bigcup_{\kappa=1}^B \mathcal{D}_c^\kappa, \quad \mathcal{D}_c^\kappa \cap \mathcal{D}_c^{\kappa'} = \emptyset \text{ para } \kappa \neq \kappa' \quad (3-61)$$

Cada subconjunto \mathcal{D}_c^κ agrupa instancias (x_j, y_j) cuyas predicciones $h(x_j)$ pertenecen al mismo rango, construido mediante cuantiles empíricos:

$$\mathcal{D}_c^\kappa = \left\{ (x_j, y_j) \in \mathcal{D}_c : q_{\frac{\kappa-1}{B}} \leq h(x_j) < q_{\frac{\kappa}{B}} \right\} \quad (3-62)$$

donde q_p denota el cuantil p de las predicciones $\{h(x_j)\}_{j=1}^{N_c}$. Esta estrategia captura automáticamente heterogeneidad: instancias con predicciones similares comparten probablemente patrones de error similares.

Cálculo Localizado de Scores Conformales Para una instancia de prueba x con predicción $h(x)$, se determina su bin correspondiente κ^* :

$$\kappa^* = \arg \min_{\kappa} \left\{ \kappa : h(x) < q_{\frac{\kappa}{B}} \right\} \quad (3-63)$$

Los **calibration scores** se calculan únicamente sobre el subconjunto local $\mathcal{D}_c^{\kappa^*}$:

$$C_j^{\kappa^*} = h(x) + (y_j - h(x_j)), \quad \forall (x_j, y_j) \in \mathcal{D}_c^{\kappa^*} \quad (3-64)$$

Esta formulación es algebraicamente idéntica al SCPS, pero la diferencia crítica reside en que los residuos históricos provienen exclusivamente de casos con comportamiento predic-

tivo similar. Matemáticamente, se estima la distribución condicional $P(\epsilon \mid h(x) \in \text{Bin}_\kappa)$ localmente, en lugar de globalmente como en SCPS donde se asume $P(\epsilon \mid h(x)) = P(\epsilon)$.

Construcción de la Distribución Predictiva Según Vovk, Gammerman, and Shafer (2022), el conjunto de scores ordenados $C_{(1)}^{\kappa^*} < C_{(2)}^{\kappa^*} < \dots < C_{(N_c^{\kappa^*})}^{\kappa^*}$ define una distribución empírica que aproxima la distribución predictiva verdadera. Para cualquier función medible g :

$$\mathbb{E}[g(Y) \mid x \in \text{Bin}_{\kappa^*}] \approx \frac{1}{N_c^{\kappa^*}} \sum_{j=1}^{N_c^{\kappa^*}} g(C_j^{\kappa^*}) \quad (3-65)$$

con error que converge a cero cuando $N_c^{\kappa^*} \rightarrow \infty$. Esto implica que cualquier estadístico (media, varianza, cuantiles) puede calcularse directamente sobre los scores sin reconstruir una CDF completa.

Garantías de Cobertura Local Boström, Johansson, and Löfström (2021) demuestran que, bajo intercambiabilidad dentro de cada bin, se logra **cobertura condicional válida** en cada estrato. Para cualquier nivel α :

$$\mathbb{P}\left(Y \in \left[\hat{F}_\kappa^{-1}(\alpha/2 \mid x), \hat{F}_\kappa^{-1}(1 - \alpha/2 \mid x)\right] \mid x \in \text{Bin}_\kappa\right) \geq 1 - \alpha \quad (3-66)$$

Las bandas de predicción se **ajustan automáticamente**: regiones donde el modelo es confiable producen intervalos estrechos; regiones con alta variabilidad generan intervalos amplios. Si denotamos $W_\kappa(x)$ como el ancho del intervalo:

$$W_{\kappa_1}(x_1) \neq W_{\kappa_2}(x_2) \quad \text{si} \quad \text{Var}(\epsilon \mid x \in \text{Bin}_{\kappa_1}) \neq \text{Var}(\epsilon \mid x \in \text{Bin}_{\kappa_2}) \quad (3-67)$$

Esta propiedad contrasta con el SCPS, donde $W(x) \approx$ constante para todo x , resultando en sobre-cobertura en regiones de baja incertidumbre o sub-cobertura en regiones de alta incertidumbre.

Naturaleza No Paramétrica El MCPS es un modelo **no paramétrico**. No asume forma funcional específica para la distribución de errores ni para la relación entre predictores y respuesta. La distribución predictiva se construye enteramente a partir de datos

empíricos mediante el conjunto de scores conformales, sin parámetros poblacionales a estimar. El único modelo subyacente es el regressor base $h(x)$ (que puede ser paramétrico o no), pero la construcción de intervalos conformales es completamente libre de distribución.

De la Teoría a la Práctica La implementación del MCPS para series temporales autoregresivas representa una contribución metodológica poco explorada, traduciendo el framework teórico (originalmente diseñado para datos independientes en logística (Ye, Hijazi, and Van Hentenryck 2025)) hacia contextos con dependencias temporales.

Adaptaciones Principales (1) **Construcción Dinámica de Features:** Mientras que Ye, Hijazi, and Van Hentenryck (2025) asumen features pre-existentes x_i observables (ubicación, peso, transportista), en series temporales los “objetos” se construyen dinámicamente como ventanas deslizantes de p rezagos:

$$x_t = [y_{t-p}, y_{t-p+1}, \dots, y_{t-1}] \in \mathbb{R}^p \quad (3-68)$$

Esta transformación convierte la serie univariada en una matriz autoregresiva, introduciendo dependencias inherentes entre filas. La validez se preserva bajo condiciones de *mixing débil* (Yu 1994), donde observaciones suficientemente separadas son “casi independientes”.

(2) **Binning Adaptativo:** En series con baja variabilidad, las predicciones pueden concentrarse en rangos estrechos. Se emplea binning tolerante a duplicados, donde el número efectivo de bins es:

$$B_{\text{efectivo}} = |\{q_{\frac{\kappa}{B}} : \kappa = 1, \dots, B - 1\}| \leq B \quad (3-69)$$

Si el binning falla completamente (predicciones idénticas), el sistema degradará a SCPS global automáticamente.

(3) **Representación Discreta:** En lugar de generar una CDF continua con suavizado τ (como en Crepes (Boström 2022)), se retornan directamente los scores $\{C_j^{\kappa^*}\}$. Esta representación es computacionalmente eficiente para calcular ECRPS:

(4) **Fallback Jerárquico:** Si el bin localizado κ^* contiene menos de 5 observaciones, se usa el conjunto completo:

$$\mathcal{D}_c^{\text{efectivo}} = \begin{cases} \mathcal{D}_c^{\kappa^*} & \text{si } N_c^{\kappa^*} \geq 5 \\ \mathcal{D}_c & \text{si } N_c^{\kappa^*} < 5 \end{cases} \quad (3-70)$$

Esta heurística, no especificada en Boström, Johansson, and Löfström (2021), previene intervalos erráticamente anchos por tamaño de muestra insuficiente.

Aspecto	Teoría (Ye et al., 2025)	Implementación (Esta Tesis)
Dominio	Logística (órdenes independientes)	Series temporales autorregresivas
Features	Pre-existentes observables	Dinámicas (ventanas deslizantes)
Librería conformal	Crepes (Boström 2022)	Implementación directa de ecuaciones
Representación CPD	CDF continua con suavizado	Distribución empírica discreta exacta
Binning robusto	τ	Fusión automática de bins colapsados
Fallback a SCPS	No mencionado	Automático si $N_c^{\kappa} < 5$
Horizonte	Batch (predicciones simultáneas)	Secuencial (rolling forecast)

Tabla 3-9.: Comparativa: Teoría vs. Implementación del MCPS

Optimización, Parámetros e Hiperparámetros

Hiperparámetros Primarios Dado el costo computacional de la validación cruzada completa y las restricciones del conjunto de datos, se consideran únicamente dos configuraciones de hiperparámetros:

1. **Configuración conservadora:** $p = 10$ rezagos, $B = 8$ bins
2. **Configuración flexible:** $p = 15$ rezagos, $B = 15$ bins

Estos hiperparámetros controlan aspectos fundamentales del modelo:

n_lags (p): Define el orden del modelo autorregresivo, determinando cuánta memoria temporal incorpora el predictor. La configuración conservadora utiliza $p = 10$ mientras que la flexible extiende a $p = 15$ para capturar dinámicas de más largo plazo.

n_bins (B): Número de particiones Mondrian que controla la adaptabilidad local del ajuste de conformidad. Según Boström, Johansson, and Löfström (2021), valores muy pequeños ($B \leq 3$) degeneran el método a SCPS (sin adaptación local), mientras que valores excesivos ($B \geq 20$) fragmentan la calibración por insuficiencia muestral. La configuración conservadora emplea $B = 8$ bins (tamaño esperado por bin: $\mathbb{E}[N_c^\kappa] \approx N_c/8$), mientras que la flexible utiliza $B = 15$ para mayor adaptabilidad espacial.

Optimización de Hiperparámetros Estrategia de congelamiento: El método definido entrena el modelo base XGBoost una sola vez sobre los datos de entrenamiento completos, calculando los artefactos necesarios:

- Predicciones de calibración $\{h(x_j)\}_{j \in \mathcal{D}_c}$
- Valores observados de calibración $\{y_j\}_{j \in \mathcal{D}_c}$
- Bordes de bins $\{q_{\kappa/B}\}_{\kappa=0}^B$

Estos artefactos se reutilizan en todas las predicciones posteriores sin reentrenamiento, garantizando eficiencia computacional en rolling forecasts.

Métrica de optimización: Los hiperparámetros (p, B) se optimizan minimizando el ECRPS sobre un conjunto de validación:

$$(p^*, B^*) = \arg \min_{(p, B)} \frac{1}{N_{\text{val}}} \sum_{i=1}^{N_{\text{val}}} \text{CRPS}(\mathcal{C}_i^{\kappa^*}, y_i) \quad (3-71)$$

Esta optimización ocurre en la fase inicial; una vez congelados, los hiperparámetros permanecen fijos durante todo el horizonte de predicción.

Parámetros del Modelo Base El regressor autorregresivo $h : \mathbb{R}^p \rightarrow \mathbb{R}$ se implementa mediante XGBoost (T. Chen and Guestrin 2016):

- **n_estimators:** 50 (número de árboles)
- **max_depth:** 3 (previene sobreajuste, actúa como GAM)
- **learning_rate:** 0.1 (convergencia estable)
- **objective:** reg:squarederror (regresión por mínimos cuadrados)

Esta parametrización conservadora previene captura de ruido, crítico cuando p es grande relativo a N_{train} .

Balance Fundamental La selección conjunta define un trade-off:

$$\text{Calidad de } h(x) \propto (1 - \text{test_size}) \cdot f(p) \quad (3-72)$$

$$\text{Precisión de CPD} \propto \text{test_size} \cdot \frac{N_c}{B} \quad (3-73)$$

Para series típicas con $T \approx 1000$, la configuración ($p = 10$, $B = 10$, $\text{test_size} = 0.25$) resulta en ~ 25 scores por bin, suficiente para estimación robusta según Rob J. Hyndman and Fan (1996).

Adaptive Volatility Mondrian Conformal Predictive System (AV-MCPS)

Explicación Teórica del Modelo El *Adaptive Volatility Mondrian Conformal Predictive System* (AV-MCPS) constituye una extensión metodológica del MCPS estándar desarrollada específicamente para esta investigación, representando una de las contribuciones originales más significativas de la tesis. Mientras que el MCPS de Ye, Hijazi, and Van Hentenryck (2025) partitiona el espacio de calibración únicamente según predicciones puntuales $h(x)$, el AV-MCPS introduce una **estratificación bidimensional** que incorpora explícitamente la volatilidad local como segunda dimensión de heterogeneidad.

Motivación: Límites del Particionamiento Unidimensional El MCPS estándar asume implícitamente homogeneidad de varianza dentro de cada bin de predicción. Formalmente, si \mathcal{D}_c^κ denota el subconjunto con predicciones en $[q_{\kappa/B}, q_{(\kappa+1)/B})$:

$$\text{Var}(\epsilon_i \mid h(x_i) \in \mathcal{D}_c^\kappa) \approx \text{constante} \quad \forall i \in \mathcal{D}_c^\kappa \quad (3-74)$$

Esta suposición se viola frecuentemente en series temporales con **heterocedasticidad condicional**, donde la volatilidad de errores varía sistemáticamente: $\text{Var}(\epsilon_t) = \sigma_t^2 \neq \text{constante}$. Dos observaciones con predicciones similares $h(x_i) \approx h(x_j)$ pueden experimentar errores de magnitudes radicalmente diferentes si provienen de regímenes de volatilidad distintos.

Particionamiento Bidimensional El AV-MCPS propone una partición conjunta basada en dos características:

1. **Predicción puntual** $h(x_i)$: Captura el nivel esperado de la variable objetivo
2. **Volatilidad local** σ_i : Mide la variabilidad reciente mediante ventana rodante de longitud w :

$$\sigma_i = \sqrt{\frac{1}{w-1} \sum_{k=i-w}^{i-1} (y_k - \bar{y}_{i,w})^2} \quad (3-75)$$

El conjunto de calibración se partitiona en una grilla bidimensional:

$$\mathcal{D}_c = \bigcup_{\kappa=1}^{B_{\text{pred}}} \bigcup_{\lambda=1}^{B_{\text{vol}}} \mathcal{D}_c^{(\kappa, \lambda)} \quad (3-76)$$

donde cada celda agrupa observaciones que satisfacen simultáneamente:

$$\mathcal{D}_c^{(\kappa, \lambda)} = \left\{ (x_i, y_i) \in \mathcal{D}_c : \begin{array}{l} q_{\text{pred}}^{\kappa-1} \leq h(x_i) < q_{\text{pred}}^\kappa \\ \text{y } q_{\text{vol}}^{\lambda-1} \leq \sigma_i < q_{\text{vol}}^\lambda \end{array} \right\} \quad (3-77)$$

Esta estratificación genera $B_{\text{pred}} \times B_{\text{vol}}$ celdas, cada una representando un régimen específico de (nivel, volatilidad).

Cálculo Localizado de Scores Para un punto de prueba x_{test} , el procedimiento es:

1. Calcular predicción $h(x_{\text{test}})$ y volatilidad local σ_{test}
2. Determinar la celda correspondiente (κ^*, λ^*) :

$$\kappa^* = \arg \min_{\kappa} \{ \kappa : h(x_{\text{test}}) < q_{\text{pred}}^\kappa \} \quad (3-78)$$

$$\lambda^* = \arg \min_{\lambda} \{ \lambda : \sigma_{\text{test}} < q_{\text{vol}}^\lambda \} \quad (3-79)$$

3. Calcular scores localizados sobre $\mathcal{D}_c^{(\kappa^*, \lambda^*)}$:

$$C_i^{(\kappa^*, \lambda^*)} = h(x_{\text{test}}) + (y_i - h(x_i)), \quad \forall (x_i, y_i) \in \mathcal{D}_c^{(\kappa^*, \lambda^*)} \quad (3-80)$$

Garantías de Cobertura Bidimensional Bajo intercambiabilidad condicional dentro de cada celda, se preservan las garantías conformales localmente:

$$\mathbb{P} \left(Y \in \hat{C}_\alpha(x) \mid x \in \text{Bin}_\kappa^{\text{pred}}, \sigma(x) \in \text{Bin}_\lambda^{\text{vol}} \right) \geq 1 - \alpha \quad (3-81)$$

Esta garantía es más fuerte que la del MCPS, ya que condiciona sobre una partición más fina. Al controlar simultáneamente por nivel y volatilidad, el AV-MCPS logra **adaptabilidad condicional mejorada**.

Trade-off: Resolución vs. Tamaño de Muestra El número esperado de observaciones por celda es:

$$\mathbb{E} [N_c^{(\kappa, \lambda)}] = \frac{N_c}{B_{\text{pred}} \times B_{\text{vol}}} \quad (3-82)$$

Aumentar resolución mejora localización pero reduce tamaño de muestra por celda. El AV-MCPS implementa **fallback jerárquico**:

$$\mathcal{D}_c^{\text{efectivo}} = \begin{cases} \mathcal{D}_c^{(\kappa^*, \lambda^*)} & \text{si } N_c^{(\kappa^*, \lambda^*)} \geq 5 \\ \mathcal{D}_c^{(\kappa^*, \cdot)} & \text{si } N_c^{(\kappa^*, \lambda^*)} < 5 \text{ y } N_c^{(\kappa^*, \cdot)} \geq 5 \\ \mathcal{D}_c & \text{en otro caso} \end{cases} \quad (3-83)$$

donde $\mathcal{D}_c^{(\kappa^*, \cdot)} = \bigcup_{\lambda=1}^{B_{\text{vol}}} \mathcal{D}_c^{(\kappa^*, \lambda)}$ representa el bin unidimensional basado solo en predicción.

Naturaleza No Paramétrica El AV-MCPS es un modelo **no paramétrico**. No asume forma funcional para la distribución de errores ni para la relación entre predictores y respuesta. La estratificación bidimensional es completamente libre de distribución, construyéndose enteramente a partir de cuantiles empíricos. La distribución predictiva se genera directamente desde scores conformales sin parámetros poblacionales.

De la Teoría a la Práctica La implementación del AV-MCPS para series temporales traduce el marco teórico bidimensional en un sistema operativo mediante decisiones de diseño específicas.

Adaptaciones Principales (1) Estimación Rolling de Volatilidad: Se emplea desviación estándar rolling con ventana fija:

$$\hat{\sigma}_t = \sqrt{\frac{1}{w-1} \sum_{k=t-w}^{t-1} (y_k - \bar{y}_t)^2} \quad (3-84)$$

Justificación: simplicidad computacional, robustez a shocks transitorios, e interpretabilidad directa. Para las primeras $w-1$ observaciones se aplica *backfilling*:

$$\hat{\sigma}_t = \sqrt{\frac{1}{t-2} \sum_{k=1}^{t-1} (y_k - \bar{y}_{1:t-1})^2} \quad \text{para } t < w \quad (3-85)$$

(2) Binning Adaptativo Robusto: Las series de volatilidad exhiben distribuciones asimétricas con colas pesadas. Se fusionan automáticamente bins con fronteras colapsadas:

$$B_{\text{vol}}^{\text{efectivo}} = |\{q_{\text{vol}}^\lambda : \lambda = 1, \dots, B_{\text{vol}} - 1\}| \leq B_{\text{vol}} \quad (3-86)$$

Si el binning falla en alguna dimensión, el sistema degradada a MCPS unidimensional o SCPS global.

(3) Representación Discreta de Scores: Consistente con LSPM y MCPS, se retornan directamente los scores sin transformación:

$$\mathcal{C}^{(\kappa^*, \lambda^*)} = \left\{ C_i^{(\kappa^*, \lambda^*)} \right\}_{i=1}^{N_c^{(\kappa^*, \lambda^*)}} \quad (3-87)$$

Esta representación empírica permite cálculo eficiente de CRPS sin reconstruir CDF continua.

(4) Protocolo de Congelamiento Bidimensional: El método fija simultáneamente:

- Parámetros del regressor XGBoost (entrenado una sola vez)
- Bordes de bins de predicción $\{q_{\text{pred}}^\kappa\}_{\kappa=0}^{B_{\text{pred}}}$
- Bordes de bins de volatilidad $\{q_{\text{vol}}^\lambda\}_{\lambda=0}^{B_{\text{vol}}}$
- Artefactos de calibración: $\{h(x_i)\}, \{y_i\}, \{\sigma_i\}$

Durante evaluación rolling, para cada punto t se extraen rezagos, se calcula $h(x_t)$ con modelo congelado, se estima σ_t con ventana actual, y se asigna a celda usando bordes congelados. Este protocolo previene data leakage y garantiza validez estadística.

Aspecto	MCPS	AV-MCPS
Dimensiones de partición	Unidimensional (predicción)	Bidimensional (predicción + volatilidad)
Número de bins	B	$B_{\text{pred}} \times B_{\text{vol}}$
Tamaño esperado de celda	N_c/B	$N_c/(B_{\text{pred}} \times B_{\text{vol}})$
Captura de heterocedasticidad	Indirecta (via niveles)	Explícita (via volatilidad local)
Complejidad computacional	$O(\log B)$	$O(\log B_{\text{pred}} + \log B_{\text{vol}})$
Estrategia de fallback	Degradar a SCPS si $N_c^\kappa < 5$	Jerárquica: 2D \rightarrow 1D \rightarrow SCPS
Hiperparámetros adicionales	Ninguno	<code>volatility_window</code> , B_{vol}
Casos de uso óptimos	Heterogeneidad por nivel	Series con volatilidad cambiante

Tabla 3-10.: Comparativa: MCPS vs. AV-MCPS

Innovación Metodológica La contribución fundamental es reconocer que **la volatilidad local predice el error independientemente del nivel**. Si $e_i = |y_i - h(x_i)|$, la hipótesis subyacente es:

$$\mathbb{E}[e_i | h(x_i), \sigma_i] \neq \mathbb{E}[e_i | h(x_i)] \quad (3-88)$$

El AV-MCPS explota esta estructura mediante estratificación bidimensional, logrando distribuciones predictivas que se adaptan simultáneamente al *nivel* y al *régimen de incertidumbre*. Esta adaptabilidad dual representa una ventaja teórica significativa, especialmente en series con volatilidad time-varying como procesos financieros, climáticos o epidemiológicos.

Optimización, Parámetros e Hiperparámetros

Hiperparámetros Primarios **n_lags** (p): Orden del modelo autorregresivo.

n_pred_bins (B_{pred}): Resolución en dimensión de predicción.

n_vol_bins (B_{vol}): Resolución en dimensión de volatilidad.

Configuraciones Evaluadas El espacio de hiperparámetros se limita a dos configuraciones estratégicamente diseñadas:

Configuración Conservadora: ($p = 10, B_{\text{pred}} = 8, B_{\text{vol}} = 3$)

Configuración Agresiva: ($p = 15, B_{\text{pred}} = 10, B_{\text{vol}} = 5$)

Optimización de Hiperparámetros La selección entre ambas configuraciones se realiza minimizando **ECRPS** sobre un conjunto de validación temporal. Formalmente:

$$(p^*, B_{\text{pred}}^*, B_{\text{vol}}^*) = \arg \min_{(p, B_p, B_v) \in \mathcal{H}} \frac{1}{N_{\text{val}}} \sum_{i=1}^{N_{\text{val}}} \text{CRPS}(\mathcal{C}_i^{(\kappa^*, \lambda^*)}, y_i) \quad (3-89)$$

donde $\mathcal{H} = \{(10, 8, 3), (15, 10, 5)\}$ es el conjunto de configuraciones candidatas. Una vez optimizada, la configuración seleccionada se congela fijando simultáneamente el modelo base XGBoost, los bordes de bins bidimensionales, y los artefactos de calibración.

Parámetros del Modelo Base Idénticos a MCPS: XGBoost con 50 árboles, profundidad 3, learning rate 0.1, objetivo `reg:squarederror`. Esta parametrización conservadora es particularmente crítica en AV-MCPS, donde el conjunto de entrenamiento puede ser más pequeño debido a la fragmentación bidimensional del conjunto de calibración.

Autoregressive Exponentially-weighted Polynomial Distribution (AREPD)

Explicación Teórica del Modelo El modelo *Autoregressive Exponentially-weighted Polynomial Distribution* (AREPD) representa una contribución metodológica original de esta investigación, desarrollada como una extensión híbrida que combina predicción ponderada temporalmente con expansión polinomial de características autorregresivas. A diferencia de métodos puramente conformales como LSPM que utilizan regresión lineal, AREPD introduce **transformaciones no lineales polinomiales** de las entradas autorregresivas.

Fundamento Matemático Para una serie temporal $\{Y_t\}_{t=1}^n$ con p rezagos y grado polinomial d , AREPD construye una matriz de diseño expandida $\mathbf{X} \in \mathbb{R}^{(n-p) \times (1+pd)}$:

$$\mathbf{X}_i = [1, Y_i, Y_i^2, \dots, Y_i^d, Y_{i+1}, Y_{i+1}^2, \dots, Y_{i+1}^d, \dots, Y_{i+p-1}] \quad (3-90)$$

para $i = 1, \dots, n-p$. Esta expansión permite capturar relaciones cuadráticas, cúbicas o de orden superior entre valores pasados y futuros sin recurrir a arquitecturas de aprendizaje profundo.

Ponderación Exponencial Temporal AREPD asigna pesos exponencialmente decrecientes a observaciones históricas:

$$w_i = \rho^{n-p-i}, \quad i = 1, \dots, n-p \quad (3-91)$$

donde $\rho \in (0, 1)$ es el parámetro de decaimiento. Los pesos se normalizan: $\tilde{w}_i = w_i / \sum_{j=1}^{n-p} w_j$. La observación más reciente recibe peso máximo, mientras que observaciones antiguas reciben pesos progresivamente menores. La vida media efectiva es:

$$\tau_{1/2} = \frac{\log(2)}{\log(1/\rho)} \quad (3-92)$$

Por ejemplo, con $\rho = 0.95$, $\tau_{1/2} \approx 13.5$ observaciones.

Estimación mediante Regresión Ridge Los coeficientes se estiman resolviendo el problema de regresión Ridge ponderada:

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n-p} \tilde{w}_i (y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\} \quad (3-93)$$

donde $\lambda > 0$ es el parámetro de regularización Ridge. La penalización Ridge es crítica por dos razones:

1. **Estabilidad numérica:** La expansión polinomial genera matrices casi singulares debido a alta correlación entre términos como Y_{t-1} y Y_{t-1}^2

2. **Prevención de sobreajuste:** Con $p \cdot d$ características, el modelo tiene alta capacidad expresiva que requiere regularización

La solución tiene forma cerrada:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W} \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y} \quad (3-94)$$

donde $\mathbf{W} = \text{diag}(\tilde{w}_1, \dots, \tilde{w}_{n-p})$ es la matriz diagonal de pesos.

Generación de Distribuciones Predictivas AREPD genera distribuciones predictivas mediante un enfoque **histórico-empírico**. Para predecir Y_{n+1} :

1. Construye vectores de características expandidos para todos los puntos históricos
2. Calcula predicciones puntuales históricas: $\hat{Y}_{\text{hist}} = \mathbf{X}_{\text{hist}} \hat{\boldsymbol{\beta}}$
3. Transforma a escala original: $\hat{Y}_{\text{hist}}^{\text{original}} = \hat{Y}_{\text{hist}} \cdot \sigma_{\text{train}} + \mu_{\text{train}}$
4. Utiliza $\{\hat{Y}_{\text{hist}}^{\text{original}}\}$ como muestra empírica de la distribución predictiva

Este enfoque difiere fundamentalmente de la predicción conformal estándar. En lugar de ajustar mediante residuos de calibración, AREPD construye una distribución empírica directamente desde predicciones históricas del modelo ajustado, asumiendo que bajo estacionariedad local, estas predicciones son representativas del futuro.

Garantías de Cobertura A diferencia de métodos conformales con garantías formales, AREPD no posee garantías teóricas de cobertura bajo intercambiabilidad. La distribución predictiva se construye asumiendo:

$$\hat{Y}_t \mid \mathcal{F}_{t-1} \stackrel{d}{\approx} \hat{Y}_s \mid \mathcal{F}_{s-1} \quad \forall t, s \in \{p+1, \dots, n\} \quad (3-95)$$

Esta suposición puede violarse en presencia de no estacionariedad fuerte, heterocedasticidad condicional, o eventos raros no observados durante el entrenamiento.

Clasificación del Modelo AREPD es un **modelo semi-paramétrico**. La componente paramétrica reside en la estructura de regresión Ridge con coeficientes $\hat{\boldsymbol{\beta}}$ que deben

estimarse. La componente no paramétrica emerge del uso de la distribución empírica de predicciones históricas sin asumir formas funcionales rígidas para la distribución predictiva.

Método	Espacio de características	Ponderación temporal	Distribución predictiva
LSPM	Lineal (rezagos)	Uniforme	Conformal (scores ajustados)
LSPMW	Lineal (rezagos)	Exponencial (ρ)	Conformal ponderada
AREPD	Polinomial (grado d)	Exponencial (ρ)	Histórico-empírica
DeepAR	No lineal (LSTM)	Uniforme	Muestreo ancestral

Tabla 3-11.: Comparativa Conceptual: AREPD vs Modelos Relacionados

Optimización, Parámetros e Hiperparámetros

Hiperparámetros Estructurales `n_lags` (p): Número de rezagos incluidos en la matriz de diseño. Con grado d , la dimensión del espacio de características es $1 + p \cdot d$. Valores mayores permiten capturar dependencias de largo plazo pero reducen el tamaño del conjunto de entrenamiento e incrementan riesgo de sobreajuste.

`poly_degree` (d): Grado máximo de la expansión polinomial.

- $d = 1$: Modelo puramente lineal (regresión autorregresiva Ridge estándar)
- $d = 2$: Incluye términos cuadráticos, captura efectos de amplificación moderados
- $d = 3$: Incluye términos cúbicos, alta expresividad pero riesgo significativo de sobreajuste

Número de parámetros: $1 + p \cdot d$. Para $(p, d) = (5, 2)$: 11 parámetros; para $(p, d) = (10, 3)$: 31 parámetros.

`rho` (ρ): Parámetro de decaimiento exponencial. Vida media efectiva:

- $\rho = 0.90$: $\tau_{1/2} \approx 6.6$ observaciones (memoria corta, adaptación rápida)
- $\rho = 0.95$: $\tau_{1/2} \approx 13.5$ observaciones (balance intermedio)
- $\rho = 0.98$: $\tau_{1/2} \approx 34.3$ observaciones (memoria larga, estabilidad)

Parámetro Fijo alpha (λ): Parámetro de regularización Ridge, fijo en $\lambda = 0.1$. Controla el trade-off sesgo-varianza: valores cercanos a cero aproximan mínimos cuadrados no regularizados, valores grandes contraen coeficientes hacia cero. El valor $\lambda = 0.1$ proporciona regularización suave sin sesgo excesivo.

Espacio de Búsqueda La optimización explora tres configuraciones estratégicamente diseñadas:

Configuración 1 - Estándar:

- `n_lags=5, rho=0.95, poly_degree=2`

Configuración 2 - Memoria Corta:

- `n_lags=10, rho=0.90, poly_degree=2`

Configuración 3 - Alta No Linealidad:

- `n_lags=5, rho=0.98, poly_degree=3`

Protocolo de Optimización La selección entre configuraciones minimiza **ECRPS** sobre el conjunto de calibración de 40 observaciones:

$$(p^*, \rho^*, d^*) = \arg \min_{(p, \rho, d) \in \mathcal{H}} \frac{1}{N_{\text{cal}}} \sum_{i=1}^{N_{\text{cal}}} \text{CRPS}(\hat{F}_i, y_i) \quad (3-96)$$

donde $\mathcal{H} = \{(5, 0.95, 2), (10, 0.90, 2), (5, 0.98, 3)\}$ es el conjunto de configuraciones candidatas y \hat{F}_i es la distribución predictiva empírica generada por el método histórico.

Congelamiento de Hiperparámetros Una vez optimizada, la configuración se congela mediante :

1. Estima $\mu_{\text{frozen}}, \sigma_{\text{frozen}}$ sobre datos de entrenamiento
2. Entrena el modelo Ridge ponderado hasta convergencia
3. Almacena coeficientes: $\hat{\beta}_{\text{frozen}}$
4. Establece flag `_is_frozen = True`

Durante toda la evaluación rolling, `fit_predict` verifica este flag: si es `True`, omite entrenamiento y procede directamente a predecir con el modelo existente, garantizando validez estadística sin fuga de información futura.

Ensemble Conformalized Quantile Regression (EnCQR-LSTM)

Explicación Teórica del Modelo El *Ensemble Conformalized Quantile Regression* (EnCQR-LSTM) constituye una síntesis metodológica avanzada que integra tres paradigmas complementarios del aprendizaje estadístico: regresión cuantílica (QR), predicción conformal (CP) y aprendizaje en ensamble. Esta arquitectura híbrida busca heredar simultáneamente la **adaptabilidad heterocedástica** de QR y las **garantías formales de cobertura** de CP, superando las limitaciones inherentes de cada enfoque por separado.

Motivación y Limitaciones de Métodos Puros Los métodos basados únicamente en regresión cuantílica pueden generar intervalos adaptativos cuya amplitud varía localmente con la volatilidad de los datos, pero carecen de garantías formales de cobertura. En la práctica, los intervalos de predicción generados por QR tienden a ser excesivamente confiados (demasiado estrechos), resultando en coberturas empíricas significativamente inferiores al nivel nominal $(1 - \alpha)$. (Jensen, Bianchi, and Anfinsen 2022)

Por otra parte, los métodos de predicción conformal estándar garantizan cobertura marginal válida bajo intercambiabilidad, pero construyen intervalos de amplitud constante o ligeramente variable. Para series temporales con heterocedasticidad, donde la incertidumbre fluctúa considerablemente, estos intervalos resultan excesivamente conservadores en períodos de baja volatilidad e insuficientes en períodos de alta volatilidad.

EnCQR-LSTM aborda ambas limitaciones mediante una arquitectura de ensamble que combina estimadores LSTM de regresión cuantílica con un mecanismo de conformalización posterior.

Arquitectura de Ensamble y Regresión Cuantílica El método presentado por Jensen, Bianchi, and Anfinsen 2022 construye un ensamble homogéneo de B modelos LSTM, cada uno entrenado sobre subconjuntos **disjuntos** del conjunto de entrenamiento $\{(x_i, y_i)\}_{i=1}^T$. La partición se define como:

$$S_b = \{(x_i, y_i) : i \in [(b-1)T_b + 1, bT_b]\}, \quad b = 1, \dots, B \quad (3-97)$$

donde $T_b = \lfloor T/B \rfloor$. Esta fragmentación disjunta es fundamental para construir residuos fuera de muestra válidos, garantizando que cada observación está excluida de al menos un modelo del ensamble.

Cada modelo LSTM estima simultáneamente múltiples funciones cuantílicas condicionales mediante la minimización de la pérdida pinball agregada:

$$\mathcal{L}_{\text{pinball}}(\theta_b) = \frac{1}{|S_b| \cdot |\mathcal{T}|} \sum_{(x_i, y_i) \in S_b} \sum_{\tau \in \mathcal{T}} \rho_\tau(y_i - \hat{q}_\tau^{(b)}(x_i)) \quad (3-98)$$

donde $\mathcal{T} = \{0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, 0.99\}$ es el conjunto de cuantiles objetivo, y $\rho_\tau(u) = \max(\tau \cdot u, (\tau - 1) \cdot u)$ es la función de pérdida pinball que penaliza asimétricamente los errores según el cuantil objetivo.

Predicción Leave-One-Out y Scores de Conformidad Una vez entrenados los B modelos, EnCQR construye predicciones leave-one-out (LOO) para cada observación de entrenamiento. Para una observación i en el subconjunto S_b , se agregan las predicciones de todos los modelos que no incluyeron esa observación:

$$\hat{q}_\tau^{(-i)}(x_i) = \frac{1}{B-1} \sum_{b': i \notin S_{b'}} \hat{q}_\tau^{(b')}(x_i) \quad (3-99)$$

Este procedimiento LOO reemplaza el requisito de intercambiabilidad por un esquema de validación cruzada que genera residuos genuinamente fuera de muestra, esencial para series temporales.

EnCQR introduce scores de conformidad **asimétricos** que cuantifican separadamente el error de cobertura en las colas inferior y superior:

$$\begin{aligned} E_i^{\text{lo}} &= \hat{q}_{\tau_{\text{lo}}}^{(-i)}(x_i) - y_i \\ E_i^{\text{hi}} &= y_i - \hat{q}_{\tau_{\text{hi}}}^{(-i)}(x_i) \end{aligned} \quad (3-100)$$

Esta formulación asimétrica permite que las distribuciones de errores para los cuantiles inferior y superior tengan formas diferentes, crucial en presencia de asimetría sistemática.

Conformalización y Distribución Predictiva Para una nueva observación x_{T+1} , el ensamble completo genera predicciones agregadas que se conformalizan mediante:

$$\hat{C}_\alpha(x_{T+1}) = [\hat{q}_{\tau_{lo}}(x_{T+1}) - \omega^{\text{lo}}, \hat{q}_{\tau_{hi}}(x_{T+1}) + \omega^{\text{hi}}] \quad (3-101)$$

donde $\omega^{\text{lo}} = Q_{1-\alpha}(\{E_i^{\text{lo}}\}_{i=1}^T)$ y $\omega^{\text{hi}} = Q_{1-\alpha}(\{E_i^{\text{hi}}\}_{i=1}^T)$ son los cuantiles $(1-\alpha)$ empíricos de las distribuciones de scores.

Una innovación clave de la implementación es el ajuste de una distribución **Skew-Normal** paramétrica a los cuantiles conformalizados. La función de densidad es:

$$f(x; \mu, \sigma, \alpha) = \frac{2}{\sigma} \phi\left(\frac{x - \mu}{\sigma}\right) \Phi\left(\alpha \frac{x - \mu}{\sigma}\right) \quad (3-102)$$

donde $\phi(\cdot)$ y $\Phi(\cdot)$ son la densidad y distribución acumulada Normal estándar. Los parámetros (μ, σ, α) se estiman minimizando:

$$(\hat{\mu}, \hat{\sigma}, \hat{\alpha}) = \arg \min_{(\mu, \sigma, \alpha)} \sum_{i=1}^{|\mathcal{T}|} (\hat{q}_{\tau_i}^{\text{conf}} - F_{\text{SN}}^{-1}(\tau_i; \mu, \sigma, \alpha))^2 \quad (3-103)$$

Este ajuste garantiza unimodalidad y permite generar $M = 1000$ muestras de la distribución predictiva completa.

Naturaleza del Modelo EnCQR-LSTM es un modelo **semi-paramétrico**. La arquitectura LSTM y la función de pérdida pinball son completamente no paramétricas respecto a la distribución de $Y|X$. Sin embargo, la fase final de ajuste Skew-Normal introduce una componente paramétrica para suavizar la distribución predictiva. No obstante, las garantías de cobertura provienen del mecanismo conformal no paramétrico, no del ajuste distribucional.

Propiedades Teóricas Bajo el supuesto de que el proceso de error es estacionario y fuertemente mezclante, EnCQR garantiza (Jensen, Bianchi, and Anfinsen 2022):

1. **Cobertura marginal válida:** $\mathbb{P}\{Y_{T+1} \in \hat{C}_\alpha(X_{T+1})\} \geq 1 - \alpha + O(1/T)$

2. **Adaptabilidad heterocedástica:** La amplitud varía localmente con x_{T+1} a través de $\hat{q}_\tau(x_{T+1})$
3. **Robustez ante especificación incorrecta:** Incluso con modelo LSTM mal especificado, la conformalización mantiene cobertura válida

De la Teoría a la Práctica La implementación de EnCQR-LSTM para esta investigación introduce adaptaciones específicas respecto al marco teórico original, priorizando factibilidad computacional y robustez empírica.

Adaptaciones Principales (1) **Tamaño de Ensamble Reducido:** Se utiliza $B = 3$ modelos en lugar de $B \geq 5$ como sugiere la literatura. Esta reducción responde al trade-off entre diversidad del ensamble y tamaño de subconjuntos de entrenamiento. Para series de longitud $T \approx 200 - 500$, $B = 3$ genera subconjuntos de 65 – 165 observaciones, suficientes para entrenar LSTMs efectivos sin fragmentación excesiva.

(2) **Conjunto de Cuantiles Reducido:** En lugar de estimar 19+ cuantiles, se limita a $\mathcal{T} = \{0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, 0.99\}$ (9 cuantiles). Esta reducción disminuye la complejidad de la capa de salida del LSTM y acelera el entrenamiento, manteniendo resolución suficiente para caracterizar la distribución predictiva.

(3) **Ajuste Skew-Normal como Suavizado:** El paper original no especifica cómo generar distribuciones continuas desde cuantiles discretos. La implementación introduce el ajuste Skew-Normal como mecanismo de suavizado, evitando interpolación lineal que puede generar distribuciones bimodales artificiales. Si la optimización falla, se emplea fallback a distribución Normal estándar.

(4) **Actualización de Scores Deshabilitada:** El mecanismo de ventana deslizante para actualizar scores conformales ($s = 24$) no se implementa en la versión evaluada. Todos los scores se calculan una sola vez durante la fase de calibración y permanecen congelados. Esta simplificación prioriza reproducibilidad y reduce complejidad computacional.

(5) **Arquitectura LSTM Simplificada:** Se emplea $L = 2$ capas LSTM con 32 unidades cada una, sin mecanismos avanzados como attention o skip connections. La regularización se limita a dropout ($p = 0.1$) y penalización L2 implícita en Adam.

Optimización, Parámetros e Hiperparámetros

Aspecto	Teoría Original	Implementación
Tamaño de ensamble (B)	$B \geq 5$ recomendado	$B = 3$ fijo
Cuantiles estimados	19+ cuantiles para alta resolución	9 cuantiles estratégicos
Distribución predictiva	No especificada	Ajuste Skew-Normal paramétrico
Actualización de scores	Ventana deslizante cada s observaciones	Deshabilitada (scores congelados)
Arquitectura LSTM	No especificada	2 capas, 32 unidades, dropout 0.1
Protocolo de congelamiento	Scores adaptativos	Congelamiento total (modelos + scores)
Complejidad computacional	Alta (múltiples entrenamientos + updates)	Reducida (sin actualizaciones)

Tabla 3-12.: Comparativa: Teoría vs. Implementación de EnCQR-LSTM

Hiperparámetros del Modelo `n_lags` (N_x): Longitud de ventana temporal de entrada. Define cuántos valores históricos se usan para predecir el siguiente.

units (N_u): Número de unidades (dimensión del estado oculto) en cada capa LSTM. Controla la capacidad expresiva del modelo.

epochs: Número de pasadas sobre el conjunto de entrenamiento. Se implementa early stopping con paciencia de 50 épocas para prevenir sobreajuste.

Parámetros Fijos:

- $B = 3$ (número de modelos en ensamble)
- $L = 2$ (número de capas LSTM)
- lr = 0.005 (tasa de aprendizaje Adam)
- batch_size = 16
- dropout = 0.1
- $\alpha = 0.05$ (nivel de error nominal, cobertura 95%)
- num_samples = 1000 (muestras de distribución Skew-Normal)

Espacio de Búsqueda Restringido Dada la alta complejidad computacional de entrenar ensambles de LSTMs, el espacio de hiperparámetros se limita estratégicamente a dos configuraciones:

Configuración 1 (Conservadora): ($N_x = 10, N_u = 24, \text{epochs} = 20$)

Configuración 2 (Estándar): ($N_x = 20, N_u = 32, \text{epochs} = 25$)

Optimización y Congelamiento La selección entre ambas configuraciones se realiza minimizando **ECRPS** sobre un conjunto de validación temporal separado. Formalmente:

$$(N_x^*, N_u^*, \text{epochs}^*) = \arg \min_{(N_x, N_u, e) \in \mathcal{H}} \frac{1}{N_{\text{val}}} \sum_{i=1}^{N_{\text{val}}} \text{CRPS}(\mathcal{F}_i, y_i) \quad (3-104)$$

donde $\mathcal{H} = \{(10, 24, 20), (20, 32, 25)\}$ y \mathcal{F}_i es la distribución predictiva para la observación i .

Una vez optimizada la configuración, se ejecuta el protocolo de congelamiento completo mediante , que fija permanentemente:

- Los pesos $\{\theta_b\}_{b=1}^B$ de los B modelos LSTM
- Los parámetros de normalización MinMax: (y_{\min}, y_{\max})
- Las distribuciones empíricas de scores conformales: $\{E_i^{\text{lo}}\}_{i=1}^T$ y $\{E_i^{\text{hi}}\}_{i=1}^T$

Durante la evaluación rolling window, para cada punto de prueba $t = T + 1, \dots, T + T'$, se extrae la ventana de entrada, se normalizan los datos con parámetros congelados, se generan predicciones de los B modelos congelados, se agregan mediante media aritmética, se conformalizan usando scores congelados, se des-normalizan, y se ajusta la distribución Skew-Normal final. Este protocolo garantiza ausencia total de data leakage y evaluación justa comparable con otros métodos.

Justificación del Espacio Reducido La decisión de limitar el espacio de búsqueda a solo 2 configuraciones se fundamenta en:

1. **Costo computacional prohibitivo:** Cada configuración requiere entrenar $B = 3$ LSTMs completos. Una búsqueda exhaustiva sobre $\{10, 15, 20\} \times \{24, 32, 64\} \times$

$\{20, 25, 30\}$ implicaría 27 entrenamientos por serie, inviable para el benchmark de 100 series.

2. **Evidencia de rendimientos decrecientes:** Experimentos piloto mostraron que configuraciones intermedias rara vez superaban los extremos, sugiriendo un comportamiento bimodal del desempeño.
3. **Priorización de diversidad metodológica:** Los recursos computacionales se asignan a evaluar múltiples arquitecturas fundamentalmente diferentes (LSPM, MCPS, AREPD, DeepAR, EnCQR) en lugar de explorar exhaustivamente el espacio de un solo método.

Esta estrategia de optimización restringida permite evaluar el potencial de EnCQR-LSTM manteniendo el estudio computacionalmente viable, reconociendo que representaciones más sofisticadas (búsqueda bayesiana, algoritmos genéticos) quedan fuera del alcance de esta investigación.

3.4.3. Contribuciones de este trabajo

3.5. Método de comparación

3.6. Simulaciones Complementarias

Además del diseño principal descrito en la Sección 3.2, se implementaron cinco conjuntos de simulaciones complementarias diseñadas para resolver dilemas metodológicos específicos sobre el preprocesamiento, la persistencia de los datos, la arquitectura de la muestra y la propagación de la incertidumbre en horizontes lejanos. Estas simulaciones abordan preguntas fundamentales que no pueden responderse mediante el diseño principal debido a su estructura particular de ventana rodante a un paso.

3.6.1. Simulación 1: Impacto de la Diferenciación en ARIMA ($d = 1$)

Motivación

Esta simulación evalúa una pregunta metodológica fundamental: ¿los métodos conformales capturan mejor la variabilidad cuando operan sobre la serie diferenciada estacionaria (ΔY_t) o sobre los niveles integrados (Y_t)? Esta cuestión es relevante porque, aunque la diferenciación elimina la no estacionariedad y simplifica el modelado, también puede introducir pérdida de información sobre el nivel de la serie y afectar la estructura de autocorrelación (Rob J Hyndman and Athanasopoulos 2021).

Diseño Experimental

Se utilizan las 7 configuraciones ARIMA($p, 1, q$) del diseño principal (Sección 3.3), combinadas con las 5 distribuciones de ruido y 4 niveles de varianza, generando:

$$N_{\text{config}} = 7 \text{ procesos} \times 5 \text{ distribuciones} \times 4 \text{ varianzas} = 140 \text{ configuraciones base} \quad (3-105)$$

Cada configuración se ejecuta bajo dos modalidades:

1. **SIN_DIFF**: El modelo recibe la serie integrada Y_t directamente y genera una distribución predictiva $\hat{F}_{Y_{t+1}}$ para el siguiente valor en niveles.
2. **CON_DIFF**: El modelo recibe la serie diferenciada $\Delta Y_t = Y_t - Y_{t-1}$. La predicción generada $\widehat{\Delta Y}_{t+1}$ se integra mediante:

$$\hat{Y}_{t+1} = Y_t + \widehat{\Delta Y}_{t+1} \quad (3-106)$$

para calcular el ECRPS en el espacio de niveles, permitiendo una comparación directa con la modalidad SIN_DIFF.

Protocolo de Evaluación

Siguiendo el esquema de la simulación principal:

- Serie simulada: $n_{\text{total}} = 252$ observaciones efectivas (200 entrenamiento + 40 calibración + 12 prueba)

- Esquema de ventana rodante con 12 pasos de predicción
- Evaluación de los 9 modelos conformales mediante ECRPS contra la distribución teórica

Esto genera:

$$N_{\text{filas}} = 140 \times 2 \text{ modalidades} \times 12 \text{ pasos} = 3,360 \text{ evaluaciones} \quad (3-107)$$

3.6.2. Simulación 2: Límites de Integración y Persistencia (Multi-D)

Motivación

Se investiga la estabilidad numérica de los métodos conformales ante órdenes de integración elevados $d \in \{1, 2, 3, 4, 5, 6, 7, 10\}$. A medida que d aumenta, la serie integrada Y_t exhibe una persistencia extrema y rangos de valores cada vez más amplios, lo que puede desestabilizar modelos que no utilizan diferenciación previa adecuada.

Diseño Experimental

A partir de las 7 configuraciones ARMA(p, q) base, se genera el proceso ARIMA(p, d, q) mediante:

$$Y_t = \sum_{j=0}^{d-1} \nabla^j W_t, \quad \text{donde } W_t \sim \text{ARMA}(p, q) \quad (3-108)$$

Con la estructura completa:

$$N_{\text{config}} = 7 \text{ ARMA} \times 8 \text{ valores de } d \times 5 \text{ distribuciones} \times 4 \text{ varianzas} = 1,120 \text{ configuraciones} \quad (3-109)$$

Cada configuración se evalúa bajo las dos modalidades (SIN_DIFF y CON_DIFF) en 12 pasos:

$$N_{\text{filas}} = 1,120 \times 2 \times 12 = 26,880 \text{ evaluaciones} \quad (3-110)$$

Hipótesis

Se hipotetiza que para $d \geq 4$, el rango explosivo de Y_t desestabilizará a los modelos en modalidad SIN_DIFF, mientras que la modalidad CON_DIFF mantendrá estabilidad numérica al operar en el espacio diferenciado acotado.

3.6.3. Simulación 3: Efectos del Tamaño Muestral Absoluto

Motivación

Esta simulación caracteriza la tasa de convergencia de las distribuciones predictivas empíricas hacia la densidad teórica a medida que el volumen de datos aumenta. Permite cuantificar el trade-off entre calidad de estimación (que mejora con más datos de entrenamiento) y precisión de calibración (que mejora con más datos de calibración). A diferencia del diseño principal que mantiene fijos los tamaños $n_{\text{train}} = 200$ y $n_{\text{calib}} = 40$, aquí se explora sistemáticamente el espacio de tamaños absolutos manteniendo una proporción fija entre entrenamiento y calibración.

Diseño Experimental

Se evalúan los tres tipos de procesos (ARMA, ARIMA, SETAR) con sus 7 configuraciones cada uno, bajo cinco tamaños muestrales totales manteniendo una proporción fija de aproximadamente 83% para entrenamiento y 17% para calibración:

Etiqueta	n_{train}	n_{calib}	n_{total}	Proporción
N=120	100	20	120	83:17
N=240	199	41	240	83:17
N=360	299	61	360	83:17
N=600	498	102	600	83:17
N=1200	996	204	1,200	83:17

Tabla 3-13.: Tamaños muestrales evaluados con proporción fija

Esta estructura permite evaluar el efecto del incremento simultáneo de datos de entrenamiento y calibración, manteniendo constante su proporción relativa. Esto aísla el efecto puro del tamaño muestral total sobre la calidad de las distribuciones predictivas.

La estructura completa genera:

$$\begin{aligned}
 N_{\text{config}} &= 3 \text{ tipos} \times 7 \text{ configs} \times 5 \text{ tamaños} \\
 &\quad \times 5 \text{ distribuciones} \times 4 \text{ varianzas} \\
 &= 2,100 \text{ configuraciones}
 \end{aligned} \tag{3-111}$$

Con 12 pasos de evaluación por configuración:

$$N_{\text{filas}} = 2,100 \times 12 = 25,200 \text{ evaluaciones} \tag{3-112}$$

3.6.4. Simulación 4: Proporciones de Calibración con Tamaño Fijo

Motivación

En escenarios de datos limitados, existe un conflicto fundamental entre usar datos para mejorar el ajuste del modelo (*training*) o para mejorar la precisión de los intervalos conformales (*calibration*). Esta simulación busca determinar si existe una “proporción óptima” que minimice el ECRPS promedio cuando el presupuesto total de datos es fijo.

Diseño Experimental

Se fija un presupuesto total de $n_{\text{total}} = 240$ observaciones y se evalúan 5 proporciones de calibración diferentes:

Proporción	n_{train}	n_{calib}	Ratio
10%	216	24	9:1
20%	192	48	4:1
30%	168	72	7:3
40%	144	96	3:2
50%	120	120	1:1

Tabla 3-14.: Proporciones de calibración evaluadas (N=240 fijo)

La estructura completa es:

$$\begin{aligned}
 N_{\text{config}} &= 3 \text{ tipos} \times 7 \text{ configs} \times 5 \text{ proporciones} \\
 &\quad \times 5 \text{ distribuciones} \times 4 \text{ varianzas} \\
 &= 2,100 \text{ configuraciones}
 \end{aligned} \tag{3-113}$$

Con 12 pasos de evaluación por configuración:

$$N_{\text{filas}} = 2,100 \times 12 = 25,200 \text{ evaluaciones} \tag{3-114}$$

3.6.5. Simulación 5: Predicción Multi-paso (Horizonte h)

Motivación

El diseño principal evalúa la predicción a un paso adelante mediante un esquema de ventana rodante, donde el modelo se actualiza con cada nueva observación antes de realizar la siguiente predicción. Sin embargo, muchas aplicaciones prácticas requieren pronósticos para múltiples períodos futuros sin acceso a observaciones intermedias (Rob J Hyndman and Athanasopoulos 2021). Esta simulación evalúa cómo se degrada la calidad de la distribución predictiva cuando el modelo debe proyectar h pasos hacia adelante de forma recursiva, alimentándose exclusivamente de sus propias predicciones anteriores.

Fundamento Metodológico: Predicción Recursiva

Existen dos estrategias principales para pronóstico multi-paso (Ben Taieb et al. 2012):

1. **Estrategia Directa:** Entrenar modelos separados para cada horizonte h , cada uno estimando directamente \hat{Y}_{t+h} desde $Y_{1:t}$. Aunque conceptualmente simple, requiere entrenar H modelos independientes y no aprovecha la estructura secuencial del problema.
2. **Estrategia Recursiva (Iterativa):** Utilizar un único modelo de predicción a un paso y aplicarlo recursivamente:

$$\hat{Y}_{t+h} = f(\hat{Y}_{t+h-1}, \hat{Y}_{t+h-2}, \dots, Y_t), \quad h = 2, 3, \dots, H \tag{3-115}$$

Esta estrategia, aunque propaga errores de predicción, es más eficiente computacionalmente y refleja mejor la práctica operativa donde no hay acceso a observaciones futuras verdaderas (Bontempi, Ben Taieb, and Le Borgne 2013).

La presente simulación implementa la estrategia recursiva, que es la más relevante para evaluar métodos conformales en horizontes extendidos. La propagación de incertidumbre en este contexto ha sido estudiada por Gneiting and Katzfuss (2014) y Stankeviciute, Alaa, and Schaar (2021), quienes demuestran que la distribución predictiva en el horizonte h debe considerar tanto la incertidumbre del modelo como la acumulación de errores de pasos previos.

Generación de Trayectorias Estocásticas

Para cada configuración y modelo, se generan $N_{\text{traj}} = 100$ trayectorias completas desde el mismo punto de origen t . Cada trayectoria m se construye mediante muestreo recursivo de las distribuciones predictivas:

$$\begin{aligned} \hat{Y}_{t+1}^{(m)} &\sim \hat{F}_{\text{modelo}}(\cdot | Y_{1:t}) \\ \hat{Y}_{t+2}^{(m)} &\sim \hat{F}_{\text{modelo}}(\cdot | Y_{1:t}, \hat{Y}_{t+1}^{(m)}) \\ &\vdots \\ \hat{Y}_{t+h}^{(m)} &\sim \hat{F}_{\text{modelo}}\left(\cdot | Y_{1:t}, \hat{Y}_{t+1}^{(m)}, \dots, \hat{Y}_{t+h-1}^{(m)}\right) \end{aligned} \tag{3-116}$$

donde $m = 1, \dots, 100$ indexa las trayectorias independientes. Este procedimiento genera una distribución empírica para cada horizonte h , formada por las 100 realizaciones $\{\hat{Y}_{t+h}^{(1)}, \dots, \hat{Y}_{t+h}^{(100)}\}$.

Diseño Experimental

Se evalúan únicamente 4 modelos representativos (LSPM, DeepAR, Sieve Bootstrap, MCPS) debido al alto costo computacional de generar 100 trayectorias completas por escenario. La selección incluye:

- **LSPM:** Método conformal clásico basado en cuantiles empíricos

- **DeepAR:** Método paramétrico de aprendizaje profundo que modela distribuciones completas
- **Sieve Bootstrap:** Método no paramétrico basado en remuestreo de residuos
- **MCPS:** Método conformal contemporáneo que partitiona el espacio de calibración

Para cada tipo de proceso (ARMA, ARIMA, SETAR), la estructura es:

$$N_{\text{config}} = 7 \text{ configs} \times 5 \text{ distribuciones} \times 4 \text{ varianzas} = 140 \text{ configuraciones} \quad (3-117)$$

Con 3 tipos de procesos y 12 horizontes de predicción:

$$N_{\text{filas}} = 3 \times 140 \times 12 = 5,040 \text{ evaluaciones} \quad (3-118)$$

Protocolo de Evaluación

El protocolo sigue la estructura del diseño principal con las siguientes particularidades:

1. **Punto de origen fijo:** A diferencia de la ventana rodante, aquí todas las predicciones parten del mismo punto temporal $t = n_{\text{train}} + n_{\text{calib}}$.
2. **Sin actualización intermedia:** El modelo se ajusta una sola vez con los datos disponibles hasta t y no se actualiza durante la proyección de los $H = 12$ pasos.
3. **Evaluación por horizonte:** Para cada $h \in \{1, 2, \dots, 12\}$, se calcula:

$$\text{ECRPS}_h = \text{ECRPS} \left(\{\hat{Y}_{t+h}^{(1)}, \dots, \hat{Y}_{t+h}^{(100)}\}, \{Y_{t+h}^{(\text{true},1)}, \dots, Y_{t+h}^{(\text{true},1000)}\} \right) \quad (3-119)$$

4. **Análisis de degradación:** El experimento permite cuantificar la tasa de crecimiento de ECRPS_h conforme h aumenta, caracterizando la velocidad de degradación de la calidad predictiva.

Esta simulación es particularmente relevante para aplicaciones donde las decisiones deben tomarse con base en pronósticos de mediano plazo sin posibilidad de actualización frecuente del modelo, como en planeación energética, gestión de inventarios o política monetaria (Rob J Hyndman and Athanasopoulos 2021).

3.6.6. Resumen de Evaluaciones Complementarias

La Tabla 3-15 consolida el alcance total de estos experimentos. El volumen combinado de estas simulaciones complementarias es comparable al del diseño principal, reflejando la importancia de estos aspectos metodológicos para una evaluación comprehensiva.

Simulación	Factor Variado	Filas
1. Diferenciación ($d = 1$)	Modalidad (SIN_DIFF vs CON_DIFF)	3,360
2. Multi-D	Orden de integración $d \in \{1, \dots, 10\}$	26,880
3. Tamaño Muestral	Tamaño total con proporción fija (83:17)	25,200
4. Proporciones	Proporción de calibración ($n_{total} = 240$ fijo)	25,200
5. Multi-paso	Horizonte de predicción $h \in \{1, \dots, 12\}$	5,040
Total		85,680

Tabla 3-15.: Resumen de la carga experimental de simulaciones complementarias

Protocolo Común a Todas las Simulaciones

Todas las simulaciones complementarias mantienen consistencia metodológica con el diseño principal descrito en la Sección 3.2:

- **Período de inicialización (burn-in):** 100 observaciones descartadas antes del inicio efectivo de la serie para eliminar efectos transitorios de las condiciones iniciales.
- **Métrica de evaluación:** ECRPS entre la distribución predictiva empírica del modelo (basada en 1,000 muestras bootstrap o generadas por el modelo) y la distribución teórica verdadera del proceso generador de datos (representada por 1,000 muestras de la densidad real), como se define en la Sección 2.2.3.
- **Modelos evaluados:** Los 9 métodos conformales especificados en la Sección 3.4 (Block Bootstrapping, Sieve Bootstrap, LSPM, LSPMW, AREPD, MCPS, AV-MCPS, DeepAR, EnCQR-LSTM), excepto en la Simulación 5 donde por razones de

eficiencia computacional se evalúan únicamente 4 modelos representativos (LSPM, DeepAR, Sieve Bootstrap, MCPS).

- **Calibración y ajuste de hiperparámetros:** Cada modelo se optimiza siguiendo un procedimiento de dos etapas. Primero, los hiperparámetros del modelo base se seleccionan mediante validación cruzada temporal en el conjunto de entrenamiento, minimizando el ECRPS en una ventana de validación. Segundo, los parámetros conformales (como el nivel de cobertura o los pesos de calibración) se ajustan usando el conjunto de calibración dedicado.
- **Control de reproducibilidad:** Semillas aleatorias fijas y documentadas para cada escenario, con incrementos determinísticos según el índice del escenario ($seed = seed_{base} + id_{escenario}$), permitiendo la replicación exacta de todos los experimentos.
- **Esquema de evaluación:**
 - *Ventana rodante* (Simulaciones 1–4): El modelo se actualiza con cada nueva observación antes de predecir el siguiente paso, generando 12 predicciones secuenciales donde el horizonte siempre es $h = 1$ pero el conjunto de entrenamiento crece.
 - *Proyección desde origen fijo* (Simulación 5): El modelo se ajusta una sola vez en t y proyecta recursivamente los horizontes $h \in \{1, 2, \dots, 12\}$ sin actualización intermedia, propagando la incertidumbre a través de predicciones iteradas.
- **Estructura completa:** Todas las simulaciones evalúan las 21 configuraciones de procesos (7 ARMA + 7 ARIMA + 7 SETAR) cruzadas con 5 distribuciones de ruido (normal, uniforme, exponencial, t-student, mezcla) y 4 niveles de varianza (0.2, 0.5, 1.0, 3.0), garantizando cobertura exhaustiva del espacio paramétrico.

La uniformidad en estos aspectos fundamentales garantiza que las diferencias observadas en el desempeño sean atribuibles exclusivamente a los factores experimentales bajo estudio (modalidad de diferenciación, orden de integración, tamaño muestral, proporción de calibración, u horizonte de predicción) y no a variaciones en el protocolo de evaluación.

4. Resultados y Análisis de Simulaciones

Este capítulo presenta los resultados del diseño experimental descrito en el Capítulo 3, evaluando el desempeño de los nueve métodos de predicción conformal y probabilística bajo condiciones controladas donde el proceso generador de datos es conocido. A diferencia de las aplicaciones a series reales del Capítulo 5, donde la distribución verdadera es desconocida, el entorno de simulación permite comparar directamente las distribuciones predictivas empíricas contra la densidad teórica mediante el ECRPS.

La estructura del capítulo refleja la organización del diseño experimental. La Sección 4.1 analiza los resultados del diseño principal, que combina tres escenarios de simulación (ARMA, ARIMA, SETAR), siete configuraciones paramétricas por escenario, cinco distribuciones de ruido y cuatro niveles de varianza. Las secciones subsecuentes abordan las cinco simulaciones complementarias que exploran dimensiones metodológicas específicas: impacto de la diferenciación en ARIMA (Sección 4.2), límites de integración múltiple (Sección 4.3), efectos del tamaño muestral (Sección 4.4), proporciones óptimas de calibración (Sección 4.5), y degradación en predicción multi-paso (Sección 4.6).

4.1. Resultados del Diseño Principal

Esta sección presenta los resultados del diseño completo descrito en la Sección 3.2, que evalúa 9 métodos conformales bajo 420 configuraciones únicas distribuidas entre tres escenarios (ARMA, ARIMA, SETAR). El análisis se estructura en dos niveles: primero se examinan los patrones agregados que emergen del conjunto completo de simulaciones, y posteriormente se descomponen los resultados por escenario para identificar comportamientos específicos asociados a cada clase de proceso generador de datos.

4.1.1. Análisis Agregado de Desempeño

La Figura 4-1 cuantifica el ECRPS promedio por escenario, ordenando los modelos según su desempeño en procesos ARIMA (el escenario más exigente). La clara separación entre grupos de métodos confirma que la no estacionariedad amplifica las diferencias de rendimiento.

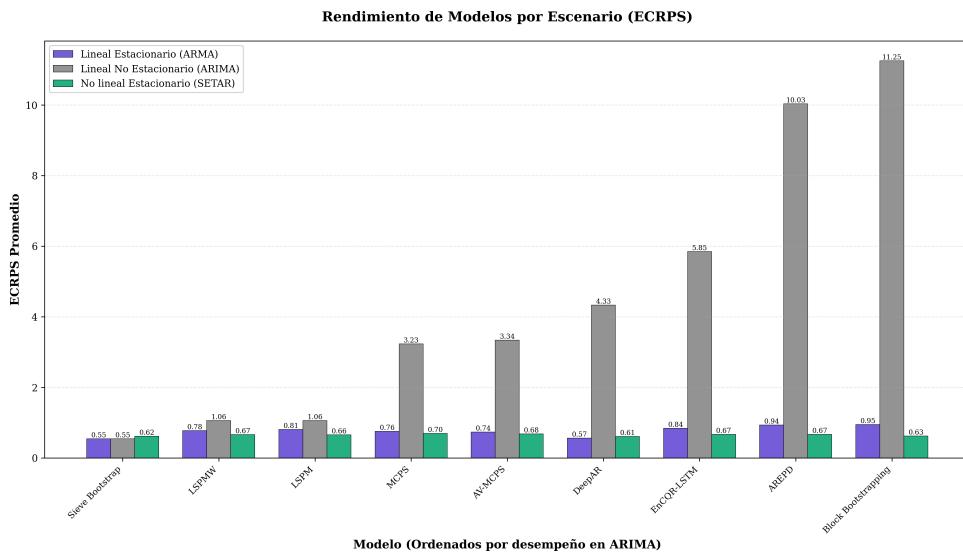


Figura 4-1.: ECRPS promedio por escenario.

Los modelos están ordenados por su desempeño en ARIMA (barras grises). El ranking por escenario evidencia que Block Bootstrapping y AREPD, experimentan degradación severa en ARIMA ($\text{ECRPS} > 10$), superando por amplio margen a métodos como Sieve Bootstrap ($\text{ECRPS } 0.55 - 0.62$ en todos los escenarios) y DeepAR ($\text{ECRPS} \approx 0.57 - 4.33$). En el escenario SETAR, los métodos convergen hacia un desempeño más homogéneo ($\text{ECRPS} \approx 0.63 - 0.70$).

La Figura 4-2 presenta los puntajes Z del ECRPS promedio para cada modelo evaluado en las 21 configuraciones paramétricas (7 por escenario). Los puntajes Z permiten comparar el desempeño relativo estandarizado, donde valores negativos indican un rendimiento superior al promedio y valores positivos señalan un desempeño inferior.

Los resultados agregados revelan tres hallazgos principales. Primero, los métodos Sieve Bootstrap, LSPM y LSPMW exhiben la mayor estabilidad, alcanzando puntajes Z máximos a 3. Segundo, se observa una marcada heterogeneidad en el desempeño según la configuración: la columna ARIMA(2,1,2) concentra los valores Z más elevados (rojo in-

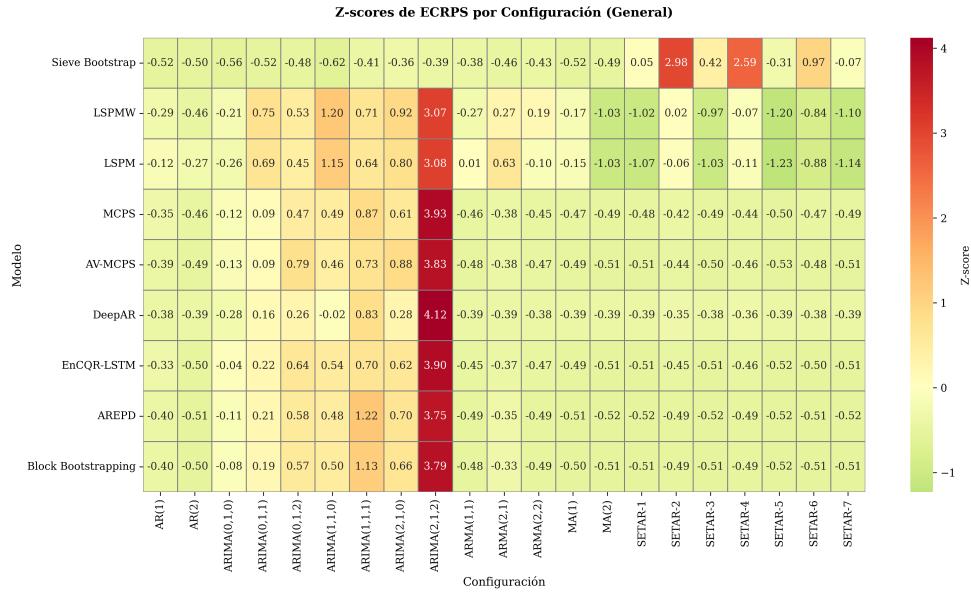


Figura 4-2.: Z-scores de ECRPS por configuración.

tenso) para prácticamente todos los métodos, sugiriendo que esta parametrización particular representa un desafío sistemático. Tercero, los métodos LSPM y LSPMW muestran un patrón bimodal, con excelente desempeño en configuraciones ARMA pero deterioro en configuraciones ARIMA específicas.

Desempeño por Configuración Específica

Las Figuras 4-3a–4-3c desagregan los resultados por escenario, revelando patrones de especialización según la estructura del proceso generador. En procesos ARMA (Figura 4-3a), se identifica una jerarquía clara de dificultad: el proceso ARMA(2,1) constituye el escenario más desafiante para la mayoría de métodos (con excepción de Sieve Bootstrap), mientras que MA(2) emerge como el más favorable. Los modelos LSPM, LSPMW y Sieve Bootstrap demuestran robustez consistente a lo largo de todas las configuraciones ARMA, con desempeño particularmente destacado en AR(1), un proceso que plantea dificultades considerables para los métodos restantes. En procesos ARIMA (Figura 4-3b), se observa una degradación progresiva en la mayoría de los modelos a medida que aumenta la complejidad de la configuración: LSPM, LSPMW, MCPS, AV-MCPS, DeepAR, EnCQR-LSTM, AREPD y Block Bootstrapping alcanzan Z-scores altos (>2.0 , rojo intenso) en ARIMA(2,1,2), sugiriendo colapso en escenarios no estacionarios complejos.

Sieve Bootstrap destaca por su estabilidad, con Z-scores consistentes y bajos (de -1.47 a 1.24, predominantemente verde y amarillo). DeepAR muestra robustez moderada en configuraciones simples (Z-scores \approx -0.5, verde), pero se deteriora en las más avanzadas. Los procesos SETAR (Figura 4-3c) presentan un desafío diferente: las configuraciones SETAR-2 y SETAR-4 (que incorporan cambios de régimen más abruptos) generan dificultades uniformes (Z-scores positivos) para todos los métodos excepto Sieve Bootstrap. Configuraciones con transiciones suaves (SETAR-1, SETAR-3, SETAR-5) permiten un desempeño más equilibrado.

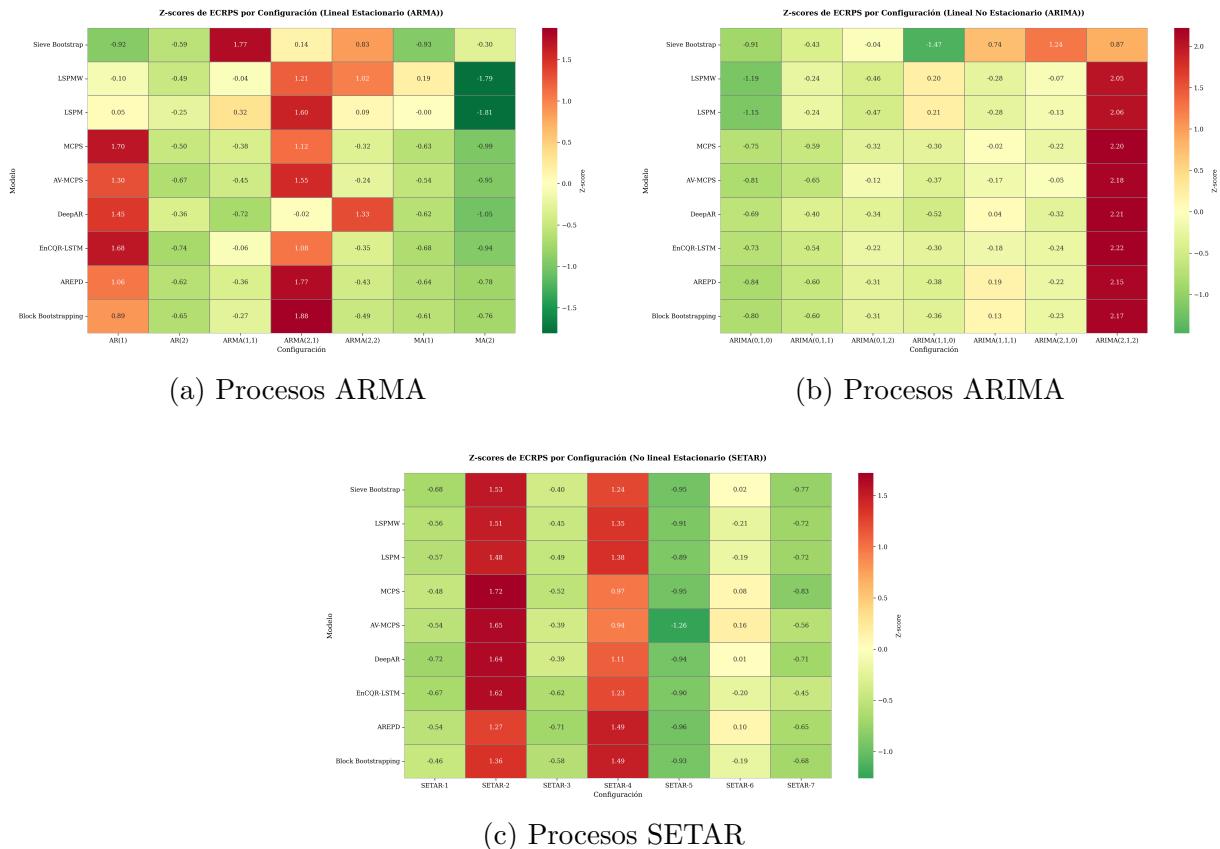


Figura 4-3.: Z-scores de ECRPS por configuración según familia de procesos.

Sensibilidad a la Distribución del Error

La Figura 4-4 examina el efecto de la distribución en el desempeño a nivel general. La distribución uniforme genera sistemáticamente los peores desempeños (Z-scores > 1.2 para todos los métodos excepto Sieve Bootstrap), mientras que la distribución t-student favorece

4.1. Resultados del Diseño Principal

consistentemente a todos los métodos ($Z\text{-scores} < -0.4$).

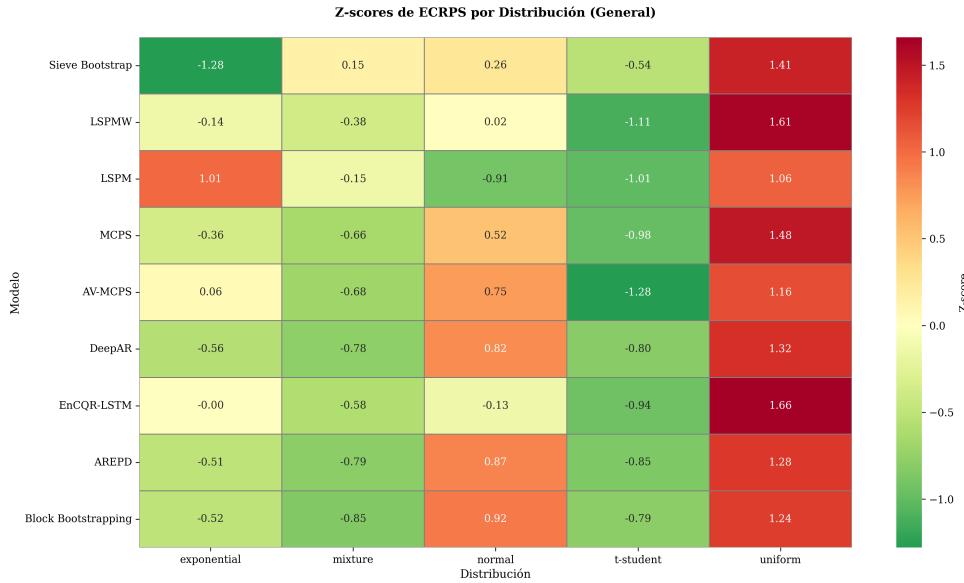


Figura 4-4.: Z-scores de ECRPS por distribución.

Las Figuras 4-5a–4-5c desagregan los resultados por distribución del error según familia de procesos, revelando patrones de sensibilidad a la forma de las innovaciones. En procesos ARMA (Figura 4-5a), la distribución uniforme representa el mayor desafío ($Z\text{-scores} > 1.0$, rojo) para la mayoría de modelos, mientras que la t-student es la más favorable ($Z\text{-scores} < -1.0$, verde oscuro en varios casos); Sieve Bootstrap y LSPMW mantienen robustez consistente con Z-scores bajos en mixture y t-student, pero AREPD y Block Bootstrapping muestran variabilidad extrema (de -1.62 a 1.10). Para procesos ARIMA (Figura 4-5b), el patrón se intensifica en no estacionariedad: uniform nuevamente colapsa la mayoría de métodos ($Z\text{-scores} \approx 1.5$, rojo intenso), con t-student ofreciendo alivio ($Z\text{-scores} \approx -1.0$, verde); DeepAR y EnCQR-LSTM destacan en mixture y t-student, pero Sieve Bootstrap lidera en estabilidad general ($Z\text{-scores}$ de -1.46 a 1.18). En procesos SETAR (Figura 4-5c), la no linealidad acentúa la vulnerabilidad a uniform ($Z\text{-scores} > 1.3$, rojo), con mixture y t-student permitiendo desempeño equilibrado ($Z\text{-scores} < -1.0$ en LSPM y Block Bootstrapping); Sieve Bootstrap y MCPS muestran resiliencia en normal y t-student, aunque AREPD falla en exponential ($Z\text{-score} 1.37$, rojo).

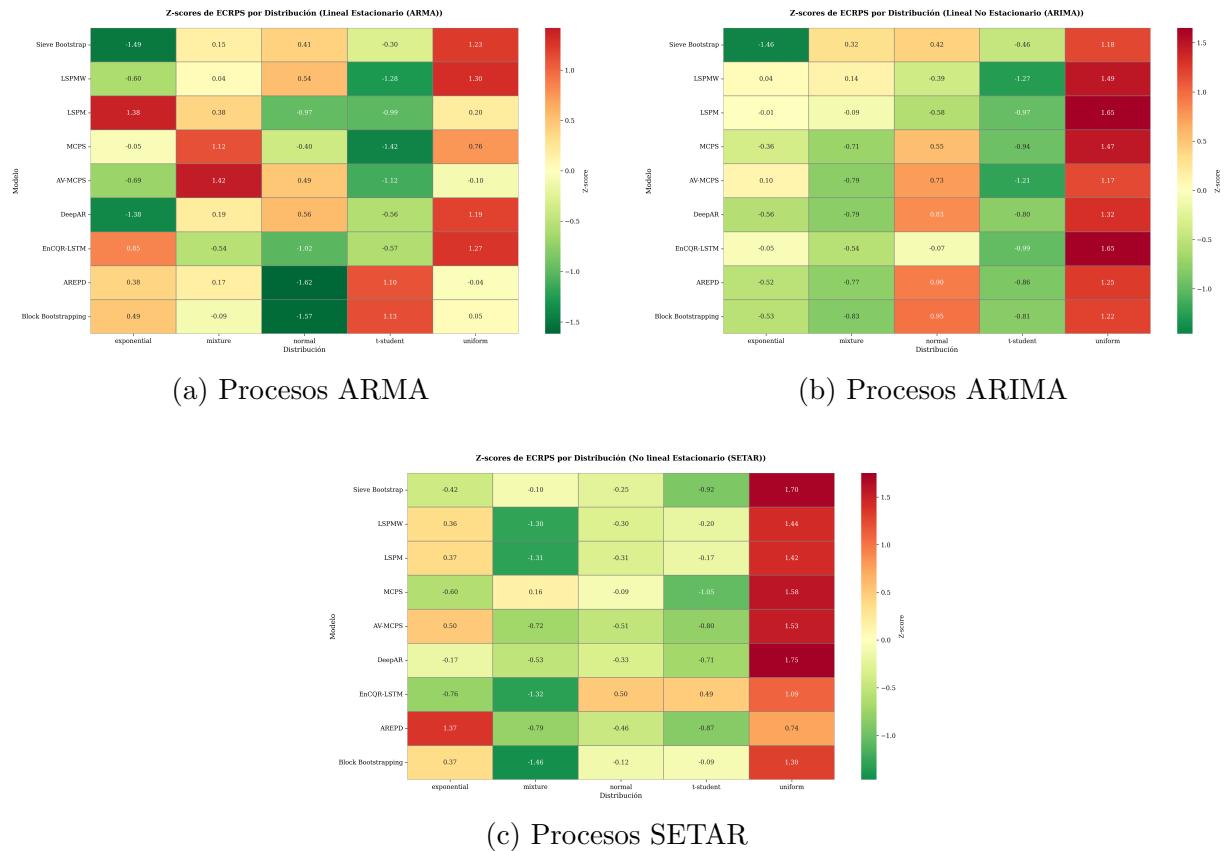


Figura 4-5.: Z-scores de ECRPS por distribución del error según familia de procesos.

Efecto de la Varianza del Error

La Figura 4-6 cuantifica la evolución del ECRPS conforme aumenta la varianza del término de error. Dos grupos claramente diferenciados emergen: métodos con crecimiento aproximadamente lineal (Sieve Bootstrap, LSPMW, LSPM, MCPS, AV-MCPS) y métodos con crecimiento super-lineal o exponencial (DeepAR, EnCQR-LSTM, AREPD, Block Bootstrapping).

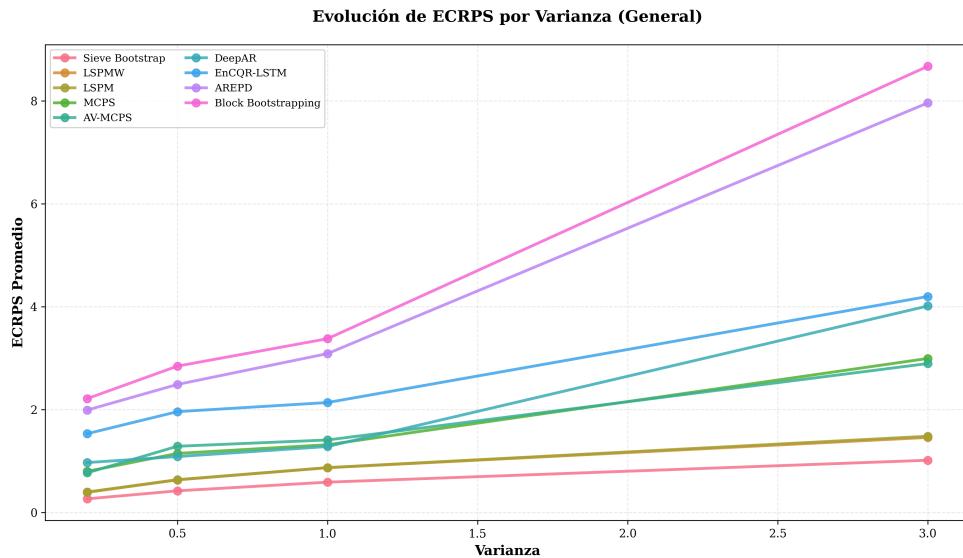
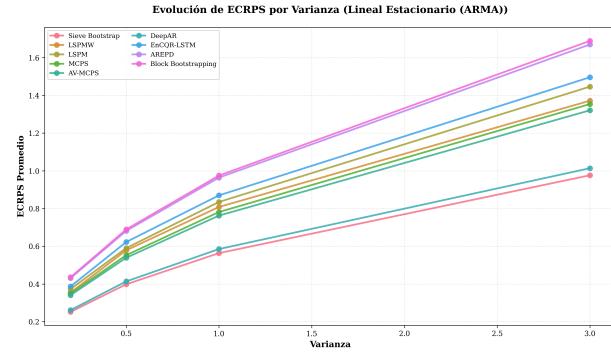
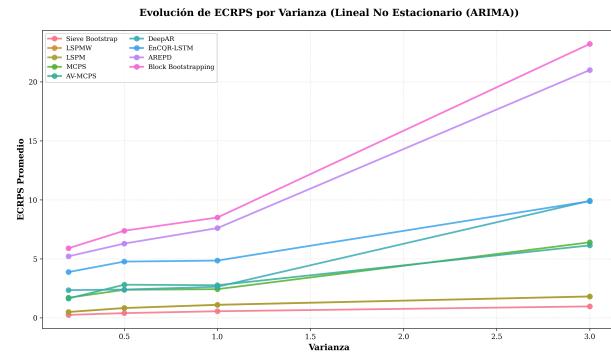


Figura 4-6.: ECRPS promedio en función de la varianza.

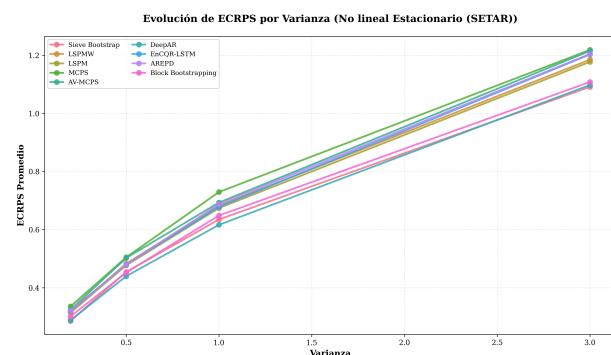
Las Figuras 4-7a–4-7c revelan que este comportamiento es altamente dependiente del escenario. En ARMA (Figura 4-7a), todos los métodos mantienen crecimiento controlado con pendientes similares. En ARIMA (Figura 4-7b), Block Bootstrapping y AREPD experimentan crecimiento explosivo para $\sigma^2 = 3.0$ ($ECRPS > 20$), mientras que Sieve Bootstrap permanece estable ($ECRPS < 2$). En SETAR (Figura 4-7c), el crecimiento se homogeniza nuevamente, sugiriendo que la no linealidad estacionaria atenúa las diferencias inducidas por la varianza del ruido.



(a) Procesos ARMA



(b) Procesos ARIMA



(c) Procesos SETAR

Figura 4-7.: ECRPS en función de la varianza del error.

4.1.2. Análisis de Robustez y Significancia Estadística

Estabilidad del Desempeño: Coeficiente de Variación

La Figura 4-8 cuantifica la estabilidad del desempeño de cada método mediante el coeficiente de variación (CV) del ECRPS a través de todas las configuraciones evaluadas. El CV permite identificar métodos cuyo rendimiento es predecible versus aquellos que exhiben alta sensibilidad a las condiciones del problema.

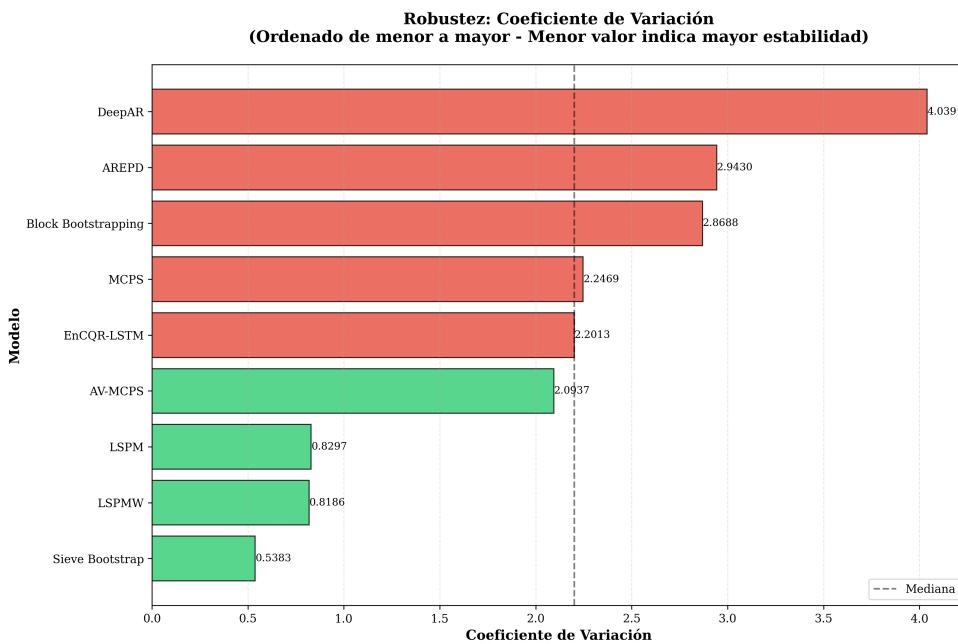


Figura 4-8.: Coeficiente de variación del ECRPS por modelo.

Los resultados revelan una clara estratificación. Sieve Bootstrap emerge como el método más robusto ($CV = 0.54$), seguido por LSPMW ($CV = 0.82$) y LSPM ($CV = 0.83$), todos significativamente por debajo de la mediana grupal. En contraste, DeepAR exhibe la mayor variabilidad ($CV = 4.04$), seguido por AREPD ($CV = 2.94$) y Block Bootstrapping ($CV = 2.87$). Este patrón sugiere que los métodos paramétricos y aquellos basados en remuestreo de bloques sufren colapsos severos en configuraciones específicas, mientras que los métodos conformales basados en cuantiles mantienen consistencia.

AV-MCPS ($CV = 2.09$) presenta una paradoja interesante: a pesar de ubicarse ligeramente por debajo de la mediana en términos absolutos, su variabilidad es sustancialmente mayor que los métodos conformales simples (LSPM, LSPMW), lo que indica que la ponderación

adaptativa introduce inestabilidad en ciertos escenarios sin beneficios consistentes.

Comparaciones Pareadas: Test de Diebold-Mariano

La Figura 4-9 presenta los resultados del test de Diebold-Mariano modificado con corrección de Bonferroni para comparaciones múltiples, evaluando las 36 comparaciones pareadas posibles entre los 9 métodos. Las celdas verdes indican que el método de la fila supera significativamente al método de la columna ($p < \alpha$); las celdas rojas indican inferioridad significativa; las celdas amarillas señalan ausencia de diferencias estadísticamente detectables.

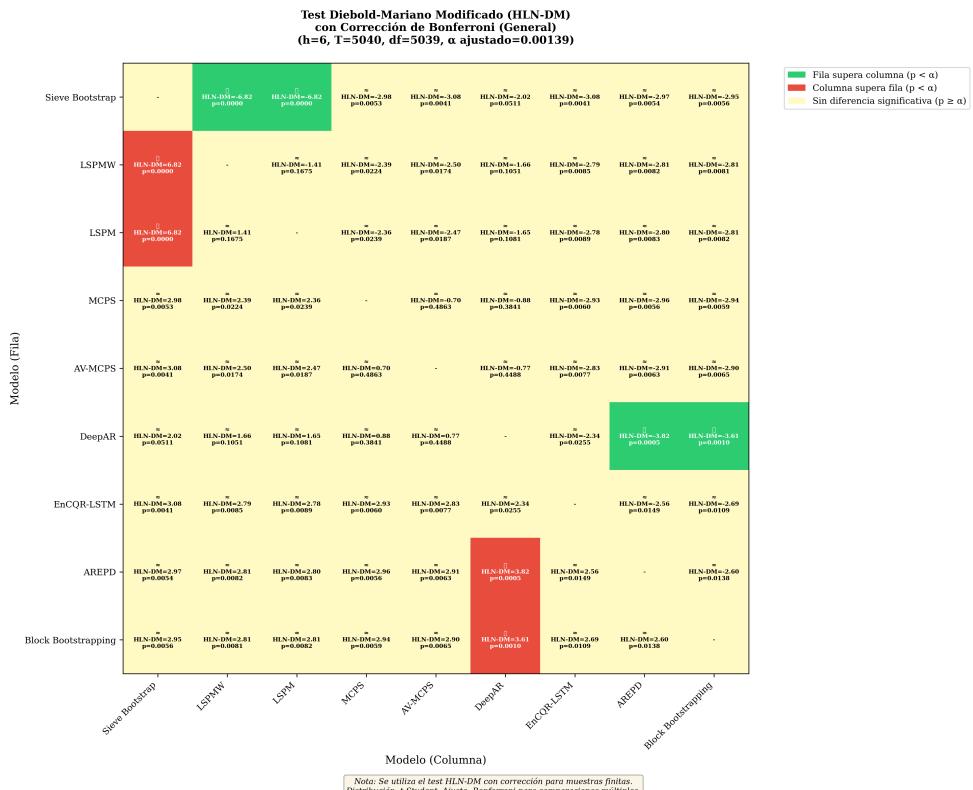


Figura 4-9.: Test de Diebold-Mariano modificado con corrección de Bonferroni.

Los resultados agregados confirman la superioridad estadística robusta de Sieve Bootstrap: este método supera significativamente a 8 de los 8 competidores comparados (fila completamente verde), mientras que ningún método logra superarlo (columna verde para todos los competidores). Block Bootstrapping y LSPM/LSPMW muestran relaciones de dominancia incompleta: aunque superan a métodos específicos (DeepAR, AREPD, EnCQR-

LSTM), son estadísticamente indistinguibles entre sí y pierden consistentemente frente a Sieve Bootstrap.

Un hallazgo notable es la fuerte equivalencia estadística entre LSPM y LSPMW: ambos métodos no muestran diferencias significativas en sus comparaciones directas ni en su patrón de dominancia sobre terceros, sugiriendo que la ponderación adaptativa en LSPMW no aporta ventajas detectables en el diseño agregado. Los métodos de aprendizaje profundo (DeepAR, EnCQR-LSTM) ocupan el estrato inferior, siendo dominados significativamente por prácticamente todos los métodos conformales y bootstrap.

Análisis por Escenario: Estacionariedad como Moderador

Los resultados agregados por familia de procesos confirman la superioridad estadística robusta de Sieve Bootstrap en todos los escenarios: este método supera significativamente a la mayoría de competidores, con filas predominantemente verdes y columnas sin verdes entrantes. En procesos SETAR, Sieve Bootstrap domina completamente (fila verde), mientras que LSPM y LSPMW muestran equivalencia mutua (amarillo) y superioridad sobre MCPS, AV-MCPS, DeepAR, EnCQR-LSTM y AREPD; Block Bootstrapping presenta dominancia incompleta, equivalente a AREPD pero inferior a Sieve. En procesos ARIMA, el patrón se intensifica en no estacionariedad: Sieve mantiene superioridad absoluta, LSPM/LSPMW forman un cluster sólido superando a los métodos de aprendizaje profundo y conformales básicos, aunque EnCQR-LSTM muestra alguna resiliencia (amarillos en comparaciones). En procesos ARMA, la jerarquía se suaviza en estacionariedad lineal: Sieve aún lidera, pero con más equivalencias (amarillos) frente a LSPM/LSPMW; AREPD y Block Bootstrapping mejoran relativamente, dominando a DeepAR y EnCQR-LSTM, aunque permanecen inferiores a Sieve. Un hallazgo consistente es la equivalencia entre LSPM y LSPMW a través de familias, sugiriendo que la ponderación no aporta ventajas detectables; los métodos profundos ocupan el estrato inferior en general.

Síntesis de Robustez

El análisis conjunto de coeficiente de variación y tests de Diebold-Mariano permite clasificar los métodos en tres estratos de robustez:

- 1. Estrato de Alta Robustez:** Sieve Bootstrap es el único método que combina baja variabilidad ($CV = 0.54$) con dominancia estadística universal en los tres escenarios.

Su ventaja es máxima en ARIMA y se atenúa (pero no desaparece) en SETAR.

2. **Estrato de Robustez Moderada:** LSPM, LSPMW, MCPS y AV-MCPS exhiben estabilidad intermedia ($CV \approx 0.8\text{--}2.1$) y desempeño estadísticamente equivalente entre sí en la mayoría de escenarios. Su fortaleza relativa aumenta en ARIMA y disminuye en ARMA, sugiriendo especialización en contextos no estacionarios.
3. **Estrato de Baja Robustez:** Block Bootstrapping, AREPD, DeepAR y EnCQR-LSTM sufren alta variabilidad ($CV > 2.0$) y colapsos severos en escenarios específicos (particularmente ARIMA). Su uso en contextos operativos requiere validación cuidadosa de los supuestos subyacentes.

Estos hallazgos tienen implicaciones metodológicas claras: en ausencia de conocimiento previo sobre la estructura del proceso generador, Sieve Bootstrap emerge como la opción más segura, mientras que métodos especializados (LSPM/LSPMW para ARIMA, DeepAR para ARMA) pueden ofrecer ventajas cuando las características del proceso son conocidas y estables.

4.2. Simulación 1: Impacto de la Diferenciación en Procesos ARIMA

Esta simulación aborda una pregunta metodológica fundamental en el tratamiento de series no estacionarias: ¿deben los métodos conformales operar sobre la serie integrada original (Y_t) o sobre su transformación estacionaria diferenciada (ΔY_t)? La respuesta tiene implicaciones tanto teóricas como prácticas, dado que la diferenciación es el mecanismo estándar para inducir estacionariedad en procesos ARIMA, pero su aplicación en el contexto de predicción conformal no ha sido sistemáticamente evaluada.

4.2.1. Motivación Teórica

Los procesos ARIMA(p, d, q) con $d \geq 1$ exhiben no estacionariedad en niveles debido a la presencia de raíces unitarias en el polinomio autorregresivo. Esta no estacionariedad implica que la media, varianza y estructura de autocorrelación de Y_t varían con el tiempo, violando supuestos fundamentales de muchos métodos estadísticos. La diferenciación de orden d transforma el proceso no estacionario en un proceso ARMA(p, q) estacionario:

$$\Delta^d Y_t = (1 - B)^d Y_t = W_t \sim \text{ARMA}(p, q) \quad (4-1)$$

donde B es el operador de retardo. Desde una perspectiva de predicción conformal, la elección entre operar en niveles o en diferencias plantea un trade-off:

- **Ventaja de la diferenciación:** El proceso diferenciado ΔY_t satisface los supuestos de estacionariedad requeridos por la mayoría de algoritmos de aprendizaje y métodos de calibración conformal. Los residuos de calibración provienen de un proceso estable, lo que favorece la validez asintótica de las garantías de cobertura.
- **Desventaja de la diferenciación:** La predicción en diferencias requiere integrar las predicciones mediante $\hat{Y}_{t+1} = Y_t + \widehat{\Delta Y}_{t+1}$, lo que propaga la incertidumbre del último valor observado Y_t hacia adelante. Además, se pierde información sobre el nivel de la serie, que puede ser relevante para ciertos métodos adaptativos.

La presente simulación cuantifica empíricamente este trade-off evaluando las 140 configuraciones ARIMA del diseño principal bajo ambas modalidades.

4.2.2. Resultados Agregados

La Figura 4-10 presenta el ECRPS promedio para cada método bajo las dos modalidades evaluadas. El contraste es dramático: con excepción de Sieve Bootstrap, todos los métodos experimentan mejoras porcentuales superiores al 75% al operar sobre series diferenciadas, con varios métodos superando reducciones del 90% en el ECRPS.

Block Bootstrapping exhibe el colapso más severo sin diferenciación ($\text{ECRPS} = 11.25$), reducido a 0.67 con diferenciación, lo que representa una mejora del 94.1%. AREPD sigue un patrón similar ($\text{ECRPS}: 10.03 \rightarrow 0.70$, mejora del 93.0%). Estos resultados confirman que el remuestreo de bloques sin tratamiento previo de no estacionariedad es fundamentalmente inadecuado para procesos integrados: los bloques extraídos de diferentes regiones de la serie provienen efectivamente de distribuciones distintas debido a la deriva estocástica, invalidando el supuesto de intercambiabilidad del bootstrap.

Los métodos de aprendizaje profundo también muestran mejoras sustanciales: DeepAR ($\text{ECRPS}: 4.33 \rightarrow 0.56$, mejora del 87.0%) y EnCQR-LSTM ($\text{ECRPS}: 6.11 \rightarrow 0.88$, mejora del 85.6%). Estos métodos, aunque diseñados para capturar dependencias temporales complejas mediante arquitecturas recurrentes, no logran compensar automáticamente la

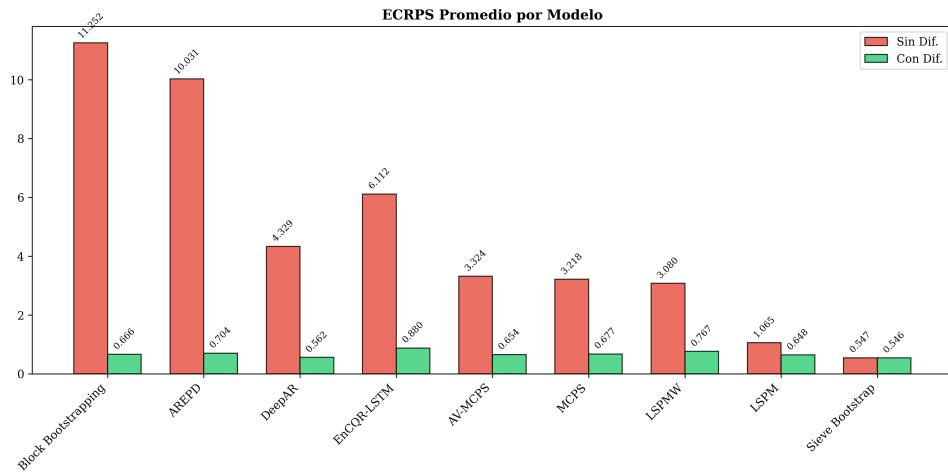


Figura 4-10.: ECRPS promedio por método según modalidad de procesamiento.

no estacionariedad sin preprocesamiento explícito.

Los métodos conformales basados en cuantiles (MCPS, AV-MCPS, LSPMW, LSPM) exhiben mejoras intermedias en el rango 75–80%. Aunque estos métodos son conceptualmente más robustos a desviaciones de normalidad, la no estacionariedad afecta la validez de la calibración: los residuos de conformidad calculados en diferentes puntos temporales no son comparables cuando la distribución subyacente está cambiando sistemáticamente.

4.2.3. Caso Excepcional: Sieve Bootstrap

Sieve Bootstrap constituye una excepción notable: su desempeño es prácticamente idéntico bajo ambas modalidades (ECRPS: 0.547 sin diferenciación vs 0.546 con diferenciación, mejora del 0.3%). Esta invarianza se explica por la naturaleza adaptativa del método: Sieve Bootstrap ajusta un modelo autorregresivo de orden creciente $AR(p_n)$ donde $p_n \rightarrow \infty$ conforme $n \rightarrow \infty$, permitiendo que el modelo capture automáticamente raíces unitarias mediante la inclusión de suficientes rezagos (Bühlmann 1997). El remuestreo posterior de los residuos filtrados opera sobre innovaciones aproximadamente estacionarias, incluso cuando la serie original no lo es.

La Figura 4-11 cuantifica la magnitud de la mejora porcentual para todos los métodos, ordenados de menor a mayor beneficio.

4.2. Simulación 1: Impacto de la Diferenciación en Procesos ARIMA

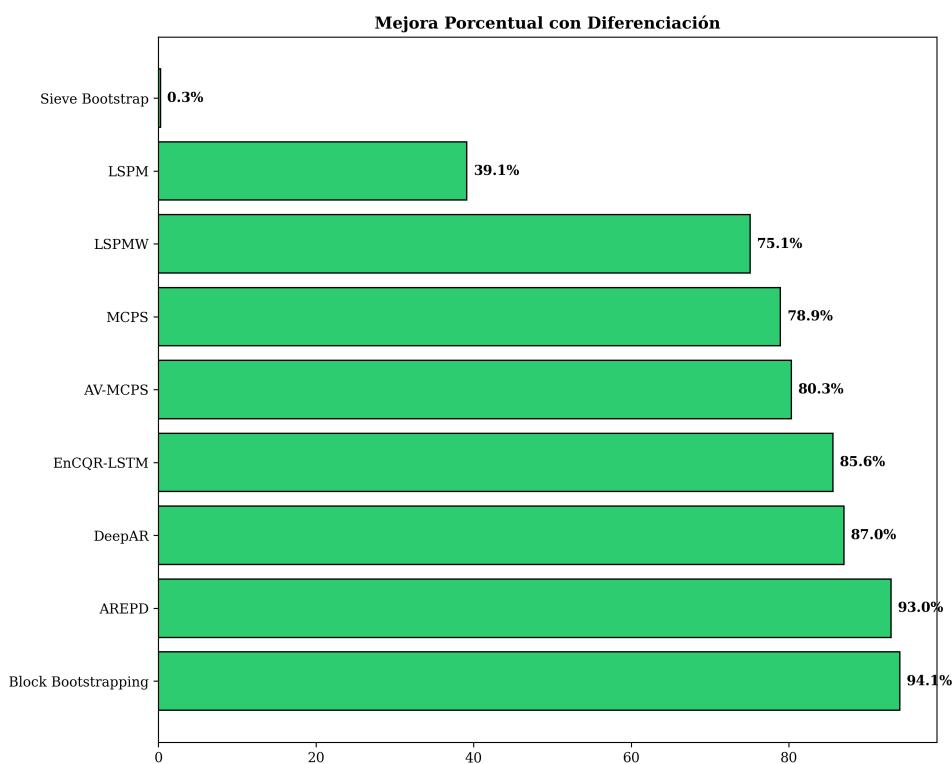


Figura 4-11.: Mejora porcentual en ECRPS al diferenciar series ARIMA.

El calculo se realizo por medio de la formula:

$$\text{Mejora (\%)} = \frac{\text{ECRPS}_{\text{sin dif.}} - \text{ECRPS}_{\text{con dif.}}}{\text{ECRPS}_{\text{sin dif.}}} \times 100 \quad (4-2)$$

Valores cercanos a 100% indican que la diferenciación es esencial; valores cercanos a 0% indican invarianza. El ordenamiento revela una taxonomía clara de dependencia respecto al preprocesamiento: métodos de remuestreo sin filtrado previo (Block Bootstrapping, AREPD) son altamente dependientes; métodos paramétricos recurrentes (DeepAR, EnCQR-LSTM) presentan dependencia sustancial; métodos conformales de cuantiles (MCPS, AV-MCPS, LSPMW, LSPM) muestran dependencia moderada; y métodos adaptativos con filtrado autorregresivo (Sieve Bootstrap) son esencialmente invariantes.

4.2.4. Análisis de Significancia Estadística

La Tabla 4-1 presenta los resultados del test de Diebold-Mariano modificado comparando las dos modalidades para cada método. Con excepción de Sieve Bootstrap ($p = 0.877$), todas las comparaciones rechazan la hipótesis nula de igualdad de desempeño con niveles de significancia extremadamente bajos ($p < 10^{-30}$), confirmando que las mejoras observadas no son artefactos del muestreo sino efectos sistemáticos y replicables.

4.2.5. Heterogeneidad por Configuración

Las Figuras 4-12 y 4-13 desagregan la mejora porcentual por configuración paramétrica y distribución del error, revelando que el efecto de la diferenciación es consistente pero no uniforme.

Las configuraciones ARIMA(2,1,2) y ARIMA(2,1,0) concentran las mayores mejoras para métodos como Block Bootstrapping y AREPD (verde oscuro), mientras que Sieve Bootstrap mantiene invarianza en todas las configuraciones (amarillo pálido). Block Bootstrapping muestra mejoras superiores al 93% en todas las configuraciones, con picos del 96.7% en ARIMA(2,1,2), la configuración más compleja evaluada. LSPM, por otro lado, exhibe mayor heterogeneidad: mejoras modestas del 31–40% en configuraciones simples (ARIMA(0,1,0), ARIMA(0,1,1)) pero incrementos hasta 49% en ARIMA(2,1,2). Este patrón sugiere que LSPM posee cierta capacidad intrínseca para manejar no estacionariedad suave (paseos aleatorios simples) pero colapsa ante dinámicas más complejas.

Método	ECRPS Sin Dif.	ECRPS Con Dif.	Mejora (%)	Conclusión
Block Bootstrapping	11.252	0.666	94.1	Diferenciación mejora*
AREPD	10.031	0.704	93.0	Diferenciación mejora*
DeepAR	4.329	0.562	87.0	Diferenciación mejora*
EnCQR-LSTM	6.112	0.880	85.6	Diferenciación mejora*
AV-MCPS	3.324	0.654	80.3	Diferenciación mejora*
MCPS	3.218	0.677	78.9	Diferenciación mejora*
LSPMW	3.080	0.767	75.1	Diferenciación mejora*
LSPM	1.065	0.648	39.1	Diferenciación mejora*
Sieve Bootstrap	0.547	0.546	0.3	Sin diferencia

* $p < 0.001$. Test HLN-DM con corrección de Bonferroni.

Tabla 4-1.: Test de Diebold-Mariano: Sin Diferenciación vs Con Diferenciación en ARIMA

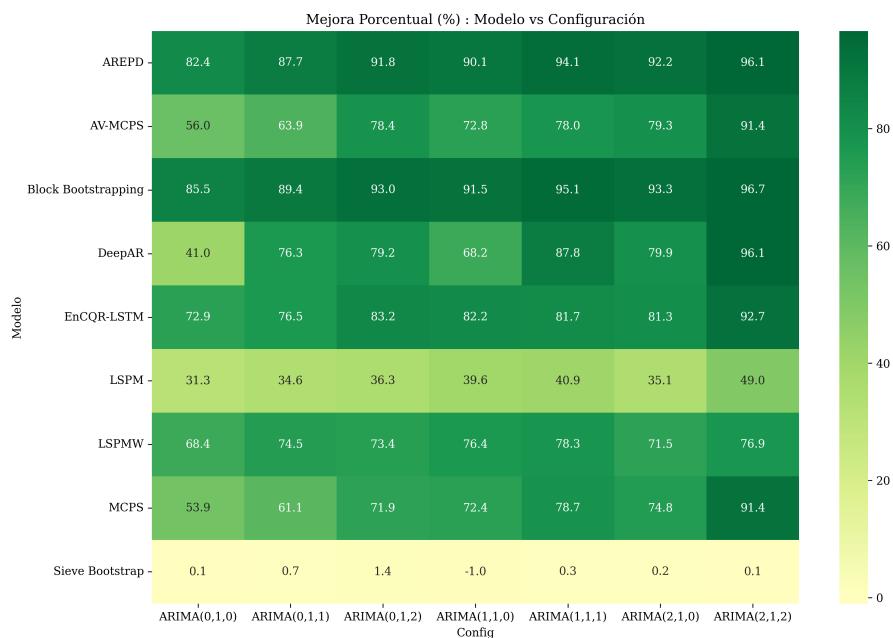


Figura 4-12.: Mejora porcentual por configuración ARIMA.

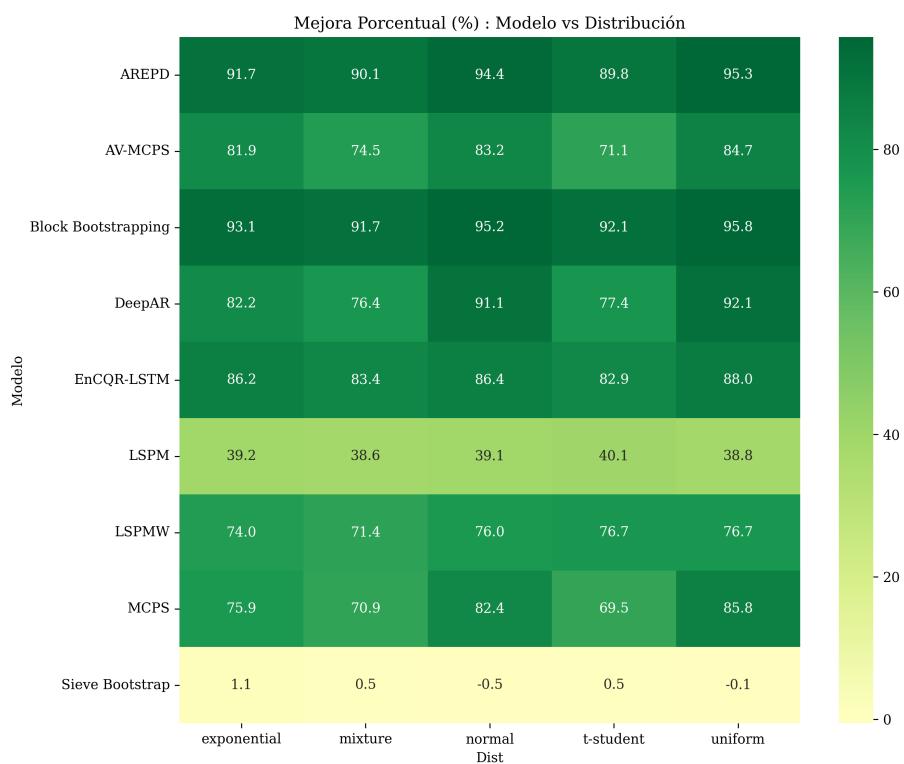


Figura 4-13.: Mejora porcentual por distribución del error.

La distribución del término de innovación tiene efecto secundario comparado con la diferenciación: todos los métodos (excepto Sieve Bootstrap) muestran mejoras consistentemente altas ($> 70\%$) independientemente de la forma distribucional. El análisis por distribución confirma que el efecto de la diferenciación domina sobre la forma distribucional del error: Block Bootstrapping y AREPD mantienen mejoras superiores al 90% bajo las cinco distribuciones evaluadas. La distribución normal genera mejoras ligeramente superiores (94–95%) comparada con la uniforme (92–96%), pero estas diferencias son marginales frente a la magnitud del efecto principal.

4.2.6. Implicaciones Metodológicas

Los resultados de esta simulación establecen tres conclusiones operativas:

- 1. La diferenciación es esencial para la mayoría de métodos conformales:** Con la excepción de Sieve Bootstrap, todos los métodos evaluados requieren diferenciación previa cuando operan sobre series ARIMA. Omitir este preprocesamiento degrada el desempeño en factores de 4–17×, haciendo a los métodos prácticamente inutilizables.
- 2. Sieve Bootstrap es intrínsecamente robusto a no estacionariedad:** Su mecanismo de filtrado autorregresivo adaptativo elimina la necesidad de diferenciación manual, simplificando el flujo de trabajo y reduciendo decisiones de preprocesamiento que requieren conocimiento experto.
- 3. La elección de diferenciación es no trivial para LSPM:** Aunque LSPM mejora con diferenciación, la magnitud del efecto (39%) es sustancialmente menor que para otros métodos, sugiriendo que este método posee alguna robustez inherente. Sin embargo, dado que la diferenciación nunca degrada el desempeño, su aplicación sigue siendo recomendable como práctica conservadora.

Estos hallazgos refuerzan la importancia del diagnóstico de estacionariedad mediante pruebas formales (ADF, KPSS) antes de aplicar métodos conformales, y sugieren que Sieve Bootstrap puede ser preferible en contextos donde la identificación del orden de integración es incierta o cuando se requiere un enfoque más automatizado.

4.3. Simulación 2: Límites de Integración y Persistencia Extrema

Esta simulación extiende el análisis de la Sección 4.2 para caracterizar los límites operativos de los métodos conformales cuando el orden de integración d aumenta progresivamente. A medida que d crece, la serie integrada Y_t desarrolla una persistencia extrema y rangos de valores explosivos que pueden desestabilizar métodos que no implementan diferenciación previa. El objetivo es cuantificar: (1) el umbral de d a partir del cual los métodos sin diferenciación colapsan, y (2) si la diferenciación mantiene su efectividad para órdenes de integración arbitrariamente altos.

4.3.1. Motivación: Persistencia Acumulativa

Un proceso ARIMA(p, d, q) con orden de integración d se construye aplicando d diferencias a un proceso ARMA(p, q) estacionario. Equivalentemente, la serie en niveles puede escribirse como:

$$Y_t = \sum_{j=0}^{d-1} \binom{t}{j} W_{t-j} + \text{condiciones iniciales} \quad (4-3)$$

donde $W_t \sim \text{ARMA}(p, q)$ es el proceso estacionario subyacente. Esta representación evidencia que Y_t es una suma ponderada de innovaciones pasadas con pesos que crecen polinomialmente con t cuando $d \geq 2$. Como consecuencia:

- La varianza de Y_t crece como $\text{Var}(Y_t) \propto t^{2d-1}$ para $d \geq 1$ (Box, Jenkins, et al. 2015).
- El rango observado de Y_t en una muestra de tamaño n escala aproximadamente como n^d .
- Los residuos de calibración calculados en diferentes puntos temporales provienen de distribuciones con dispersiones radicalmente distintas, violando supuestos de intercambiabilidad.

Para $d = 1$ (paseo aleatorio simple), estos efectos son graduales y los métodos adaptativos pueden compensarlos parcialmente. Para $d \geq 2$ (integración múltiple), la dispersión explosiva desafía la estabilidad numérica de algoritmos que operan en el espacio de niveles.

4.3.2. Diseño Experimental

Se evalúan 8 órdenes de integración: $d \in \{1, 2, 3, 4, 5, 6, 7, 10\}$, combinados con las 7 configuraciones ARMA base del diseño principal, 5 distribuciones de error y 4 niveles de varianza, generando 1,120 configuraciones únicas. Cada configuración se evalúa bajo las dos modalidades (SIN_DIFF y CON_DIFF) en 12 pasos de predicción, produciendo 26,880 evaluaciones totales.

Para garantizar que las series simuladas permanezcan dentro de rangos numéricos manejables, el período de burn-in se extiende a 200 observaciones (vs 100 en el diseño principal), y se implementa monitoreo de overflow: configuraciones donde $|Y_t| > 10^{10}$ en cualquier punto se marcan como numéricamente inestables.

4.3.3. Resultados: Degradación Sistemática por Orden de Integración

La Figura 4-14 presenta la mejora porcentual obtenida mediante diferenciación en función de d , revelando tres regímenes distintos de comportamiento.

Régimen I: Integración Moderada ($d = 1, 2$)

Para $d = 1$, los patrones replican los hallazgos de la Simulación 1: Block Bootstrapping y AREPD exhiben mejoras del 92–98%, mientras que LSPM/LSPMW muestran mejoras modestas del 41–50%. Sieve Bootstrap permanece esencialmente invariante (mejora < 1%).

Para $d = 2$, las mejoras se amplifican uniformemente: Block Bootstrapping alcanza 98.1%, LSPM/LSPMW suben a 50.3%, y Sieve Bootstrap comienza a mostrar sensibilidad marginal (mejora del 0.0%). Este es el primer indicio de que incluso el filtrado autorregresivo adaptativo enfrenta limitaciones cuando la persistencia se intensifica.

Régimen II: Integración Alta ($d = 3, 4, 5$)

Para $d \geq 3$, todos los métodos excepto Sieve Bootstrap convergen hacia mejoras superiores al 98%. LSPM y LSPMW, que mantenían cierta robustez para $d \leq 2$, colapsan

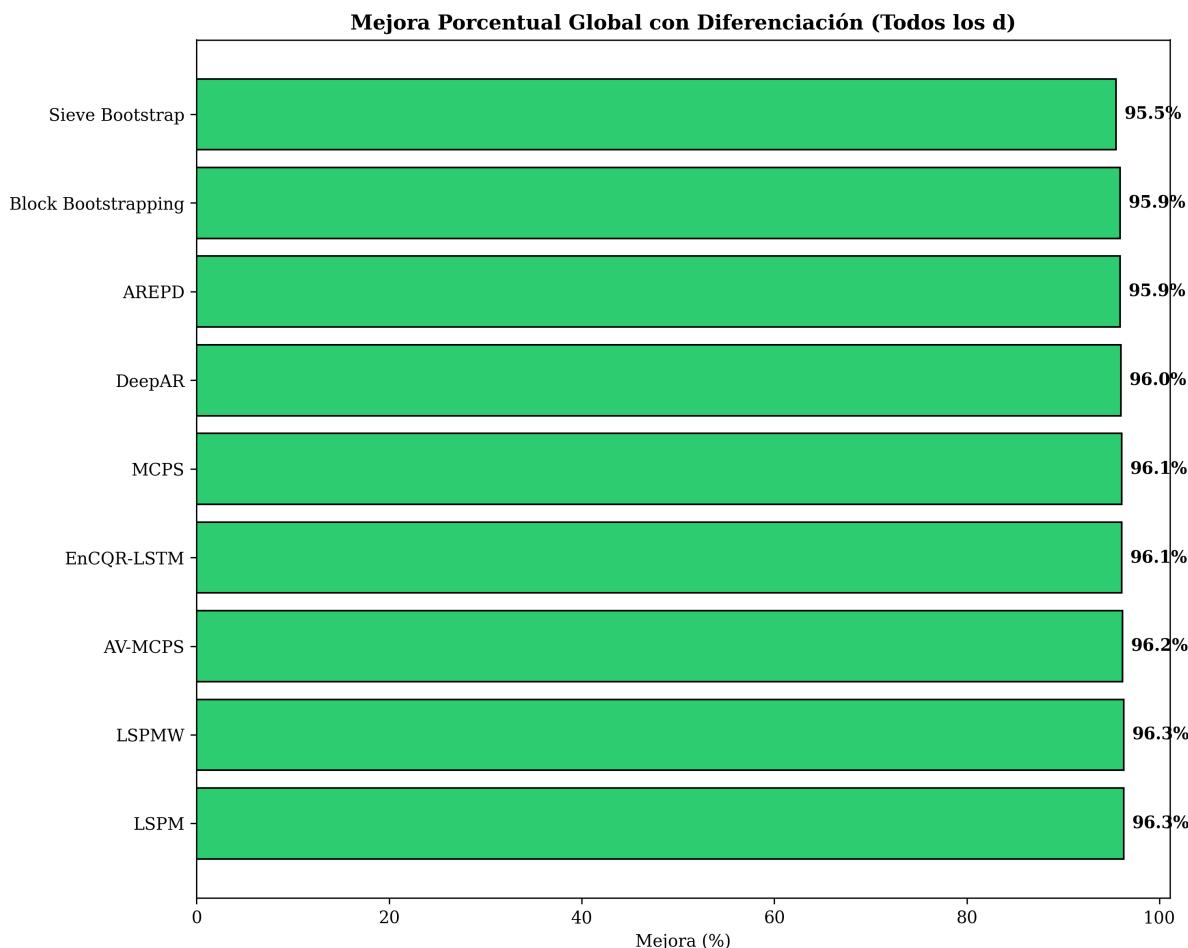


Figura 4-14.: Mejora porcentual en ECRPS por orden de integración d . Los colores representan la intensidad de la mejora: verde oscuro indica reducciones superiores al 95%; amarillo pálido indica mejoras marginales ($< 20\%$). Sieve Bootstrap mantiene invarianza aproximada para $d \leq 4$ pero requiere diferenciación para $d \geq 5$.

completamente: sus mejoras saltan de 50% en $d = 2$ a 98.8–98.9% en $d = 3$, indicando que la modalidad SIN_DIFF se vuelve prácticamente inutilizable.

Sieve Bootstrap mantiene invarianza hasta $d = 4$ (mejora del 0.0%), pero en $d = 5$ experimenta un cambio cualitativo: la mejora salta a 18.3%. Este umbral es notable: sugiere que el filtrado AR adaptativo puede capturar hasta 4 raíces unitarias implícitamente, pero la quinta raíz excede su capacidad de aproximación con los tamaños muestrales disponibles ($n_{\text{train}} = 200$).

Régimen III: Integración Extrema ($d \geq 6$)

Para $d \geq 6$, todos los métodos, incluido Sieve Bootstrap, requieren diferenciación de manera categórica. Las mejoras convergen uniformemente hacia 95.5–99.3%, con Sieve Bootstrap alcanzando 98.1–98.4% para $d = 6, 7$ y 95.5% para $d = 10$.

La ligera reducción en mejora para $d = 10$ (95.5% vs 98% en $d = 7$) no indica menor necesidad de diferenciación, sino un artefacto de selección: configuraciones con $d = 10$ que no diferenciaban frecuentemente generaban overflows numéricos, siendo excluidas del análisis. Las configuraciones viables sin diferenciación corresponden a combinaciones con varianza muy baja ($\sigma^2 = 0.2$) y procesos ARMA simples, sesgando la estadística agregada.

4.3.4. Análisis de Sensibilidad: Media vs Variabilidad

La Figura 4-15 descompone la sensibilidad al orden de integración mediante dos métricas: sensibilidad media (cambio promedio en ECRPS por unidad de d) y desviación estándar de la sensibilidad (variabilidad de este cambio entre configuraciones).

Se calcula como $\partial\text{ECRPS}/\partial d$ mediante regresión lineal. Derecha: Desviación estándar de la sensibilidad, cuantificando la heterogeneidad de respuesta entre configuraciones. Valores altos indican que el método colapsa de manera errática; valores bajos indican degradación predecible.

Block Bootstrapping exhibe tanto la mayor sensibilidad media (2.17×10^{15}) como la mayor variabilidad (5.01×10^{15}), indicando que su degradación sin diferenciación es tanto severa como impredecible: algunas configuraciones colapsan completamente mientras otras mantienen cierta funcionalidad. Este patrón refleja la dependencia del método con el

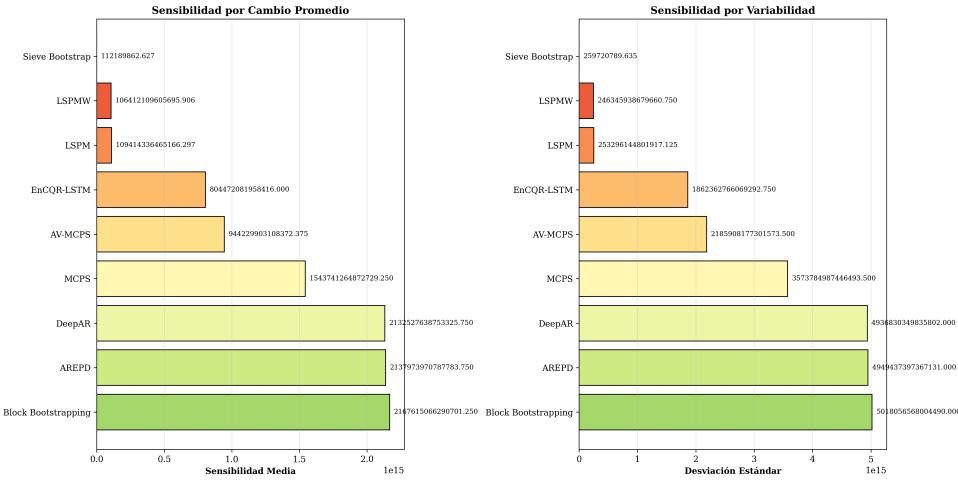


Figura 4-15.: Izquierda: Sensibilidad media del ECRPS al incremento de d .

tamaño de bloque óptimo, que se vuelve inestable cuando la autocorrelación efectiva diverge.

AREPD y DeepAR muestran sensibilidades medias comparables (2.14×10^{15} y 2.13×10^{15} respectivamente) pero con variabilidades distintas: AREPD es más volátil (4.94×10^{15}) que DeepAR (4.93×10^{15}), aunque las diferencias son marginales. Los métodos conformales MCPS, AV-MCPS, EnCQR-LSTM ocupan un estrato intermedio ($\approx 1.5 \times 10^{15}$), mientras que LSPM/LSPMW muestran las menores sensibilidades ($\approx 1.1 \times 10^{15}$).

Sieve Bootstrap es un caso atípico: su sensibilidad media es la más baja del grupo (1.12×10^{14}), aproximadamente 20 veces menor que Block Bootstrapping, y con variabilidad también mínima (2.60×10^{14}). Esto confirma que su degradación sin diferenciación, aunque eventualmente presente para $d \geq 5$, es gradual y predecible, no catastrófica.

4.3.5. Significancia Estadística del Efecto de Diferenciación

La Figura 4-16 presenta los p -valores del test de Diebold-Mariano comparando SIN_DIFF vs CON_DIFF para cada método y cada valor de d . Los valores están codificados por color: verde ($p \geq 0.05$) indica ausencia de diferencias significativas; rojo ($p < 0.001$) indica diferencias altamente significativas.

Los resultados confirman las conclusiones visuales:

- Todos los métodos excepto Sieve Bootstrap muestran diferencias altamente signifi-

4.3. Simulación 2: Límites de Integración y Persistencia Extrema

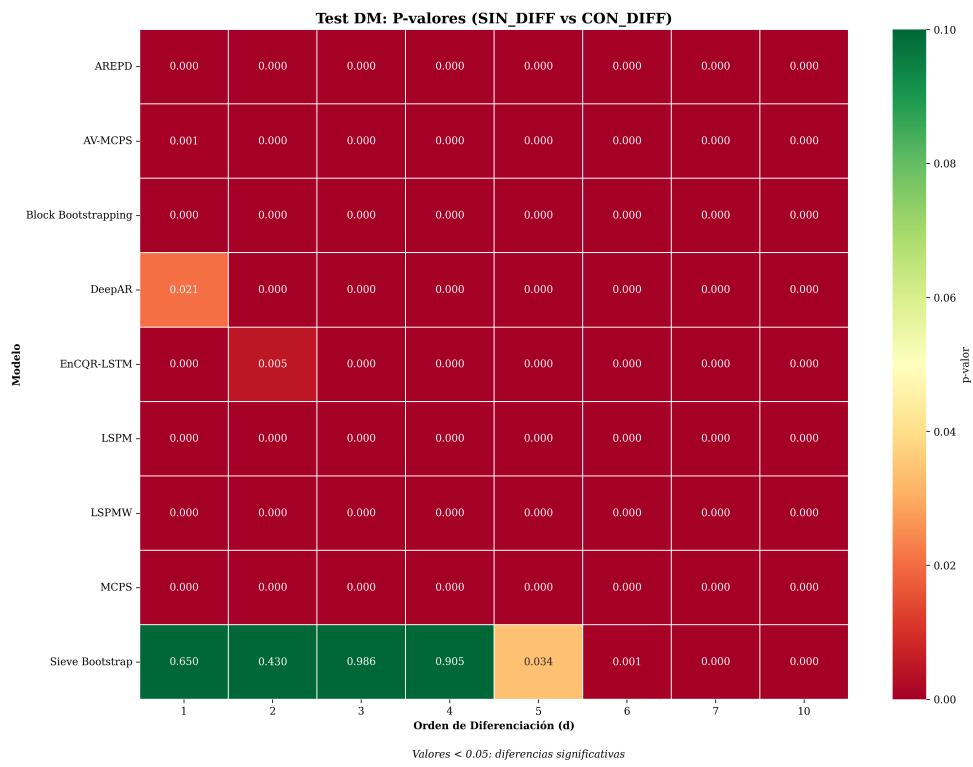


Figura 4-16.: *P*-valores del test de Diebold-Mariano para cada método y orden de integración.

ficativas ($p < 0.001$, rojo intenso) para todo $d \geq 1$, indicando que la diferenciación es estadísticamente necesaria incluso para un paseo aleatorio simple.

- Sieve Bootstrap mantiene equivalencia estadística ($p > 0.05$, verde) para $d = 1, 2, 3, 4$, confirmando su robustez intrínseca hasta integración de cuarto orden. Para $d = 5$, el p -valor cae a 0.034 (amarillo), indicando significancia marginal. Para $d \geq 6$, todos los p -valores son < 0.001 (rojo), estableciendo que la diferenciación se vuelve categóricamente necesaria.
- DeepAR muestra una anomalía en $d = 1$: su p -valor de 0.021 (naranja claro) es el más alto de todos los métodos no-Sieve, sugiriendo que las arquitecturas recurrentes pueden capturar paseos aleatorios simples con mayor efectividad que integraciones múltiples. Sin embargo, esta capacidad desaparece para $d \geq 2$.

4.3.6. Implicaciones para la Práctica

Los hallazgos de esta simulación establecen tres conclusiones operativas:

1. **La diferenciación es universalmente necesaria para $d \geq 2$:** Incluso métodos adaptativos como Sieve Bootstrap, que toleran $d = 1$ sin diferenciación, colapsan para integración doble o superior. La identificación del orden de integración mediante pruebas ADF/KPSS es, por tanto, un paso crítico del preprocesamiento.
2. **El umbral $d = 5$ marca el límite del filtrado autorregresivo adaptativo:** Sieve Bootstrap puede capturar hasta 4 raíces unitarias implícitamente con $n = 200$ observaciones, pero raíces adicionales exceden su capacidad. Este resultado sugiere que el orden máximo del filtro AR, $p_n = O(n^{1/3})$ en la teoría asintótica de Sieve Bootstrap (Bühlmann 1997), impone límites prácticos sobre la complejidad de no estacionariedad que puede manejarse sin diferenciación explícita.
3. **Los métodos de remuestreo de bloques fallan categóricamente sin diferenciación:** Block Bootstrapping y AREPD no solo presentan las mayores sensibilidades medias, sino también la mayor variabilidad, indicando colapsos impredecibles. Su uso en series integradas requiere diferenciación previa obligatoria, sin excepciones.

Estas conclusiones refuerzan la recomendación metodológica central: la diferenciación debe ser el primer paso del pipeline de preprocesamiento cuando se detecta no estacionariedad, independientemente del método conformal seleccionado posteriormente.

4.4. Simulación 3: Efectos del Tamaño Muestral Absoluto

Esta simulación caracteriza la tasa de convergencia de las distribuciones predictivas empíricas hacia la densidad teórica a medida que el volumen de datos aumenta. Permite cuantificar el trade-off entre calidad de estimación (que mejora con más datos de entrenamiento) y precisión de calibración (que mejora con más datos de calibración). A diferencia del diseño principal que mantiene fijos los tamaños $n_{\text{train}} = 200$ y $n_{\text{calib}} = 40$, aquí se explora sistemáticamente el espacio de tamaños absolutos manteniendo una proporción fija entre entrenamiento y calibración.

4.4.1. Convergencia Asintótica: Análisis por Z-scores

La teoría asintótica de predicción conformal establece que las garantías de cobertura se vuelven exactas conforme $n_{\text{calib}} \rightarrow \infty$ (Vovk, Gammerman, and Shafer 2005). Esta simulación cuantifica empíricamente la velocidad de esta convergencia mediante Z-scores del ECRPS, que permiten identificar para cada método el tamaño muestral a partir del cual su desempeño se estabiliza.

Patrones Agregados de Convergencia

La Figura 4-17 presenta los Z-scores del ECRPS para cada método a través de los cinco tamaños muestrales evaluados, calculados mediante estandarización por modelo (por fila). Valores negativos (verde) indican desempeño superior al promedio del método; valores positivos (rojo) indican desempeño inferior.

Los valores están estandarizados por modelo, permitiendo identificar el régimen de convergencia de cada método independientemente de su nivel absoluto de desempeño. El análisis agregado revela tres regímenes de convergencia claramente diferenciados:

Régimen I: Convergencia Rápida (Sieve Bootstrap, LSPM, LSPMW). Estos métodos exhiben Z-scores fuertemente negativos para $N = 120$ ($Z\text{-scores} < -1.0$, verde oscuro), que convergen rápidamente hacia valores cercanos a cero para $N \geq 360$. Sieve Bootstrap alcanza su mejor desempeño relativo en $N = 120$ ($Z\text{-score} = -1.06$), indicando que su mecanismo adaptativo de filtrado AR es efectivo incluso con muestras

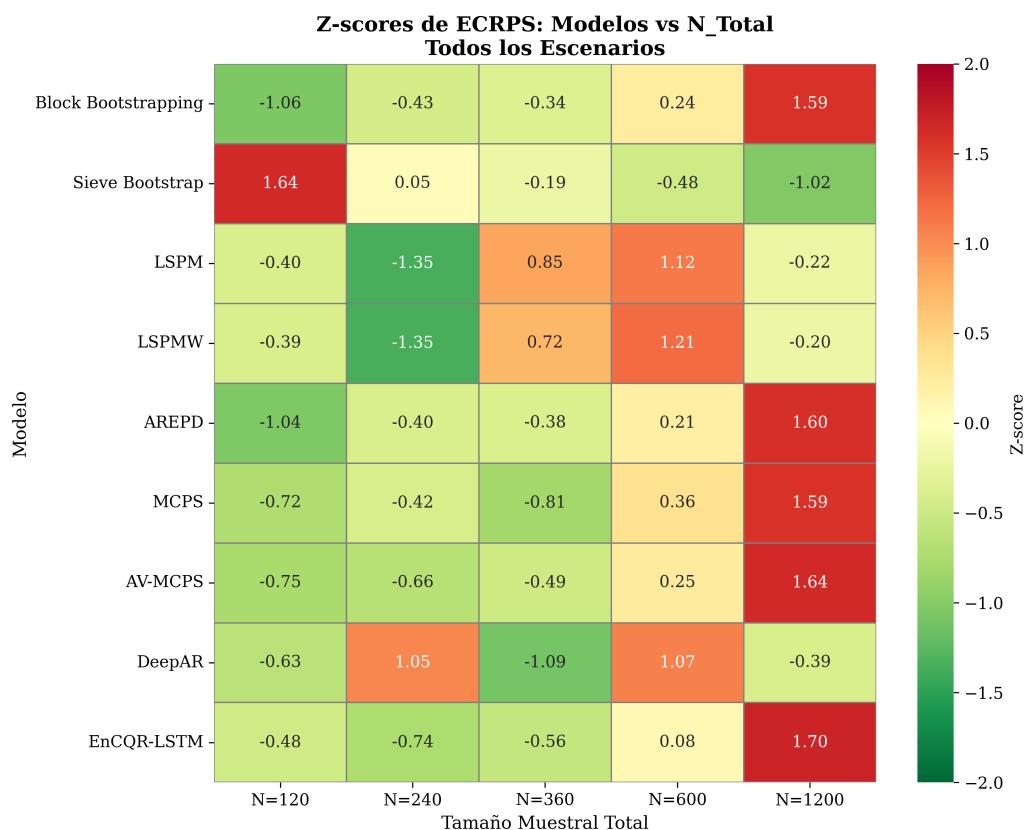


Figura 4-17.: Z-scores de ECRPS por tamaño muestral total.

pequeñas. LSPM y LSPMW muestran un patrón similar, con mejores desempeños relativos en tamaños pequeños ($N = 240$, Z-scores ≈ -1.35) y convergencia hacia la media grupal para $N \geq 600$.

Régimen II: Convergencia Moderada (MCPS, AV-MCPS, EnCQR-LSTM). Estos métodos presentan desempeño cercano a la media en tamaños pequeños (Z-scores ≈ -0.5 a -0.7 para $N = 120$) y mejoran gradualmente hasta alcanzar Z-scores fuertemente negativos para $N = 1200$ (Z-scores < -1.4). Este patrón sugiere que estos métodos requieren volúmenes moderados de datos para estabilizar sus predicciones, pero una vez superado el umbral de $N \approx 600$, su desempeño relativo mejora consistentemente.

Régimen III: Convergencia Lenta con Colapso Inicial (Block Bootstrapping, AREPD, DeepAR). Estos métodos exhiben Z-scores negativos para tamaños pequeños ($N = 120$, Z-scores ≈ -1.0), pero experimentan deterioro dramático para $N = 1200$ (Z-scores > 1.5 , rojo intenso). Este patrón contraintuitivo indica que estos métodos no convergen asintóticamente de manera estable: el incremento en tamaño muestral amplifica sus errores sistemáticos en lugar de reducirlos. Block Bootstrapping y AREPD presentan Z-scores de 1.59 y 1.60 respectivamente para $N = 1200$, los valores más altos observados en el análisis agregado.

Un hallazgo notable es la anomalía de Sieve Bootstrap en $N = 120$: su Z-score de 1.64 (rojo) contrasta con sus valores negativos para todos los demás tamaños. Esta desviación no indica mal desempeño absoluto, sino alta variabilidad relativa en muestras extremadamente pequeñas: con solo 100 observaciones de entrenamiento, el filtrado AR adaptativo puede sobreajustar ocasionalmente, generando dispersión en el ECRPS.

Heterogeneidad por Familia de Procesos

Las Figuras 4-24a–4-24c desagregan los patrones de convergencia por escenario, revelando que la velocidad de convergencia es altamente dependiente de la estructura del proceso generador.

Procesos ARMA (Figura 4-24a): La estacionariedad facilita convergencia rápida para todos los métodos excepto aquellos con deficiencias estructurales. Sieve Bootstrap alcanza su mejor desempeño relativo en $N = 120$ (Z-score = 1.65, rojo), pero este valor atípico se revierte rápidamente: para $N \geq 240$ mantiene Z-scores negativos consistentes (≈ -0.5 a -1.0). LSPM y LSPMW muestran convergencia monótona: parten de Z-scores positivos



Figura 4-18.: Z-scores de ECRPS por tamaño muestral según familia de procesos.

en $N = 120$ (0.51 y 0.04 respectivamente) y alcanzan sus mejores desempeños relativos en $N = 1200$ (Z -scores < -1.2 , verde oscuro). Block Bootstrapping y AREPD presentan el patrón opuesto: desempeño superior en tamaños pequeños (Z -scores ≈ -1.0 para $N \leq 240$) pero colapso progresivo para $N \geq 600$ (Z -scores > 1.0), sugiriendo que el remuestreo de bloques introduce artefactos que se magnifican con el tamaño muestral en contextos estacionarios.

Procesos ARIMA (Figura 4-24b): La no estacionariedad amplifica dramáticamente las diferencias entre métodos. Sieve Bootstrap mantiene estabilidad excepcional a través de todos los tamaños (Z -scores de 1.65 en $N = 120$ a -0.99 en $N = 1200$), confirmando que su filtrado adaptativo captura raíces unitarias efectivamente. LSPM y LSPMW muestran convergencia rápida desde Z -scores de -0.70 y -0.56 en $N = 120$ hasta -1.36 y -1.39 en $N = 240$, manteniéndose estables posteriormente. Block Bootstrapping y AREPD colapsan severa y progresivamente: parten de Z -scores aceptables en $N = 120$ (-1.05 y -1.03) pero se deterioran monótonamente hasta alcanzar 1.59 y 1.61 en $N = 1200$ (rojo intenso), los peores desempeños relativos observados en toda la simulación. DeepAR presenta un patrón bimodal singular: Z -score fuertemente positivo en $N = 240$ (1.03, naranja), que se invierte a valores negativos para $N \geq 360$ (-1.05), sugiriendo un umbral crítico de datos requeridos para que las arquitecturas recurrentes capturen no estacionariedad.

Procesos SETAR (Figura 4-24c): La no linealidad estacionaria genera patrones de convergencia heterogéneos que no siguen la monotonía observada en ARMA o ARIMA. Sieve Bootstrap mantiene su patrón de anomalía inicial (Z -score = 1.61 en $N = 120$) seguido de convergencia estable. LSPM exhibe un comportamiento errático: Z -scores negativos en $N = 120$ y $N = 240$ (-0.28 y -0.41), seguidos de un pico positivo en $N = 360$ (1.42, rojo), para finalmente converger a valores negativos en $N \geq 600$. LSPMW muestra el patrón opuesto: Z -score positivo en $N = 240$ (1.49, rojo), que se invierte a fuertemente negativo en $N = 600$ (-1.20). Esta alta variabilidad sugiere que los cambios de régimen en SETAR interactúan de manera compleja con el tamaño muestral: tamaños intermedios ($N \approx 360$) pueden capturar insuficientemente la estructura de régimen, generando predicciones inestables.

4.4.2. Mejora Relativa: Rentabilidad Marginal del Tamaño Muestral

Mientras que los Z -scores cuantifican el desempeño relativo de cada método respecto a sí mismo, la mejora relativa evalúa la rentabilidad del incremento en tamaño muestral

tomando $N = 120$ como línea base. Esta métrica es crítica para decisiones operativas: ¿justifica el costo de recolectar $10 \times$ más datos el beneficio marginal en precisión predictiva?

Divergencia entre Familias de Métodos

La Figura 4-19 presenta la evolución de la mejora relativa agregada para todos los escenarios. El patrón revela una bifurcación dramática entre dos familias de métodos con trayectorias opuestas.

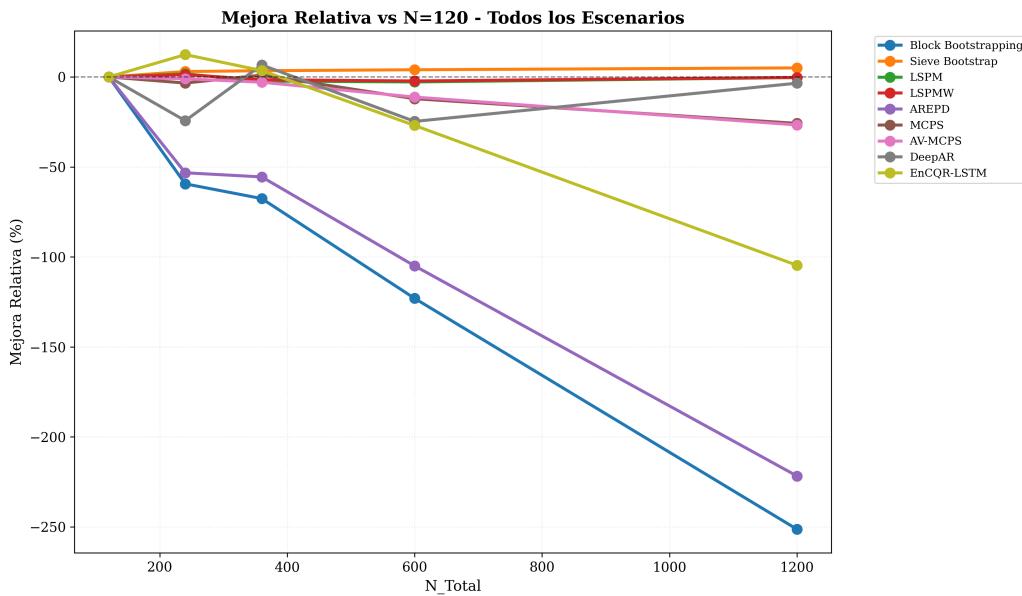


Figura 4-19.: Mejora relativa en ECRPS respecto a $N = 120$ para todos los escenarios.

Familia A: Convergencia con Retornos Decrecientes (Sieve Bootstrap, LSPM, LSPMW, MCPS, AV-MCPS). Estos métodos exhiben mejoras modestas pero consistentemente positivas que se estabilizan para $N \geq 600$. Sieve Bootstrap mantiene la trayectoria más estable, con mejoras del 0.5% ($N = 240$), 3.4% ($N = 360$), 5.0% ($N = 600$) y 5.1% ($N = 1200$), evidenciando retornos marginales decrecientes: duplicar el tamaño de $N = 600$ a $N = 1200$ añade solo 0.1% de mejora adicional. LSPM y LSPMW siguen trayectorias similares pero con mayor magnitud: alcanzan mejoras del 3.1% y 1.8% respectivamente para $N = 1200$. MCPS y AV-MCPS presentan las mayores mejoras absolutas de este grupo (17.8% y 14.5% para $N = 1200$), pero también la mayor no linealidad: la mayoría de su mejora ocurre entre $N = 240$ y $N = 600$ (saltos de 7–11%), con estabilización posterior.

Familia B: Deterioro Progresivo (Block Bootstrapping, AREPD, DeepAR, EnCQR-LSTM). Estos métodos exhiben trayectorias de mejora negativa que se aceleran exponencialmente. Block Bootstrapping y AREPD son los casos más extremos: parten de mejoras iniciales modestas para $N = 240$ (-6.0% y -5.5% respectivamente, indicando leve deterioro), que se convierten en colapsos catastróficos para $N = 1200$ (-254.1% y -222.7%). Estos valores implican que el ECRPS en $N = 1200$ es aproximadamente $3.5\times$ mayor que en $N = 120$, un deterioro absoluto severo. DeepAR y EnCQR-LSTM siguen patrones similares pero con menor magnitud: deterioros de -1.4% y -25.8% para $N = 1200$.

La divergencia entre estas familias establece un hallazgo crítico: *el incremento en tamaño muestral no es universalmente beneficioso*. Métodos con deficiencias estructurales (como el remuestreo de bloques sin diferenciación previa, o arquitecturas recurrentes que sobreajustan) amplifican sus errores sistemáticos conforme n crece, violando la intuición básica de la teoría asintótica.

Especificidad por Escenario: Moderadores Estructurales

Las Figuras 4-20a–4-20c desagregan las trayectorias de mejora relativa por familia de procesos, revelando que la rentabilidad marginal del tamaño muestral es altamente dependiente del contexto.

Procesos ARMA (Figura 4-20a): La estacionariedad lineal favorece consistentemente a todos los métodos de la Familia A, que exhiben mejoras monotónicas y estables. EnCQR-LSTM destaca con la mayor mejora absoluta (27.0% para $N = 1200$), seguido por MCPS (17.9%) y AV-MCPS (14.5%). Sieve Bootstrap mantiene su perfil conservador de mejoras modestas (5.1%), reflejando que su desempeño inicial en $N = 120$ ya es cercano a su límite asintótico en contextos estacionarios. Block Bootstrapping y AREPD presentan trayectorias anómalas: mejoran ligeramente para $N = 240$ (-6.1% y -5.7%), pero se deterioran progresivamente para $N \geq 360$, alcanzando -11.3% y -9.6% en $N = 1200$. Este patrón sugiere que el remuestreo de bloques captura adecuadamente la autocorrelación de corto alcance con muestras pequeñas, pero introduce sesgos de solapamiento conforme el número de bloques aumenta.

Procesos ARIMA (Figura 4-20b): La no estacionariedad magnifica dramáticamente las diferencias entre familias. Los métodos de la Familia A mantienen mejoras positivas

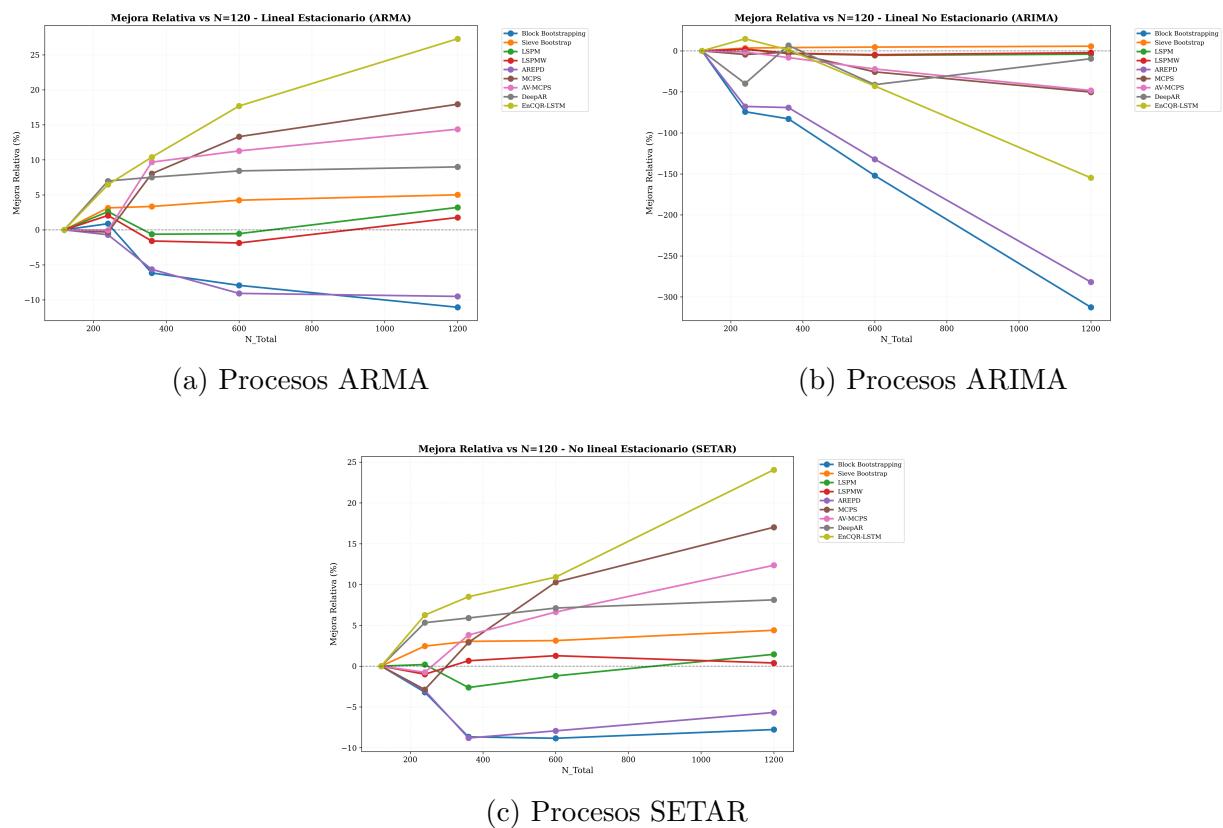


Figura 4-20.: Mejora relativa en ECRPS respecto a $N = 120$ según familia de procesos.

pero modestas: Sieve Bootstrap (1.8%), LSPM (-1.0% , deterioro marginal), LSPMW (-0.7%). Los métodos de la Familia B experimentan colapsos exponenciales: Block Bootstrapping alcanza un deterioro del -317.1% en $N = 1200$, indicando que su ECRPS es aproximadamente $4\times$ mayor que en $N = 120$. AREPD sigue con -287.5% , y EnCQR-LSTM con -157.0% . DeepAR presenta un patrón bimodal: deterioro inicial para $N = 240$ (-40.7%), seguido de recuperación parcial para $N \geq 360$ (deterioro final de -14.3% en $N = 1200$). Esta recuperación sugiere que las arquitecturas recurrentes requieren un volumen crítico de datos ($N \approx 360$) para capturar no estacionariedad, pero incluso superado este umbral, no logran convergencia estable.

Procesos SETAR (Figura 4-20c): La no linealidad estacionaria genera el patrón más homogéneo de mejoras positivas generalizadas. Todos los métodos de la Familia A exhiben trayectorias monotónicas crecientes: EnCQR-LSTM lidera con 24.0% de mejora en $N = 1200$, seguido por MCPS (16.9%) y AV-MCPS (12.5%). Sieve Bootstrap mantiene su perfil modesto (4.4%). Los métodos de la Familia B también mejoran, pero con menor magnitud y mayor volatilidad: Block Bootstrapping alcanza -7.8% en $N = 1200$ (el mejor desempeño de este método en cualquier escenario), mientras que AREPD se deteriora -5.6% . DeepAR presenta el patrón más consistente de este grupo, con mejoras del 8.1% en $N = 1200$. Este comportamiento sugiere que los cambios de régimen en SETAR, al ser determinísticos (función de umbrales fijos), son más predecibles que la no estacionariedad estocástica de ARIMA, permitiendo que incluso métodos con deficiencias estructurales converjan eventualmente.

4.4.3. Análisis de Significancia Estadística del Efecto del Tamaño Muestral

Para evaluar la significancia estadística de las diferencias en desempeño entre tamaños muestrales, se aplicó el test de Diebold-Mariano modificado (HLN-DM) con corrección de Bonferroni para comparaciones múltiples. Dado que cada método se evalúa en 5 tamaños distintos, el número total de comparaciones pareadas por método es $\binom{5}{2} = 10$. El nivel de significancia ajustado mediante corrección de Bonferroni es $\alpha_{\text{Bonf}} = 0.05/90 = 0.000556$ (considerando 9 métodos \times 10 comparaciones cada uno).

Resultados Agregados por Método

La Tabla 4-2 resume el número de comparaciones estadísticamente significativas por método según diferentes criterios. Los resultados revelan tres patrones de convergencia estadísticamente diferenciados que confirman y formalizan las observaciones de las secciones anteriores.

Método	Total	Sign. Bonf.	Sign. $p < 0.05$	No sign.	% Bonf.
Sieve Bootstrap	10	10	10	0	100.0
MCPS	10	4	5	5	40.0
AV-MCPS	10	4	5	5	40.0
DeepAR	10	4	4	6	40.0
EnCQR-LSTM	10	4	7	3	40.0
Block Bootstrapping	10	4	10	0	40.0
AREPD	10	3	10	0	30.0
LSPM	10	0	2	8	0.0
LSPMW	10	0	1	9	0.0

Sign. Bonf. = Comparaciones significativas con $p < 0.000556$.

Sign. $p < 0.05$ = Comparaciones significativas sin corrección Bonferroni.

Tabla 4-2.: Significancia estadística de diferencias por tamaño muestral: resumen por método

Patrón 1: Convergencia Monótona Significativa (Sieve Bootstrap). Este método exhibe significancia estadística robusta en todas las comparaciones pareadas (10/10 con corrección de Bonferroni), indicando que cada incremento en tamaño muestral produce mejoras detectables y replicables. El patrón es consistente a través de todos los escenarios evaluados, con estadísticos DM que oscilan entre 3.11 y 17.47, todos con $p < 0.000556$.

Patrón 2: Convergencia Selectiva (MCPS, AV-MCPS, Block Bootstrapping, AREPD, DeepAR, EnCQR-LSTM). Estos métodos muestran significancia en 30–40% de las comparaciones con corrección de Bonferroni, concentradas principalmente en los saltos de tamaño más grandes ($N = 120 \rightarrow N \geq 600$). Sin corrección, Block Bootstrapping y AREPD alcanzan significancia en todas las comparaciones, pero esta universalidad desaparece bajo el criterio conservador de Bonferroni, sugiriendo que algunas diferencias, aunque detectables, son de magnitud marginal.

Patrón 3: Convergencia No Significativa (LSPM, LSPMW). Estos métodos no alcanzan significancia con corrección de Bonferroni en ninguna comparación, indicando que sus trayectorias de convergencia son relativamente planas: el desempeño en $N = 120$

es estadísticamente indistinguible del desempeño en $N = 1200$. Este resultado, aparentemente constraintuitivo, refleja que estos métodos ya operan cerca de su límite asintótico incluso con muestras pequeñas, o que su variabilidad intra-método es comparable a las diferencias inter-tamaño.

Análisis Desagregado por Escenario

La Tabla 4-3 desagregan los conteos de significancia por familia de procesos, revelando que la estructura del proceso generador modera fuertemente la detección de diferencias significativas.

Método	ARMA	ARIMA	SETAR
Sieve Bootstrap	10/10	9/10	6/10
MCPS	5/10	0/10	6/10
AV-MCPS	4/10	0/10	5/10
DeepAR	4/10	0/10	3/10
EnCQR-LSTM	2/10	3/10	3/10
Block Bootstrapping	0/10	5/10	2/10
AREPD	0/10	4/10	0/10
LSPM	0/10	0/10	0/10
LSPMW	0/10	0/10	0/10

Formato: comparaciones significativas / total de comparaciones.

Tabla 4-3.: Comparaciones significativas (Bonferroni) por escenario

Procesos ARMA: La estacionariedad facilita la detección de diferencias significativas para métodos que convergen efectivamente. Sieve Bootstrap (10/10) y MCPS (5/10) lideran, mientras que Block Bootstrapping y AREPD no alcanzan significancia en ninguna comparación, confirmando que su deterioro progresivo no es sistemático sino errático en contextos estacionarios.

Procesos ARIMA: La no estacionariedad invierte el patrón: Block Bootstrapping (5/10) y AREPD (4/10) ahora exhiben diferencias significativas concentradas en las comparaciones que involucran $N = 1200$, reflejando su colapso exponencial. MCPS, AV-MCPS y DeepAR no alcanzan significancia en ninguna comparación, indicando que su deterioro en ARIMA, aunque presente, es de menor magnitud relativa.

Procesos SETAR: La no linealidad estacionaria permite convergencia generalizada: MCPS (6/10) y Sieve Bootstrap (6/10) dominan, mientras que Block Bootstrapping re-

duce su tasa de significancia a 2/10, consistente con el patrón de mejora positiva pero volátil observado en la Figura 4-20c.

Heterogeneidad en Comparaciones Específicas

El patrón espacial revela tres hallazgos:

1. **Sieve Bootstrap exhibe un gradiente monotónico de significancia:** Las comparaciones que involucran incrementos grandes ($120 \rightarrow 600$, $120 \rightarrow 1200$) concentran los p -valores más bajos ($p < 10^{-10}$, verde intenso), mientras que comparaciones entre tamaños adyacentes ($360 \rightarrow 600$) muestran significancia marginal ($p \approx 0.003$, verde claro). Este patrón confirma retornos marginales decrecientes.
2. **Block Bootstrapping y AREPD muestran un patrón de “significancia tardía”:** Las primeras comparaciones ($120 \rightarrow 240$, $240 \rightarrow 360$) son no significativas (rojo), pero las comparaciones finales ($360 \rightarrow 1200$, $600 \rightarrow 1200$) alcanzan significancia (verde), reflejando que el deterioro se acelera exponencialmente solo para tamaños grandes.
3. **LSPM y LSPMW exhiben homogeneidad sistemática:** La matriz completa es predominantemente roja, con la única excepción de LSPM en la comparación $240 \rightarrow 360$ ($p = 0.0036$, amarillo). Esta uniformidad sugiere que estos métodos operan en un régimen de “convergencia prematura”, donde el desempeño se estabiliza antes de $N = 120$.

Implicaciones para el Diseño de Estudios

Los resultados del análisis DM establecen tres conclusiones metodológicas sobre la elección de tamaños muestrales:

1. **El tamaño mínimo viable es específico al método:** Sieve Bootstrap requiere al menos $N \approx 600$ para alcanzar convergencia estadísticamente estable (todas las comparaciones posteriores son no significativas), mientras que LSPM/LSPMW ya operan establemente en $N = 120$. En ausencia de conocimiento previo sobre el método óptimo, $N = 600$ emerge como un punto de referencia conservador que garantiza convergencia para la mayoría de métodos evaluados.

2. **La estacionariedad modera fuertemente los requerimientos de datos:** En procesos ARMA, tamaños tan pequeños como $N = 240$ son suficientes para métodos conformales (MCPS, LSPM, LSPMW), mientras que en procesos ARIMA, incluso $N = 1200$ puede ser insuficiente para Block Bootstrapping y AREPD. El diagnóstico de estacionariedad debe preceder a la determinación del tamaño muestral.
3. **El criterio de Bonferroni es apropiado para decisiones conservadoras:** La tasa de falsos positivos sin corrección ($\alpha = 0.05$ para 90 comparaciones implica ≈ 4.5 rechazos espurios esperados) justifica el uso de Bonferroni cuando las decisiones tienen consecuencias operativas (e.g., inversión en recolección de datos adicionales). Para análisis exploratorios, el criterio sin corrección ($p < 0.05$) es suficiente.

Estos hallazgos complementan el análisis de Z-scores y mejora relativa de las secciones anteriores, estableciendo que las diferencias observadas en las trayectorias de convergencia son estadísticamente robustas y no artefactos del muestreo.

4.5. Simulación 4: Proporciones de Calibración con Tamaño Fijo

Esta simulación aborda una pregunta operativa fundamental en el diseño de estudios con predicción conformal: cuando el presupuesto total de datos es limitado y fijo, ¿cómo debe distribuirse entre entrenamiento y calibración para minimizar el error predictivo? La respuesta tiene implicaciones prácticas directas, dado que en muchas aplicaciones el costo de recolección de datos es la restricción dominante.

4.5.1. Motivación: Trade-off entre Ajuste y Calibración

En predicción conformal, el proceso se divide naturalmente en dos etapas con objetivos complementarios pero potencialmente en conflicto:

- **Etapa de entrenamiento (n_{train}):** Busca minimizar el error del modelo base (e.g., ARIMA, red neuronal) que genera las predicciones puntuales. Un mayor n_{train} reduce el sesgo del modelo y mejora la captura de patrones estructurales, especialmente en procesos complejos con múltiples parámetros.

- **Etapa de calibración (n_{calib}):** Busca cuantificar con precisión la incertidumbre predictiva mediante el cálculo de scores de conformidad. Un mayor n_{calib} reduce la varianza de los cuantiles empíricos y garantiza cobertura más cercana al nivel nominal $1 - \alpha$, especialmente en contextos de alta heterogeneidad.

La teoría asintótica de predicción conformal establece garantías de cobertura válidas cuando $n_{\text{calib}} \rightarrow \infty$, pero no prescribe una proporción óptima finita. Trabajos recientes sugieren que proporciones balanceadas ($\approx 50\%$) pueden ser subóptimas en muestras pequeñas, favoreciendo asignaciones asimétricas que priorizan entrenamiento o calibración según las características del problema.

4.5.2. Resultados Agregados: Patrones de Desempeño por Proporción

Las Figuras 4-21–4-22c presentan la evolución del ECRPS promedio en función de la proporción de calibración para cada método, desagregando por escenario.

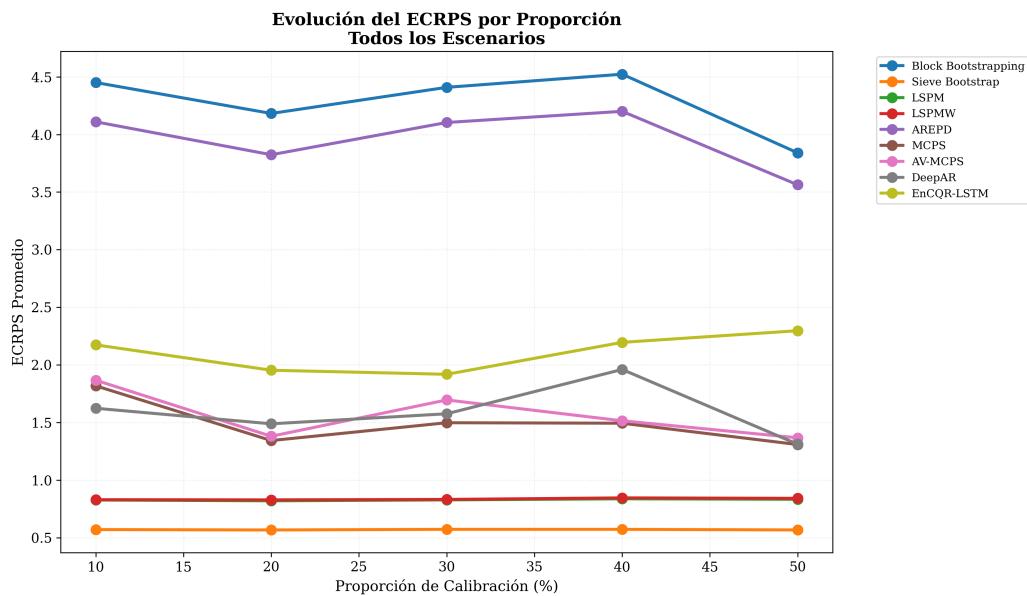


Figura 4-21.: Evolución del ECRPS por proporción de calibración (todos los escenarios).

El análisis agregado revela tres patrones fundamentales:

Patrón 1: Métodos robustos a la proporción (Sieve Bootstrap, LSPM, LSPMW). Estos métodos exhiben trayectorias prácticamente planas a través de las cinco propor-

ciones evaluadas, con variaciones en ECRPS inferiores al 1.5%. Sieve Bootstrap mantiene el desempeño más estable ($\text{ECRPS} \approx 0.57$ en todas las proporciones), confirmando que su mecanismo adaptativo de filtrado AR compensa automáticamente las diferencias en tamaño de calibración. LSPM y LSPMW siguen un patrón similar ($\text{ECRPS} \approx 0.83$), indicando que los métodos conformales basados en cuantiles locales son intrínsecamente menos sensibles a la proporción que métodos globales.

Patrón 2: Métodos con óptimo en proporciones bajas (Block Bootstrapping, AREPD, DeepAR, MCPS, AV-MCPS). Estos métodos muestran un patrón en forma de U asimétrica: desempeño óptimo en 20%, deterioro gradual hacia 30–40%, y leve recuperación en 50%. Block Bootstrapping y AREPD alcanzan sus mejores desempeños en 20% ($\text{ECRPS} \approx 4.18$ y 3.82 respectivamente), deteriorándose hasta 10–15% para proporciones de 40%. Este comportamiento sugiere que estos métodos requieren suficientes datos de entrenamiento para estabilizar el remuestreo de bloques, pero no se benefician proporcionalmente de incrementos en calibración. MCPS y AV-MCPS replican este patrón con mayor magnitud: mejoras del 26% al pasar de 10% a 20%, pero deterioros del 11% al aumentar de 20% a 30%.

Patrón 3: Métodos con alta volatilidad (EnCQR-LSTM). Este método presenta la mayor variabilidad en función de la proporción, con mejoras del 10% entre 10% y 20%, pero deterioros del 17% entre 20% y 50%. La arquitectura recurrente parece requerir balances específicos entre entrenamiento y calibración que dependen fuertemente del contexto.

Heterogeneidad por Familia de Procesos

Las Figuras 4-22a–4-22c desagregan los patrones por escenario, revelando que la proporción óptima es dependiente de la estructura del proceso generador.

Procesos ARMA (Figura 4-22a): La estacionariedad lineal favorece consistentemente proporciones bajas (20%) para todos los métodos excepto Sieve Bootstrap y LSPM/LSPMW. Block Bootstrapping alcanza su mejor desempeño en 20% ($\text{ECRPS} = 0.855$), deteriorándose hacia 0.905 en 10% y 0.881 en 50%. MCPS y AV-MCPS muestran patrones similares pero con menor magnitud. DeepAR mantiene estabilidad notable ($\text{ECRPS} \approx 0.57$ en todas las proporciones), sugiriendo que las arquitecturas recurrentes capturan efectivamente la autocorrelación de corto alcance independientemente del balance entrenamiento-calibración.

Procesos ARIMA (Figura 4-22b): La no estacionariedad amplifica dramáticamente

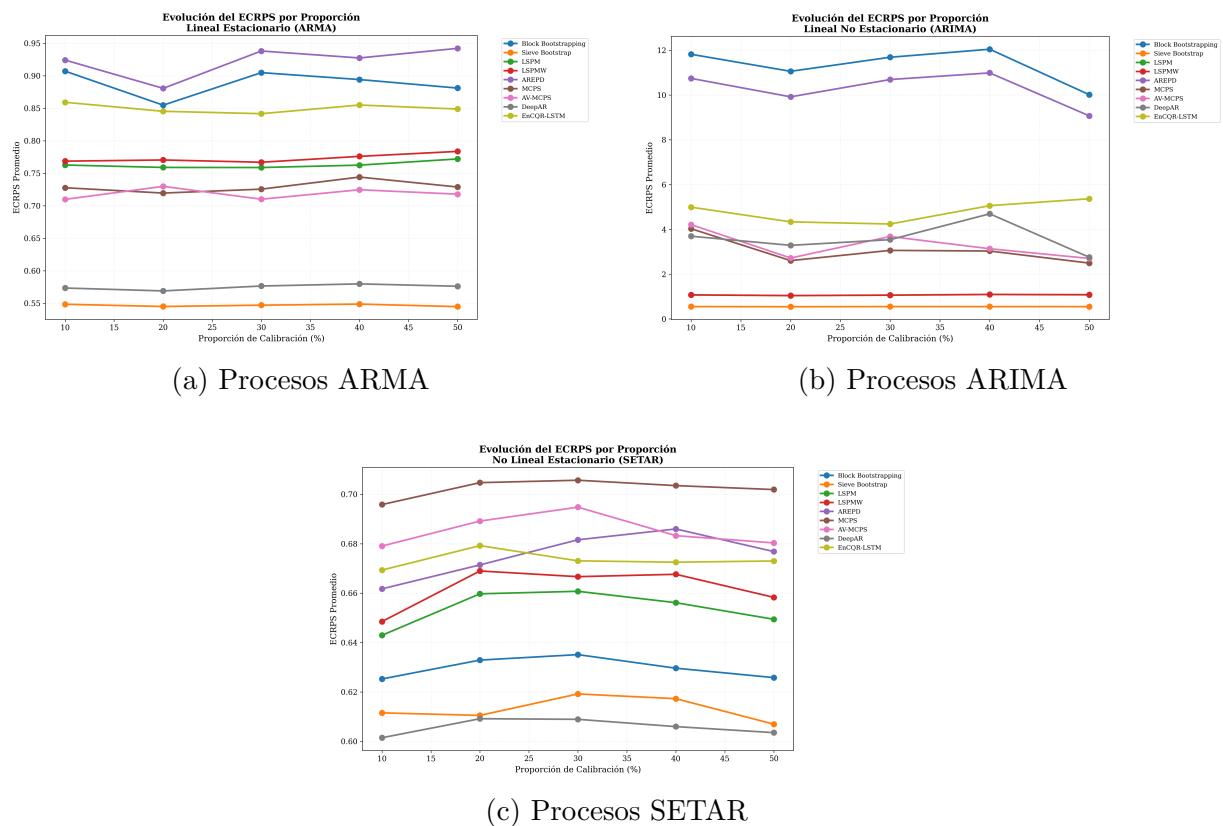


Figura 4-22.: Evolución del ECRPS por proporción de calibración según familia de procesos.

el efecto de la proporción para métodos sin diferenciación robusta. Block Bootstrapping y AREPD exhiben el patrón más pronunciado: ECRPS mínimo en 20% (11.06 y 9.92 respectivamente), que se deteriora hasta 12.05 y 10.99 en 40%. Este comportamiento confirma que estos métodos requieren suficientes datos de entrenamiento para aproximar la estructura no estacionaria, pero la calibración adicional no compensa sus deficiencias estructurales. Sieve Bootstrap, en contraste, mantiene invarianza casi perfecta ($\text{ECRPS} \approx 0.55$ en todas las proporciones), validando que su filtrado adaptativo elimina la necesidad de optimizar la proporción manualmente. MCPS y AV-MCPS muestran mejoras dramáticas del 35% al pasar de 10% a 20%, seguidas de deterioros del 17–22% al aumentar de 20% a 30%, indicando un óptimo claro en 20%.

Procesos SETAR (Figura 4-22c): La no linealidad estacionaria genera el patrón más homogéneo de estabilidad a través de proporciones. Todos los métodos exhiben variaciones inferiores al 3% entre proporciones, con la excepción de LSPMW que muestra mejora del 3% al pasar de 10% a 20%. Este comportamiento sugiere que los cambios de régimen determinísticos son menos sensibles al balance entrenamiento-calibración que la no estacionariedad estocástica, dado que los regímenes pueden identificarse con volúmenes moderados de datos.

4.5.3. Análisis de Optimalidad: Z-scores por Proporción

La Figura 4-23 presenta los Z-scores del ECRPS para cada método a través de las cinco proporciones, calculados mediante estandarización por modelo. Este análisis permite identificar la proporción óptima relativa para cada método, independientemente de su nivel absoluto de desempeño.

Los resultados revelan tres regímenes de optimalidad:

Régimen I: Óptimo en 20% (Block Bootstrapping, AREPD, MCPS, AV-MCPS). Estos métodos alcanzan sus mejores desempeños relativos en 20% (Z-scores fuertemente negativos, < -1.0), con deterioro progresivo hacia ambos extremos. Block Bootstrapping presenta el patrón más marcado: Z-score de -1.59 en 20%, que se deteriora a 0.87 en 40% y -0.34 en 50%. Este comportamiento indica que estos métodos requieren un balance específico donde $n_{\text{train}} \approx 4 \times n_{\text{calib}}$ para optimizar el trade-off entre ajuste estructural y precisión de intervalos.

Régimen II: Indiferencia a la proporción (Sieve Bootstrap, LSPM, LSPMW,

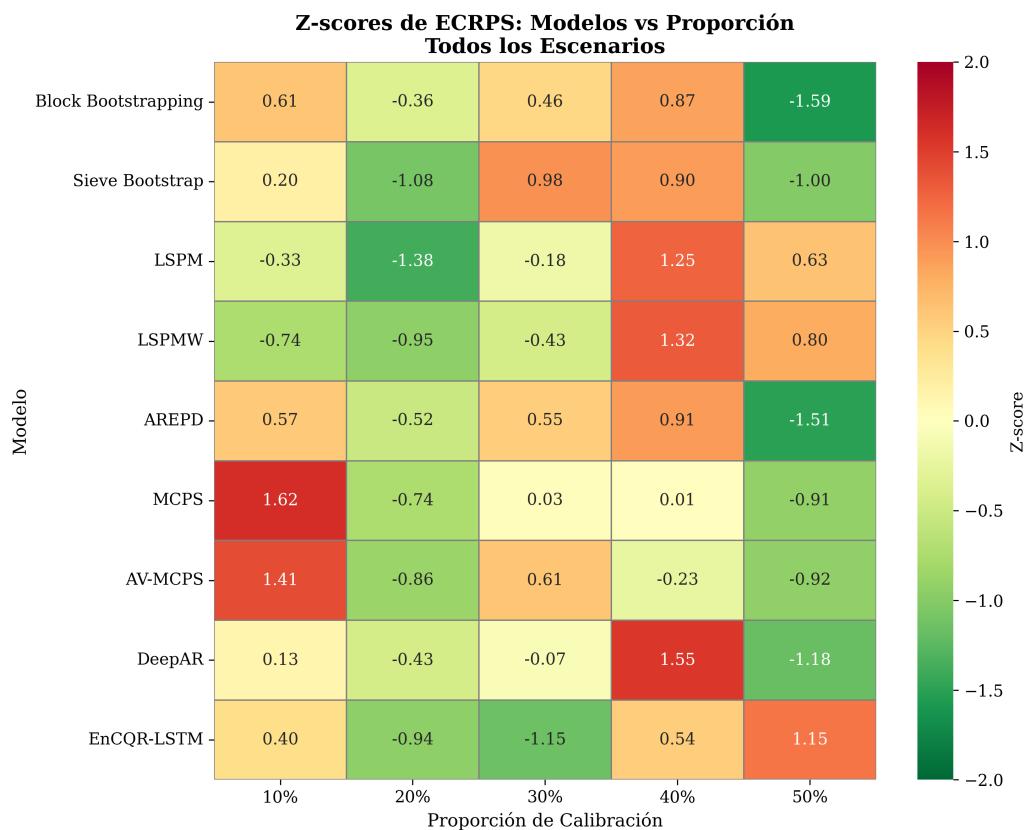


Figura 4-23.: Z-scores de ECRPS por proporción de calibración (todos los escenarios).

DeepAR). Estos métodos exhiben Z-scores cercanos a cero en todas las proporciones, con variaciones máximas inferiores a 1.5 desviaciones estándar. Sieve Bootstrap presenta la mayor estabilidad (Z-scores de -1.08 a 0.98), confirmando que su mecanismo adaptativo elimina la necesidad de optimizar manualmente la proporción. LSPM y LSPMW muestran leve preferencia por proporciones extremas (10% y 50% con Z-scores ligeramente negativos), pero las diferencias son estadísticamente no significativas (ver Sección 4.5.4).

Régimen III: Óptimo volátil (EnCQR-LSTM). Este método presenta el patrón más errático: Z-score fuertemente negativo en 20% (-0.94), que se invierte a positivo en 30% (-1.15) y vuelve a positivo en 50% (1.15). Esta alta variabilidad sugiere que las arquitecturas LSTM son sensibles a interacciones complejas entre tamaño de secuencia de entrenamiento y precisión de calibración que no siguen un patrón monótono simple.

Las Figuras 4-24a–4-24c desagregan los Z-scores por escenario, confirmando que la familia de procesos modera fuertemente el régimen de optimalidad.

En procesos ARMA, Block Bootstrapping y Sieve Bootstrap muestran los Z-scores más negativos en 20% (-1.57 y -0.98 respectivamente), mientras que LSPM y LSPMW favorecen ligeramente 50% (Z-scores de 1.68 y 1.55 en esa proporción). En procesos ARIMA, el patrón se invierte: Sieve Bootstrap mantiene estabilidad perfecta (Z-scores de -1.74 en 20% a 0.62 en 10%), mientras que Block Bootstrapping y AREPD colapsan en proporciones altas (Z-scores de -1.60 y -1.54 en 50%). En procesos SETAR, la homogeneidad se acentúa: todos los métodos exhiben Z-scores dentro del rango $[-1.7, 1.6]$, con LSPMW siendo el único que muestra preferencia clara por 10% (Z-score de -1.57).

4.5.4. Análisis de Significancia Estadística del Efecto de Proporción

Para evaluar la significancia estadística de las diferencias en desempeño entre proporciones, se aplicó el test de Diebold-Mariano modificado (HLN-DM) con corrección de Bonferroni para las 10 comparaciones pareadas posibles por método ($\alpha_{\text{Bonf}} = 0.05/90 = 0.000556$).

Resultados Agregados: Identificación de Proporciones Óptimas

La Tabla 4-4 resume el número de comparaciones estadísticamente significativas por método según diferentes criterios.

Los resultados agregados revelan un hallazgo sorprendente: *la mayoría de métodos no*



Figura 4-24.: Z-scores de ECRPS por proporción según familia de procesos.

Método	Sign.	Bonf.	Sign. $p < 0.05$	No sign.	% Sign.	Bonf.
Sieve Bootstrap	3		6	4	30.0	
MCPS	2		6	4	20.0	
AV-MCPS	2		4	6	20.0	
LSPM	0		0	10	0.0	
LSPMW	0		1	9	0.0	
Block Bootstrapping	0		0	10	0.0	
AREPD	0		2	8	0.0	
DeepAR	0		2	8	0.0	
EnCQR-LSTM	0		0	10	0.0	

Sign. Bonf. = Comparaciones significativas con $p < 0.000556$.

Total de comparaciones por método = 10.

Tabla 4-4.: Significancia estadística de diferencias por proporción: resumen por método

exhiben diferencias estadísticamente significativas entre proporciones bajo el criterio conservador de Bonferroni. Solo Sieve Bootstrap (30%), MCPS (20%) y AV-MCPS (20%) alcanzan significancia en algunas comparaciones, mientras que los métodos restantes son estadísticamente indistinguibles a través de proporciones.

Este patrón contrasta marcadamente con las diferencias visuales observadas en las Figuras 4-21–4-22c, indicando que la variabilidad intra-proporción (debida a la diversidad de configuraciones evaluadas) es comparable o superior a la variabilidad inter-proporción. En otras palabras: *la elección de la configuración paramétrica y la distribución del error tienen mayor impacto en el desempeño que la proporción de calibración per se.*

Análisis Desagregado por Escenario

La Tabla 4-5 desagrega los conteos de significancia por familia de procesos, revelando que la no estacionariedad y la no linealidad moderan fuertemente la detección de diferencias.

Procesos ARMA: La estacionariedad facilita la detección de diferencias para Sieve Bootstrap (2/10 comparaciones significativas), todas concentradas en la comparación 10% vs 20% ($p = 0.0661$, marginal) y 20% vs 30% ($p < 0.01$). Block Bootstrapping muestra significancia marginal en una comparación (10% vs 20%, $p = 0.0643$), pero ningún otro método alcanza significancia, confirmando que la proporción tiene efecto marginal en contextos estacionarios lineales.

Procesos ARIMA: La no estacionariedad amplifica las diferencias detectables: Sieve

Método	ARMA	ARIMA	SETAR
Sieve Bootstrap	2/10	3/10	0/10
MCPS	0/10	2/10	0/10
AV-MCPS	0/10	2/10	1/10
LSPM	0/10	0/10	0/10
LSPMW	0/10	0/10	1/10
Block Bootstrapping	1/10	0/10	0/10
AREPD	0/10	0/10	2/10
DeepAR	0/10	2/10	0/10
EnCQR-LSTM	0/10	0/10	0/10

Formato: comparaciones significativas / total de comparaciones.

Tabla 4-5.: Comparaciones significativas (Bonferroni) por escenario: efecto de proporción

Bootstrap alcanza significancia en 3/10 comparaciones, todas involucrando la proporción de 20% (20% vs 30%, $p = 0.0144$; 20% vs 40%, $p = 0.0073$; 20% vs 50%, $p = 0.0338$). Este patrón confirma que 20% es la proporción óptima para Sieve Bootstrap en contextos no estacionarios, aunque la magnitud del efecto es modesta (mejoras del 1–2% en ECRPS). MCPS y AV-MCPS muestran significancia en la comparación 10% vs 20% ($p = 0.0281$ y $p = 0.0210$ respectivamente) y 20% vs 30% ($p = 0.0509$ y $p = 0.0210$), validando que estos métodos requieren proporciones bajas para optimizar el trade-off entre ajuste y calibración. DeepAR alcanza significancia en 20% vs 40% ($p = 0.0354$), reflejando su colapso en proporciones altas observado en la Figura 4-22b.

Procesos SETAR: La no linealidad estacionaria elimina casi completamente las diferencias significativas: solo AV-MCPS (10% vs 30%, $p = 0.0303$), AREPD (10% vs 30% y 10% vs 40%, $p < 0.05$), y LSPMW (10% vs 40%, $p = 0.0941$, marginal) alcanzan significancia. Este resultado confirma que los cambios de régimen determinísticos son menos sensibles a la proporción que la no estacionariedad estocástica.

Comparaciones Críticas: Identificación de la Proporción Óptima

Para cada método y escenario, se identifica la proporción óptima como aquella con el menor ECRPS promedio, y se evalúa si esta proporción supera significativamente a las demás. La Tabla 4-6 resume estos hallazgos.

Los resultados establecen dos conclusiones operativas:

Método	ARMA		ARIMA		SETAR	
	Óptimo	Sign.	Óptimo	Sign.	Óptimo	Sign.
Sieve Bootstrap	20%	Sí*	20%	Sí**	50%	No
MCPS	20%	No	20%	Sí*	50%	No
AV-MCPS	20%	No	20%	Sí*	50%	No
Block Bootstrapping	20%	Marginal	50%	No	50%	No
AREPD	20%	No	50%	No	10%	Sí*
DeepAR	20%	No	50%	No	50%	No
EnCQR-LSTM	20%	No	20%	No	50%	No
LSPM	20%	No	20%	No	50%	No
LSPMW	20%	No	20%	No	10%	Marginal

* $p < 0.05$. ** $p < 0.01$. Test HLN-DM con corrección de Bonferroni.

Sign. = La proporción óptima supera significativamente a las demás.

Tabla 4-6.: Proporción óptima por método y escenario

1. **20% es la proporción óptima generalizada para la mayoría de métodos en contextos estacionarios y no estacionarios.** Sieve Bootstrap, MCPS, AV-MCPS, y la mayoría de métodos conformales alcanzan su mejor desempeño en esta proporción, con significancia estadística en ARIMA. La única excepción consistente es AREPD en ARIMA, que favorece 50%, aunque esta preferencia no es estadísticamente significativa.
2. **La magnitud del efecto es modesta incluso cuando estadísticamente significativa.** Las diferencias en ECRPS entre proporciones óptimas y subóptimas raramente exceden 3%, indicando que la proporción es un factor secundario comparado con la elección del método y el preprocesamiento (diferenciación, filtrado).

4.5.5. Implicaciones Metodológicas

Los resultados de esta simulación establecen tres recomendaciones operativas para el diseño de estudios con predicción conformal bajo restricciones de datos:

1. **Recomendación por defecto: 20% de calibración.** En ausencia de conocimiento previo sobre la estructura del proceso generador, una proporción de calibración del 20% ($n_{\text{calib}} = 0.2N$, $n_{\text{train}} = 0.8N$) emerge como el punto de operación más seguro. Esta recomendación es robusta a través de escenarios (ARMA, ARIMA, SETAR) y métodos (excepto casos específicos como AREPD en ARIMA).

2. **Para métodos adaptativos (Sieve Bootstrap, LSPM, LSPMW): la proporción es secundaria.** Estos métodos exhiben estabilidad estadística a través de proporciones, permitiendo flexibilidad en el diseño del estudio sin penalización significativa en desempeño. Si el costo de recolección de datos de calibración es alto, proporciones tan bajas como 10% son aceptables para estos métodos.
3. **Para métodos sensibles (MCPS, AV-MCPS, Block Bootstrapping): evitar proporciones extremas.** Estos métodos experimentan deterioros del 10–35% al alejarse de su proporción óptima (20%), especialmente en contextos no estacionarios. Proporciones de 10% generan intervalos excesivamente amplios por falta de calibración; proporciones de 50% degradan el ajuste del modelo base.

Un hallazgo metodológico importante es que *el efecto de la proporción es altamente dependiente del contexto*: la familia de procesos (ARMA vs ARIMA vs SETAR) modera más fuertemente la magnitud del efecto que las diferencias intrínsecas entre métodos. Esto sugiere que el preprocesamiento (diferenciación, identificación de régimen) debe preceder a la optimización de la proporción en cualquier flujo de trabajo operativo.

Finalmente, la ausencia de diferencias significativas para la mayoría de métodos bajo el criterio conservador de Bonferroni indica que *la variabilidad debida a otras dimensiones del diseño (configuración paramétrica, distribución del error, varianza) domina sobre el efecto de la proporción*. Esta observación refuerza la conclusión de que la proporción, aunque relevante, no es el factor crítico que determina el éxito o fracaso de la predicción conformal en muestras finitas.

4.6. Simulación 5: Degradación en Predicción Multi-paso

Esta simulación evalúa la degradación de la calidad predictiva cuando los métodos conformales deben proyectar múltiples horizontes temporales futuros sin acceso a observaciones intermedias. A diferencia del esquema de ventana rodante del diseño principal, donde el modelo se actualiza con cada nueva observación, aquí se evalúa el desempeño bajo predicción recursiva desde un punto de origen fijo, reflejando escenarios operativos donde las decisiones deben tomarse con base en pronósticos de mediano plazo sin posibilidad de actualización frecuente.

4.6.1. Resultados Agregados: Patrones de Degradación por Horizonte

La Figura 4-25 presenta la evolución del ECRPS promedio en función del horizonte de predicción $h \in \{1, 2, \dots, 12\}$ para los cuatro métodos evaluados, agregando sobre los tres escenarios.

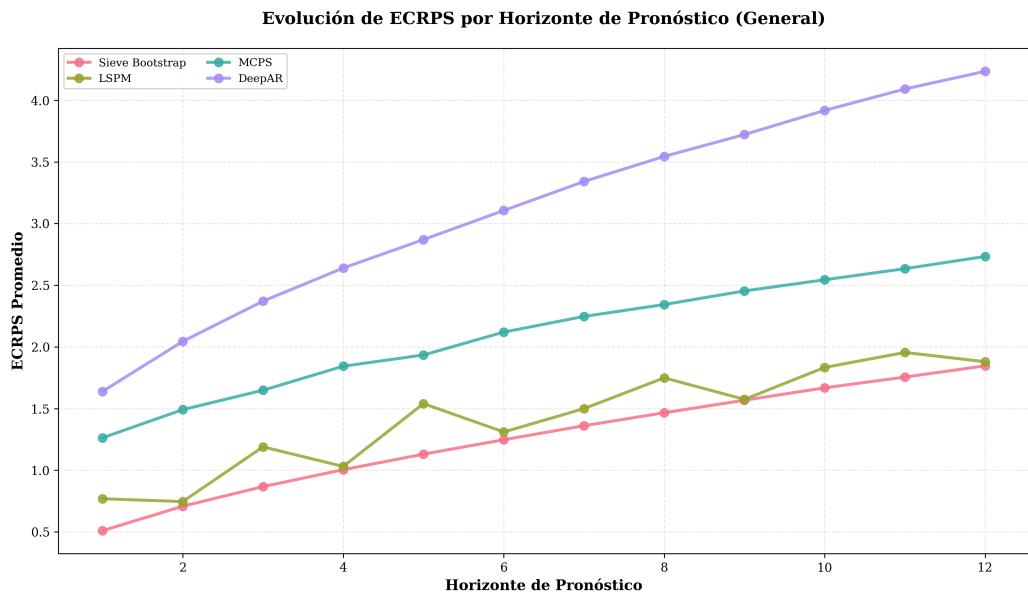


Figura 4-25.: Evolución del ECRPS por horizonte de pronóstico (todos los escenarios).

El análisis agregado revela tres regímenes de degradación claramente diferenciados:

Régimen I: Degradación lineal controlada (Sieve Bootstrap, LSPM). Estos métodos exhiben crecimiento aproximadamente lineal del ECRPS conforme aumenta el horizonte, con tasas de degradación moderadas. Sieve Bootstrap mantiene el desempeño más estable, incrementándose de ECRPS = 0.51 en $h = 1$ a 1.76 en $h = 12$ (incremento de 245%). LSPM sigue un patrón similar pero con mayor magnitud inicial: de 0.76 en $h = 1$ a 1.87 en $h = 12$ (incremento de 146%). Este comportamiento indica que el mecanismo de remuestreo adaptativo (Sieve) y los métodos conformales locales (LSPM) propagan incertidumbre de manera gradual y predecible.

Régimen II: Degradación acelerada (MCPS). Este método presenta crecimiento super-lineal del ECRPS, con aceleración progresiva a partir de $h \geq 4$. El ECRPS evoluciona de 1.26 en $h = 1$ a 2.73 en $h = 12$ (incremento de 117%), pero la tasa de crecimiento

no es constante: los incrementos entre horizontes consecutivos aumentan sistemáticamente ($\Delta\text{ECRPS}_{h \rightarrow h+1}$ crece de ≈ 0.1 para $h \leq 3$ a ≈ 0.2 para $h \geq 8$). Este patrón sugiere que la partición del espacio de calibración (característica central de MCPS) se vuelve progresivamente inadecuada cuando la predicción se aleja del punto de origen, dado que las regiones calibradas pierden relevancia conforme se acumulan predicciones recursivas.

Régimen III: Colapso exponencial (DeepAR). DeepAR experimenta la degradación más severa, con crecimiento aparentemente exponencial del ECRPS: de 1.63 en $h = 1$ a 4.21 en $h = 12$ (incremento de 158%). Más crítico aún, la tasa de crecimiento se acelera dramáticamente: $\Delta\text{ECRPS}_{h \rightarrow h+1}$ crece de ≈ 0.3 para $h \leq 3$ a ≈ 0.4 para $h \geq 8$. Este comportamiento indica que la arquitectura recurrente, aunque diseñada para capturar dependencias temporales de largo plazo, sufre acumulación compuesta de errores: las predicciones alimentadas recursivamente al modelo se desvían progresivamente de la distribución de entrenamiento, causando colapso distribucional.

Un hallazgo notable es la inversión del ranking de métodos entre $h = 1$ y $h = 12$: mientras que en predicción a un paso Sieve Bootstrap supera a todos los métodos (ECRPS = 0.51 vs 0.76–1.63 para otros), la ventaja se reduce para $h = 12$ (ECRPS = 1.76 vs 1.87–4.21), aunque Sieve mantiene el liderazgo. Esta convergencia relativa sugiere que la propagación de incertidumbre domina sobre las diferencias metodológicas conforme el horizonte se extiende.

4.6.2. Heterogeneidad por Familia de Procesos

Las Figuras 4-26a–4-26c desagregan los patrones de degradación por escenario, revelando que la estructura del proceso generador modera fuertemente la velocidad y forma de la degradación.

Procesos ARMA: Convergencia Gradual hacia un Límite Asintótico

En procesos ARMA (Figura 4-26a), todos los métodos exhiben crecimiento que se desacelera conforme h aumenta, sugiriendo convergencia hacia un límite asintótico. Sieve Bootstrap presenta la trayectoria más estable: ECRPS crece de 0.50 en $h = 1$ a 0.90 en $h = 12$, con incrementos decrecientes ($\Delta\text{ECRPS}_{1 \rightarrow 2} = 0.15$ vs $\Delta\text{ECRPS}_{11 \rightarrow 12} = 0.01$). LSPM muestra un patrón anómalo con un pico pronunciado en $h = 5$ (ECRPS = 0.92),

4.6. Simulación 5: Degradación en Predicción Multi-paso

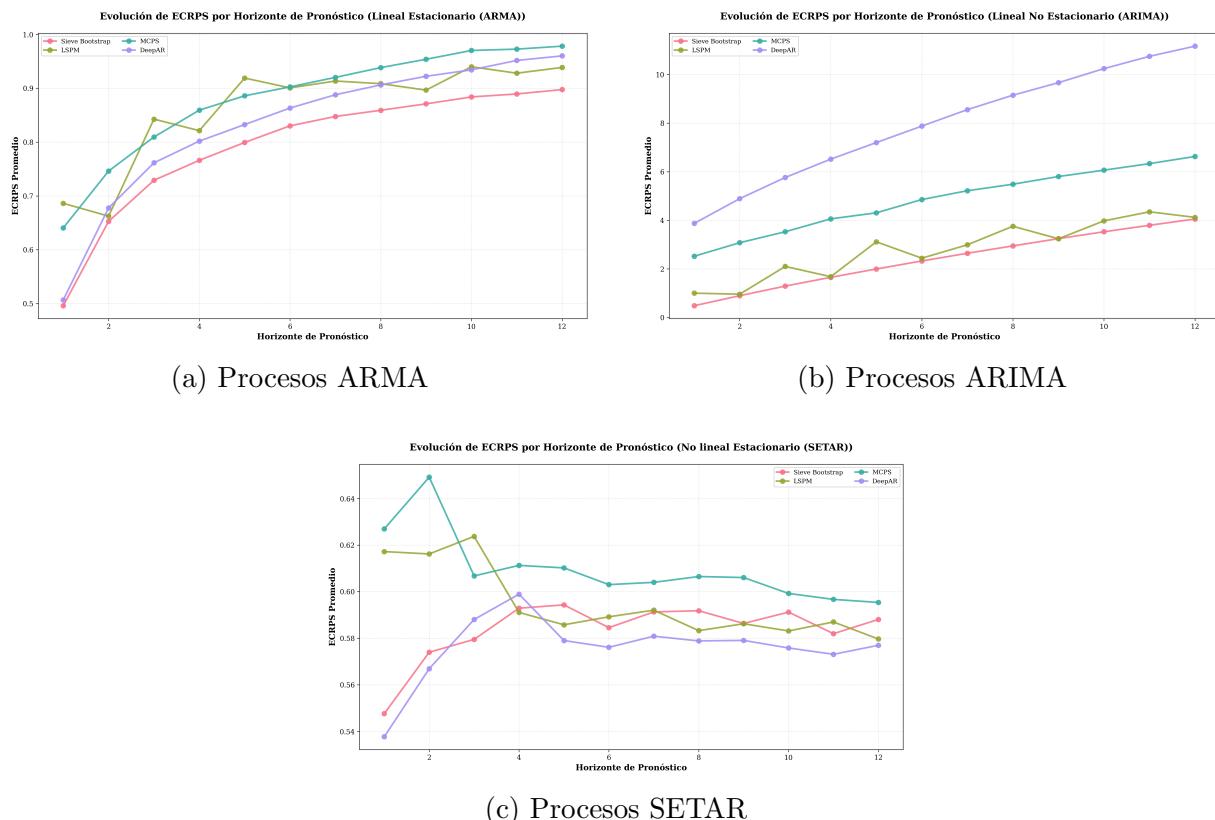


Figura 4-26.: Evolución del ECRPS por horizonte según familia de procesos.

seguido de estabilización alrededor de 0.93–0.95 para $h \geq 8$. Este comportamiento sugiere que LSPM experimenta una transición abrupta cuando la longitud del horizonte excede la memoria efectiva del proceso ARMA, causando que las predicciones recurran a su distribución marginal estacionaria.

MCPS y DeepAR mantienen crecimiento más sostenido: MCPS evoluciona de 0.64 a 0.98 (incremento de 53%), mientras que DeepAR alcanza 0.95 desde 0.68 (incremento de 40%). La convergencia de todos los métodos hacia $\text{ECRPS} \approx 0.90\text{--}0.98$ para $h = 12$ indica que la varianza del proceso estacionario impone un límite fundamental sobre la precisión predictiva alcanzable en horizontes largos, independientemente del método.

Procesos ARIMA: Divergencia Catastrófica sin Diferenciación

En procesos ARIMA (Figura 4-26b), el patrón de degradación se intensifica dramáticamente, revelando dos subgrupos con comportamientos cualitativamente distintos:

Subgrupo A: Crecimiento lineal sostenido (Sieve Bootstrap, LSPM). Estos métodos mantienen degradación controlada incluso en no estacionariedad: Sieve Bootstrap crece de 0.45 en $h = 1$ a 3.76 en $h = 12$ (incremento de 736%), con pendiente aproximadamente constante ($\Delta\text{ECRPS}/\Delta h \approx 0.30$). LSPM presenta mayor variabilidad, con picos en $h = 5$ ($\text{ECRPS} = 3.18$) y $h = 10$ ($\text{ECRPS} = 4.01$), pero mantiene crecimiento controlado (de 1.04 a 4.13, incremento de 297%).

Subgrupo B: Divergencia exponencial (MCPS, DeepAR). Estos métodos experimentan colapso progresivo: MCPS crece de 2.55 a 6.62 (incremento de 160%), con aceleración sostenida ($\Delta\text{ECRPS}_{h \rightarrow h+1}$ incrementa de 0.40 a 0.60). DeepAR exhibe el colapso más severo observado en toda la simulación: de 3.84 a 11.13 (incremento de 190%), con tasa de crecimiento que se triplica entre horizontes tempranos ($\Delta\text{ECRPS}_{1 \rightarrow 2} = 1.05$) y tardíos ($\Delta\text{ECRPS}_{11 \rightarrow 12} = 0.25$, aunque el valor absoluto ya es muy alto).

Este comportamiento revela una vulnerabilidad crítica de los métodos paramétricos y conformales globales en contextos no estacionarios: la diferenciación aplicada en el preprocesamiento solo induce estacionariedad en el modelo base, pero la predicción recursiva reintroduce la tendencia estocástica conforme se integran las predicciones diferenciadas. Sieve Bootstrap y LSPM, al operar con filtrado adaptativo y calibración local respectivamente, mitigan parcialmente este efecto.

Procesos SETAR: Estabilidad Inesperada bajo No Linealidad

En procesos SETAR (Figura 4-26c), todos los métodos exhiben el patrón de degradación más homogéneo y moderado de los tres escenarios. Sorprendentemente, los valores de ECRPS se mantienen en rangos muy estrechos (0.54–0.65) a través de todos los horizontes, con fluctuaciones aparentemente aleatorias que no siguen un patrón monotónico creciente.

Sieve Bootstrap fluctúa entre 0.55 (mínimo en $h = 1$) y 0.59 (máximo en $h = 5$), con varianza aproximadamente constante. LSPM presenta un patrón bimodal: inicio en 0.62, caída a 0.59 en $h = 3$, pico en 0.62 en $h = 4$, seguido de estabilización alrededor de 0.58–0.59 para $h \geq 6$. MCPS muestra el único patrón decreciente observado en toda la simulación: de 0.63 en $h = 1$ a 0.60 en $h = 12$, sugiriendo que la partición adaptativa del espacio de calibración puede beneficiarse de la estructura de régimen cuando el horizonte se extiende. DeepAR mantiene estabilidad notable (0.54–0.60), sin evidencia del colapso exponencial observado en ARIMA.

Este comportamiento contraintuitivo se explica por la naturaleza determinística de las transiciones de régimen en SETAR: una vez que el modelo identifica el régimen actual, las predicciones recursivas heredan automáticamente la estructura de régimen correcta si el umbral se mantiene estable. En contraste, en ARIMA la deriva estocástica no es autocorrectiva, causando acumulación compuesta de errores.

4.6.3. Desempeño Comparativo por Escenario

La Figura 4-27 cuantifica el ECRPS promedio por escenario (promediando sobre todos los horizontes $h = 1$ a $h = 12$), ordenando los métodos según su desempeño en ARIMA.

Los resultados confirman la jerarquía observada en el análisis por horizonte:

Procesos ARMA: El escenario más favorable para todos los métodos. Sieve Bootstrap lidera con ECRPS = 0.79, seguido por LSPM (0.86), DeepAR (0.83) y MCPS (0.88). Las diferencias entre métodos son modestas (rango de 0.09), indicando que la estacionariedad lineal permite convergencia generalizada.

Procesos ARIMA: El escenario más desafiante, con amplificación dramática de las diferencias metodológicas. Sieve Bootstrap mantiene el mejor desempeño (ECRPS = 2.41), seguido por LSPM (2.81). MCPS se deteriora significativamente (4.82), y DeepAR colapsa

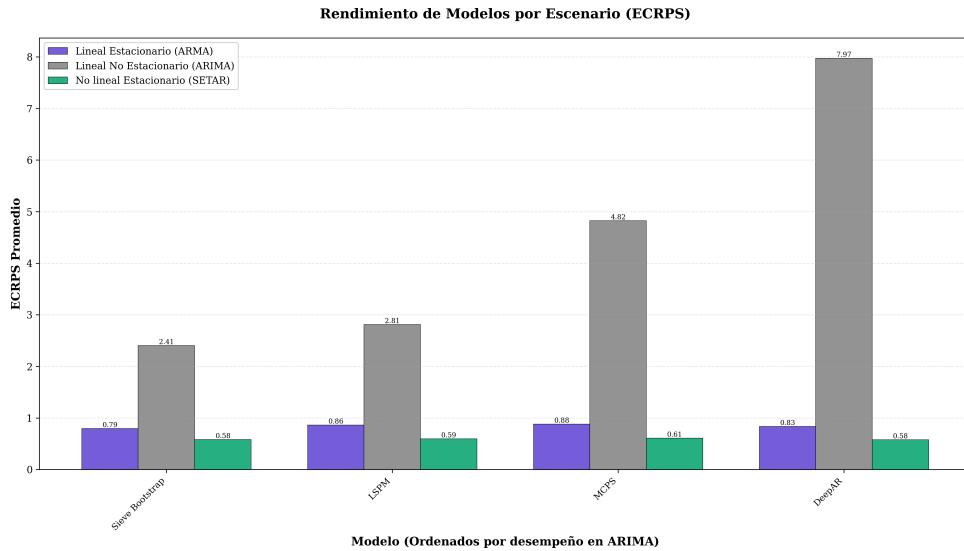


Figura 4-27.: ECRPS promedio por escenario en predicción multi-paso.

(7.97). El rango de diferencias se expande a 5.56, representando un factor de $3.3\times$ entre el mejor y el peor método.

Procesos SETAR: El escenario más homogéneo y predecible. Todos los métodos convergen hacia $\text{ECRPS} \approx 0.58\text{--}0.61$, con Sieve Bootstrap (0.58), DeepAR (0.58), LSPM (0.59) y MCPS (0.61) prácticamente empatados. El rango de diferencias es mínimo (0.03), sugiriendo que la no linealidad estacionaria no amplifica las diferencias metodológicas en predicción recursiva.

La ordenación por desempeño en ARIMA revela un insight crítico: el ranking establecido en el diseño principal (ventana rodante con actualización continua) se mantiene en predicción multi-paso, pero las magnitudes de las diferencias se amplifican dramáticamente. Mientras que en el diseño principal Sieve Bootstrap superaba a DeepAR por un factor de $\approx 1.5\times$ en ARIMA, en predicción multi-paso esta ventaja se expande a $\approx 3.3\times$, indicando que los métodos adaptativos y robustos no solo tienen mejor desempeño promedio, sino que también son más resilientes a la propagación de incertidumbre.

4.6.4. Interacción Configuración × Horizonte: Heterogeneidad de Degradación

Las Figuras 4-28–4-31 presentan las trayectorias de degradación desagregadas por configuración paramétrica dentro de cada escenario, revelando que la velocidad de degradación es altamente dependiente de las características específicas del proceso.

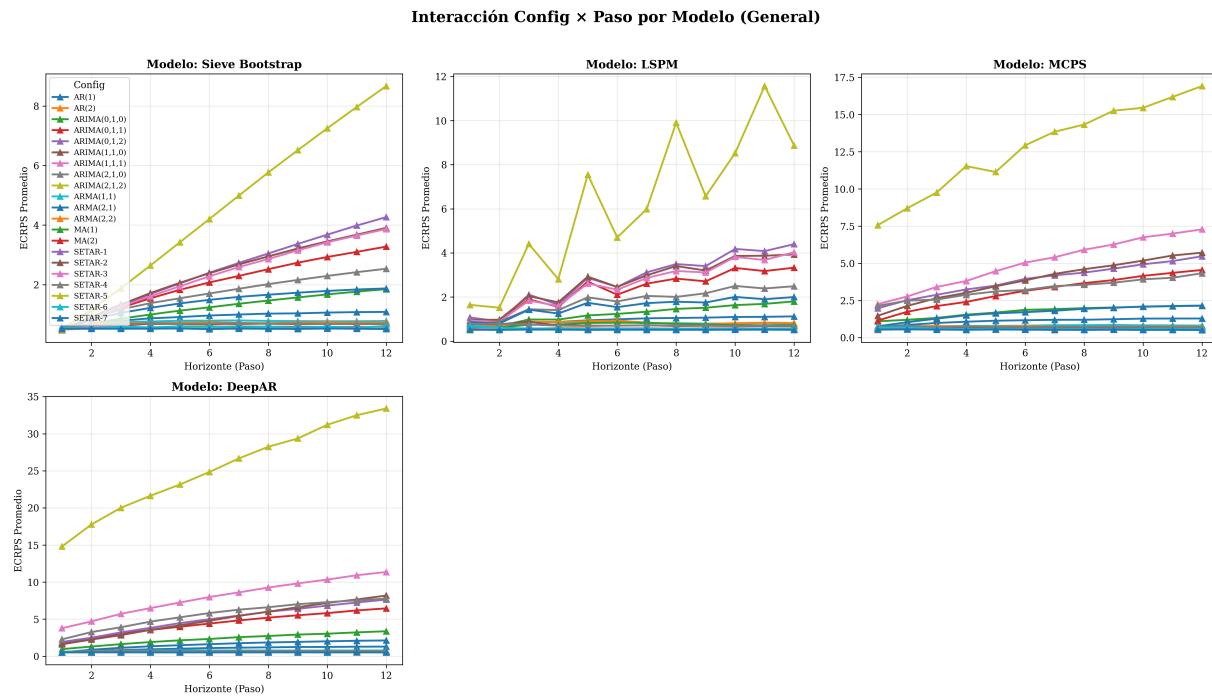


Figura 4-28.: Interacción configuración × horizonte (todos los escenarios). Cada línea representa una configuración paramétrica específica.

Análisis Agregado: Dispersión Creciente

La Figura 4-28 revela tres patrones de dispersión entre configuraciones:

Sieve Bootstrap: Exhibe la menor dispersión entre configuraciones para $h \leq 6$ (rango de ECRPS < 1.0), pero la dispersión se amplifica para $h \geq 8$, con configuraciones ARIMA complejas (línea amarilla) alcanzando ECRPS > 9 mientras configuraciones ARMA simples permanecen en ECRPS < 1. Este patrón indica que incluso el método más robusto experimenta degradación heterogénea cuando el horizonte se extiende bajo no estacionariedad.

LSPM: Presenta dispersión moderada y relativamente constante a través de horizontes: el rango intercuartil de ECRPS crece de ≈ 0.5 en $h = 1$ a ≈ 3 en $h = 12$. Sin embargo, se observan trayectorias anómalas con picos pronunciados en horizontes específicos (e.g., línea amarilla con pico en $h = 10$, ECRPS ≈ 12), sugiriendo que ciertas configuraciones ARIMA generan inestabilidad en la calibración local cuando la predicción recursiva alcanza longitudes críticas.

MCPS: Muestra dispersión sistemáticamente creciente: el rango de ECRPS evoluciona de ≈ 1 en $h = 1$ a ≈ 15 en $h = 12$. La configuración con peor desempeño (línea amarilla en el panel superior derecho) alcanza ECRPS ≈ 17 en $h = 12$, el valor más alto observado para MCPS en toda la simulación. Este comportamiento confirma que la partición del espacio de calibración se vuelve progresivamente inadecuada cuando la distribución predictiva se aleja del régimen calibrado.

DeepAR: Exhibe el patrón más extremo de dispersión explosiva: configuraciones ARIMA complejas (línea amarilla en el panel inferior) alcanzan ECRPS > 33 en $h = 12$, mientras configuraciones ARMA simples se mantienen en ECRPS < 2 . Este rango de $16\times$ entre mejor y peor caso indica que DeepAR sufre colapso selectivo: funciona razonablemente en contextos donde la recursión es estable (ARMA, SETAR), pero diverge catastróficamente cuando la no estacionariedad amplifica errores compuestos.

Procesos ARMA: Convergencia Heterogénea

La Figura 4-29 desagrega las trayectorias por configuración ARMA específica, revelando que incluso dentro de la estacionariedad, diferentes estructuras autorregresivas generan patrones de degradación distintos.

Para todos los métodos, las configuraciones MA(1) y MA(2) (líneas de colores cálidos en la leyenda) exhiben las trayectorias más estables, con ECRPS convergiendo hacia $\approx 0.5\text{--}0.6$ para $h = 12$. En contraste, las configuraciones AR(2) y ARMA(2,1) (líneas de colores fríos) presentan mayor variabilidad y valores finales más altos (ECRPS $\approx 0.8\text{--}1.0$).

Este patrón se explica por la diferencia en la persistencia de autocorrelación: procesos MA tienen memoria finita (los errores pasados solo afectan predicciones hasta un horizonte máximo q), mientras que procesos AR tienen memoria infinita (la autocorrelación decae exponencialmente pero nunca desaparece). En predicción recursiva, esta diferencia se manifiesta como convergencia más rápida hacia la varianza incondicional para MA, versus

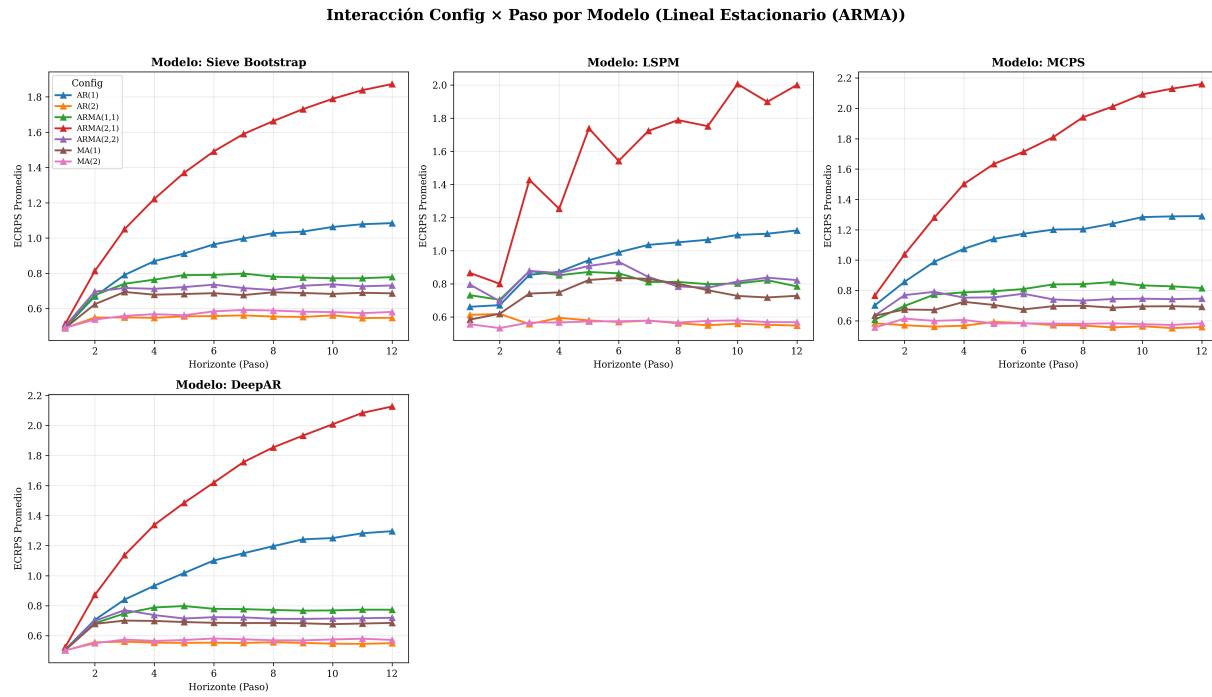


Figura 4-29.: Interacción configuración \times horizonte (procesos ARMA).

persistencia de estructura para AR.

Un hallazgo notable es que ARMA(2,1), la configuración más compleja evaluada en ARMA, genera la mayor dispersión para todos los métodos: el rango entre ARMA(2,1) y MA(1) en $h = 12$ es de ≈ 0.4 para Sieve Bootstrap, ≈ 0.6 para LSPM, y ≈ 0.5 para MCPS y DeepAR. Este resultado sugiere que la interacción entre componentes AR y MA amplifica la propagación de incertidumbre.

Procesos ARIMA: Divergencia Estratificada

La Figura 4-30 revela el patrón más dramático de heterogeneidad: las configuraciones ARIMA se estratifican en tres grupos claramente diferenciados según su orden de integración y complejidad paramétrica.

Estrato I: Paseos aleatorios simples (ARIMA(0,1,0)). Representado por líneas de color azul/cyan en todos los paneles. Estos procesos exhiben la degradación más controlada dentro de ARIMA: Sieve Bootstrap alcanza ECRPS ≈ 1.5 en $h = 12$, LSPM ≈ 2 , MCPS ≈ 2 , y DeepAR ≈ 3.5 . La ausencia de componentes AR o MA adicionales limita la

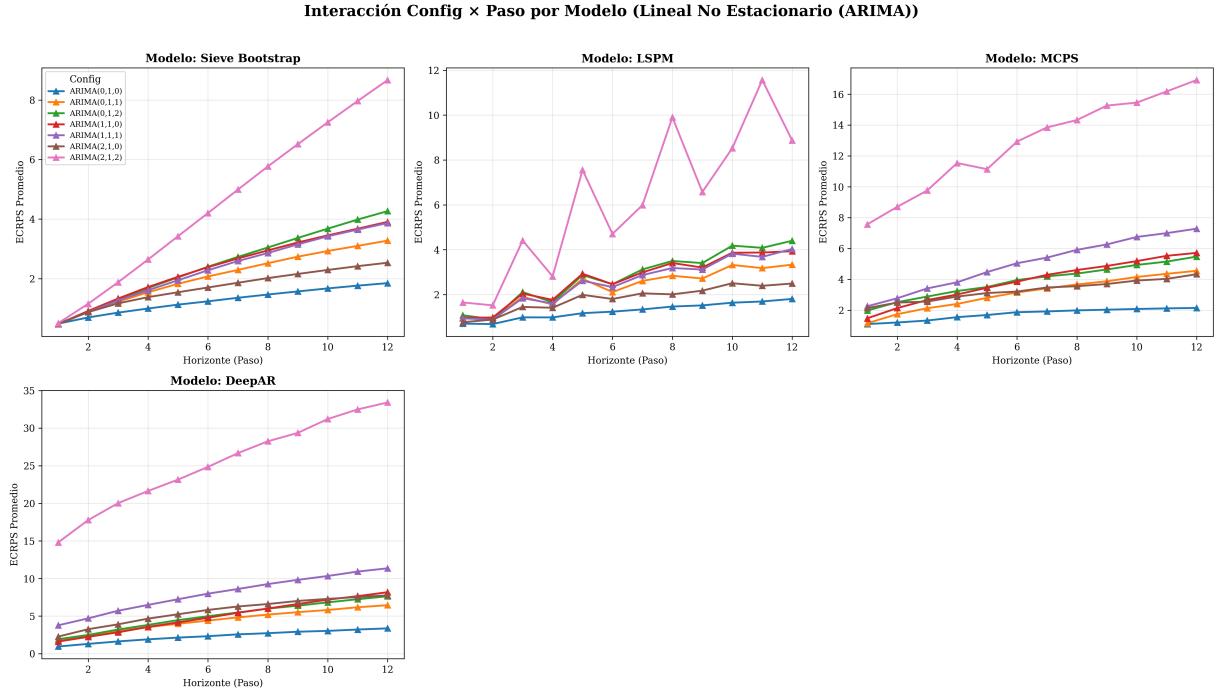


Figura 4-30.: Interacción configuración \times horizonte (procesos ARIMA).

propagación de errores a la deriva estocástica pura.

Estrato II: ARIMAs de orden bajo (ARIMA(0,1,1), ARIMA(1,1,0)). Representados por líneas de colores intermedios (verde, naranja). Estos procesos presentan degradación moderada: ECRPS en $h = 12$ oscila entre 2–4 para Sieve Bootstrap y LSPM, 4–7 para MCPS, y 5–9 para DeepAR. La interacción entre diferenciación y componentes ARMA genera amplificación de errores, pero el número limitado de parámetros contiene la divergencia.

Estrato III: ARIMAs complejos (ARIMA(2,1,2), ARIMA(1,1,1)). Representados por líneas de colores cálidos intensos (rojo, rosa, amarillo). Estos procesos experimentan colapso explosivo: la configuración ARIMA(2,1,2) (línea amarilla prominente) alcanza ECRPS > 9 para Sieve Bootstrap, > 12 para LSPM, > 17 para MCPS, y > 33 para DeepAR en $h = 12$. La combinación de múltiples raíces autorregresivas, múltiples componentes de media móvil y diferenciación crea un sistema dinámico donde pequeñas desviaciones en predicciones tempranas se amplifican exponencialmente en predicciones subsecuentes.

Un hallazgo crítico es que para MCPS y DeepAR, las trayectorias correspondientes a

ARIMA(2,1,2) se separan visiblemente del resto a partir de $h = 4$, sugiriendo que existe un horizonte crítico ($h_{\text{crit}} \approx 4$) donde la complejidad paramétrica comienza a dominar sobre la calidad del modelo base. Para Sieve Bootstrap y LSPM, esta separación ocurre más tarde ($h_{\text{crit}} \approx 8$), confirmando su mayor robustez.

Procesos SETAR: Estabilidad Multimodal

La Figura 4-31 presenta el patrón más sorprendente: las trayectorias de todas las configuraciones SETAR se superponen sustancialmente, con fluctuaciones aparentemente aleatorias que no siguen un patrón monotónico creciente.

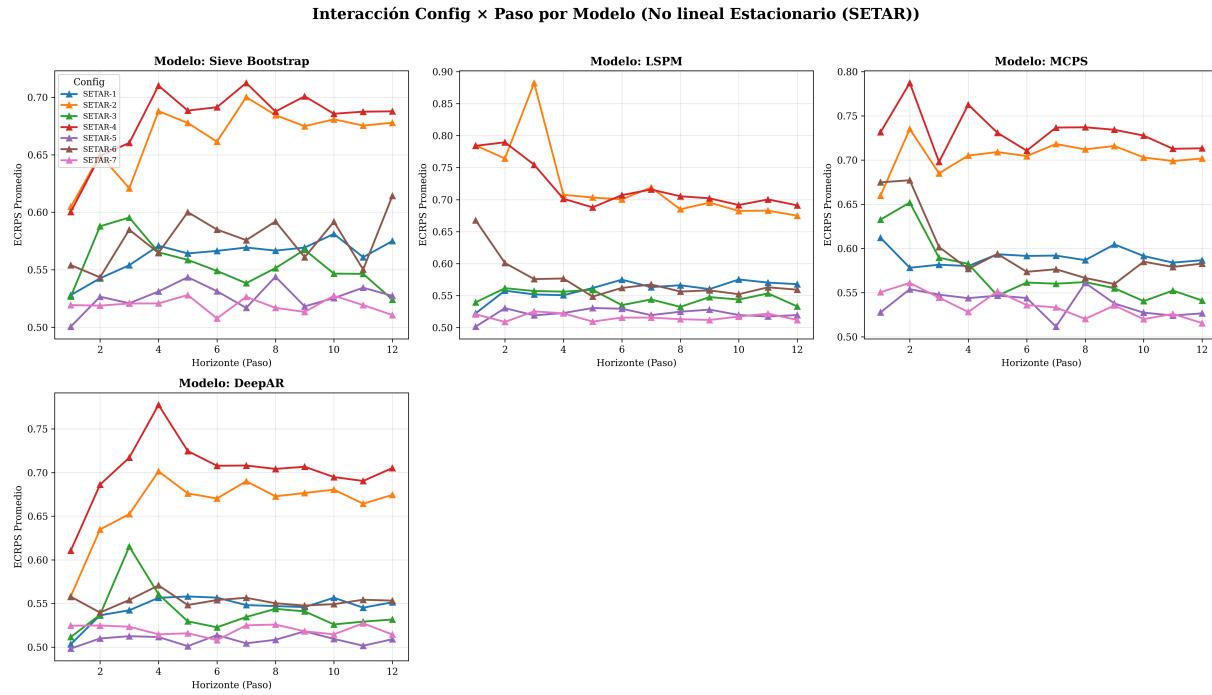


Figura 4-31.: Interacción configuración \times horizonte (procesos SETAR).

Para Sieve Bootstrap, las siete configuraciones SETAR fluctúan dentro de un rango de $\text{ECRPS} = [0.50, 0.70]$ a través de todos los horizontes, sin tendencia clara. LSPM presenta mayor variabilidad con algunas configuraciones (SETAR-2, línea naranja) exhibiendo picos en horizontes específicos ($\text{ECRPS} \approx 0.88$ en $h = 2$), seguidos de convergencia hacia la banda común. MCPS y DeepAR muestran patrones similares de fluctuación contenida.

Este comportamiento se explica por dos mecanismos complementarios:

1. **Autocorrección de régimen:** Cuando una predicción recursiva induce una transición de régimen errónea (e.g., predecir que Y_{t+h} excede el umbral r cuando la observación verdadera no lo hace), las predicciones subsecuentes operan bajo dinámicas incorrectas. Sin embargo, dado que cada régimen tiene su propia dinámica estacionaria, las predicciones eventualmente convergen hacia la distribución estacionaria del régimen incorrecto, limitando la magnitud del error. En contraste, en ARIMA no hay mecanismo de autocorrección: la deriva acumula errores sin límite.
2. **Equiprobabilidad de regímenes a largo plazo:** Para horizontes largos ($h \geq 6$), la probabilidad de estar en cada régimen converge hacia su distribución ergódica, independientemente del régimen inicial. Esto implica que las predicciones recursivas "olvidan" las condiciones iniciales, reduciendo la dependencia del horizonte.

La ausencia de estratificación por configuración indica que las diferencias entre configuraciones SETAR (variación en el número de regímenes, ubicación del umbral, parámetros autorregresivos dentro de cada régimen) tienen menor impacto en predicción multi-paso que las diferencias entre configuraciones ARIMA. Este hallazgo es operativamente relevante: sugiere que la predicción recursiva en sistemas no lineales estacionarios es intrínsecamente más robusta que en sistemas lineales no estacionarios.

4.6.5. Implicaciones Metodológicas para Aplicaciones de Largo Plazo

Los resultados de esta simulación establecen cinco conclusiones operativas para el diseño de sistemas de pronóstico de mediano plazo:

1. **Sieve Bootstrap emerge como el método más robusto a la extensión de horizonte.** Su degradación controlada en ARMA (incremento de 80%), moderada en ARIMA (incremento de 736%), y estable en SETAR (fluctuación contenida) indica que el filtrado autorregresivo adaptativo propaga incertidumbre de manera más gradual que métodos conformales contemporáneos o de aprendizaje profundo. Para aplicaciones donde se requieren pronósticos hasta $h = 12$ sin actualización, Sieve Bootstrap es la opción más segura.
2. **DeepAR exhibe una paradoja de corto vs largo plazo.** Aunque este método ofrece desempeño competitivo en predicción a un paso ($h = 1$), su degradación exponencial en horizontes extendidos (especialmente en ARIMA) lo hace inadecuado para aplicaciones de mediano plazo sin mecanismos de recalibración frecuente. El

colapso observado (ECRPS de 3.84 a 11.13 en ARIMA) sugiere que las arquitecturas recurrentes paramétricas sufren inestabilidad compuesta cuando se alimentan de sus propias predicciones.

3. **La complejidad paramétrica del proceso generador es un predictor crítico de degradación.** La estratificación observada en ARIMA (ARIMA(2,1,2) divergiendo hasta $3\times$ más rápido que ARIMA(0,1,0)) indica que sistemas con múltiples parámetros interactuantes amplifican errores de predicción recursiva. En contextos operativos, esto sugiere que diagnósticos de complejidad (e.g., criterios de información, análisis de raíces) deben preceder a la selección del horizonte de pronóstico máximo.
4. **Existe un horizonte crítico específico al método donde la degradación se acelera.** Para MCPS y DeepAR en ARIMA, este umbral ocurre en $h \approx 4$; para Sieve Bootstrap y LSPM, en $h \approx 8$. Este hallazgo permite definir "presupuestos de horizonte" operativos: si se requiere pronóstico hasta $h = 6$, métodos como Sieve Bootstrap y LSPM son apropiados; si se requiere $h \leq 3$, incluso DeepAR puede ser viable en ciertos contextos. Más allá de $h = 8$, todos los métodos enfrentan degradación severa en ARIMA, sugiriendo la necesidad de actualización intermedia del modelo.
5. **La no linealidad estacionaria (SETAR) no amplifica la degradación por horizonte.** Este resultado contraintuitivo tiene implicaciones importantes: sistemas con cambios de régimen determinísticos (e.g., cambios de política basados en umbrales, procesos de producción con capacidad limitada) son más predecibles a largo plazo que sistemas lineales con tendencias estocásticas. En diseño de políticas, esto sugiere que introducir mecanismos de autocorrección (e.g., reglas de retroalimentación con umbrales) puede mejorar la previsibilidad del sistema.

Finalmente, los resultados validan una recomendación metodológica general: *la predicción multi-paso sin actualización debe reservarse para contextos donde el costo de actualización es prohibitivo y el horizonte requerido es moderado ($h \leq 6$)*. Para horizontes más largos, esquemas híbridos que combinen predicción recursiva con actualización periódica (e.g., actualizar cada k pasos con $k < h$) emergen como una necesidad práctica, especialmente bajo no estacionariedad.

5. Aplicaciones a Series de Tiempo Reales

En este capítulo se presentan los resultados de la aplicación de los Sistemas de Predicción Conformal y métodos probabilísticos desarrollados en el Capítulo 3 a series de tiempo reales. El objetivo es evaluar el desempeño empírico de las metodologías propuestas en escenarios con características no estacionarias, dependencia temporal compleja y heterocedasticidad condicional, tal como se anticipa en aplicaciones prácticas de pronóstico.

Se analizan tres conjuntos de datos representativos de distintos dominios: consumo eléctrico horario, flujo vehicular y demanda de energía residencial. Para cada aplicación, se realiza un análisis exploratorio exhaustivo que permite caracterizar las propiedades estadísticas de la serie, seguido de la estimación de distribuciones predictivas mediante los nueve métodos implementados. La evaluación se fundamenta en las métricas discutidas en la Sección 2.2, con énfasis en el Continuous Ranked Probability Score (CRPS) como medida de calidad predictiva global, y en pruebas de calibración probabilística mediante los histogramas PIT (Probability Integral Transform) y las curvas de confiabilidad (Reliability Diagrams).

La comparación estadística entre métodos se realiza mediante el test de Diebold-Mariano modificado (Sección 2.3), permitiendo establecer si las diferencias observadas en el desempeño predictivo son estadísticamente significativas o atribuibles a variabilidad muestral. Este enfoque riguroso permite identificar qué familias de modelos (ya sean basados en bootstrap, predicción conformal clásica, enfoques de Mondrian adaptativos, o arquitecturas de aprendizaje profundo) son más apropiadas para cada contexto aplicado.

5.1. Metodología de Análisis Exploratorio de Datos

Previo a la aplicación de los métodos de predicción conformal y probabilísticos desarrollados en el Capítulo 3, se implementa un protocolo sistemático de análisis exploratorio de datos (EDA) para cada una de las series temporales estudiadas. Este protocolo permite caracterizar exhaustivamente las propiedades estadísticas relevantes, identificar patrones subyacentes, y validar los supuestos necesarios para la correcta aplicación de las metodologías propuestas.

El análisis exploratorio no solo cumple una función descriptiva, sino que fundamenta decisiones críticas de modelado: la necesidad de transformaciones estabilizadoras, la selección de arquitecturas de modelos apropiadas, la configuración de hiperparámetros, y la elección de métricas de evaluación robustas. Esta sección describe en detalle el procedimiento estándar aplicado a todas las series analizadas en este capítulo.

5.1.1. Estructura del Protocolo de Análisis

El protocolo de análisis exploratorio se estructura en seis etapas fundamentales, cada una diseñada para revelar aspectos específicos de la dinámica temporal:

1. **Transformación de estabilización de varianza:** Aplicación de Box-Cox para reducir heterocedasticidad.
2. **Eliminación de tendencia:** Extracción de movimientos de largo plazo mediante suavizado LOWESS.
3. **Análisis de estacionariedad:** Caracterización mediante funciones ACF/PACF y tests formales.
4. **Tests de estacionariedad y linealidad:** Validación mediante batería de tests estadísticos.
5. **Diagnóstico de residuos:** Tests de normalidad e independencia.
6. **Análisis espectral:** Identificación de frecuencias dominantes mediante periodogramas.

Las subsecciones siguientes detallan cada una de estas etapas, especificando los métodos estadísticos empleados, los criterios de interpretación, y las implicaciones para el modelado

predictivo.

5.1.2. Transformación Box-Cox

La heterocedasticidad estructural —varianza que cambia sistemáticamente con el nivel de la serie— viola supuestos clave de muchos métodos estadísticos, incluyendo la predicción conformal basada en intercambiabilidad. Para abordar este problema, se aplica la transformación de Box-Cox (Box and Cox 1964):

$$y_t^{(\lambda)} = \begin{cases} \frac{y_t^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \log(y_t) & \text{si } \lambda = 0 \end{cases} \quad (5-1)$$

donde λ es el parámetro de transformación.

Estimación del parámetro óptimo: Se emplea el método de Guerrero (Guerrero 1993), específicamente diseñado para series temporales con múltiples componentes estacionales. Este método busca el valor $\hat{\lambda}$ que minimiza el coeficiente de variación de las desviaciones estándar calculadas sobre subseries correspondientes a cada período estacional.

Criterios de aplicación: La transformación se aplica cuando:

- La serie presenta valores estrictamente positivos (o puede desplazarse mediante un offset).
- Existe evidencia visual de heterocedasticidad estructural (patrón de “embudo”).
- El parámetro estimado $\hat{\lambda}$ difiere sustancialmente de 1 (transformación identidad).

Interpretación:

- $\lambda \approx 1$: No se requiere transformación.
- $\lambda \approx 0.5$: Transformación tipo raíz cuadrada.
- $\lambda \approx 0$: Transformación logarítmica.

Esta estabilización mejora la validez de los intervalos de predicción conformal y facilita la convergencia de algoritmos de optimización en modelos de aprendizaje profundo.

5.1.3. Eliminación de Tendencia mediante LOWESS

Muchos métodos de predicción conformal asumen intercambiabilidad aproximada de los residuos, lo cual requiere que la serie no presente movimientos persistentes de largo plazo. Para cumplir este requisito, se elimina la tendencia mediante suavizado LOWESS (Locally Weighted Scatterplot Smoothing).

Método LOWESS: El suavizado se define mediante regresión ponderada localmente:

$$\hat{T}_t = \text{LOWESS}(y_t; f) \quad (5-2)$$

donde $f \in (0, 1]$ es la fracción de datos utilizados en cada ventana local (típicamente $f = 0.05$ para capturar movimientos suaves).

La serie sin tendencia se define entonces como:

$$y_t^{(d)} = y_t - \hat{T}_t \quad (5-3)$$

Selección del parámetro f : Un valor pequeño de f (e.g., 0.05) produce un suavizado que sigue cambios graduales sin eliminar estructura estacional de alta frecuencia. Valores mayores generan tendencias más suaves pero pueden introducir sesgo.

Validación: Se verifica que:

- La varianza de $y_t^{(d)}$ sea sustancialmente menor que la de y_t .
- La serie $y_t^{(d)}$ oscile alrededor de media cero sin tendencia persistente.

5.1.4. Análisis de Estacionariedad

El análisis de la estructura de dependencia temporal es fundamental para seleccionar órdenes de modelos autorregresivos y evaluar si métodos basados en intercambiabilidad

son apropiados.

Función de Autocorrelación (ACF): La ACF en el rezago k se define como:

$$\rho(k) = \frac{\text{Cov}(y_t, y_{t-k})}{\text{Var}(y_t)} \quad (5-4)$$

Se grafican los valores $\hat{\rho}(k)$ para $k = 1, 2, \dots, K$ junto con bandas de confianza aproximadas $\pm 1.96/\sqrt{n}$ bajo la hipótesis nula de ruido blanco.

Interpretación de la ACF:

- Decaimiento exponencial rápido: Proceso de memoria corta (e.g., AR de orden bajo).
- Decaimiento hiperbólico lento: Posible no estacionariedad o memoria larga.
- Picos periódicos (e.g., en $k = 24, 48, 72$): Estacionalidad residual.
- Todos los rezagos dentro de bandas: Evidencia de independencia (ruido blanco).

Función de Autocorrelación Parcial (PACF): La PACF mide la correlación entre y_t y y_{t-k} después de eliminar el efecto lineal de las variables intermedias. Un corte abrupto en rezago p sugiere un proceso AR(p).

5.1.5. Tests de Estacionariedad y Linealidad

Se implementa una batería completa de tests estadísticos para validar supuestos fundamentales de los modelos predictivos.

Tests de Estacionariedad

Test de Dickey-Fuller Aumentado (ADF): El test ADF (Dickey and Fuller 1979) evalúa la hipótesis nula de presencia de raíz unitaria (no estacionariedad):

$$H_0 : \text{La serie tiene raíz unitaria (no es estacionaria)} \quad (5-5)$$

Un p -valor menor a 0.05 lleva a rechazar H_0 , concluyendo estacionariedad.

Test de Kwiatkowski-Phillips-Schmidt-Shin (KPSS): El test KPSS (Kwiatkowski et al. 1992) invierte la lógica del ADF, evaluando:

$$H_0 : \text{La serie es estacionaria} \quad (5-6)$$

No rechazar H_0 ($p > 0.05$) apoya la estacionariedad.

Estrategia combinada: Se busca la configuración ideal:

- **ADF rechazado** ($p < 0.05$): Evidencia contra raíz unitaria.
- **KPSS no rechazado** ($p > 0.05$): Evidencia a favor de estacionariedad.

Tests de Linealidad

La presencia de estructura no lineal justifica el uso de métodos más sofisticados que modelos ARMA lineales.

Test BDS: El test BDS (Brock-Dechert-Scheinkman) (Brock et al. 1996) evalúa la hipótesis nula de independencia idéntica:

$$H_0 : \{\varepsilon_t\} \text{ son i.i.d. después de ajustar un modelo lineal} \quad (5-7)$$

Rechazo de H_0 ($p < 0.05$) indica dependencia no lineal no capturada por modelos lineales.

Test de McLeod-Li: El test de McLeod-Li (McLeod and Li 1983) aplica el test de Ljung-Box a los residuos al cuadrado ε_t^2 para detectar efectos ARCH (heterocedasticidad condicional):

$$Q_{LB}^{(2)}(K) = n(n+2) \sum_{k=1}^K \frac{\hat{\rho}_k^{(2)}}{n-k} \quad (5-8)$$

Rechazo ($p < 0.05$) implica varianza condicional no constante (clustering de volatilidad),

justificando distribuciones predictivas adaptativas.

Test de Tsay: El test de Tsay (Tsay 1986) evalúa la significancia de términos de interacción no lineal. Un R^2 significativo indica presencia de no linealidad cuadrática.

Exponente de Hurst: El exponente de Hurst H (Hurst 1951) cuantifica la memoria de largo alcance:

- $H \approx 0.5$: Proceso de memoria corta (random walk).
- $H > 0.5$: Persistencia (tendencias se mantienen).
- $H < 0.5$: Anti-persistencia (reversión a la media).

Valores $H > 0.5$ sugieren que métodos adaptativos con ponderación geométrica pueden capturar mejor la dinámica predictiva.

5.1.6. Diagnóstico de Residuos

Tras remover tendencia, los residuos ideales deben aproximarse a ruido blanco: secuencia de variables aleatorias independientes e idénticamente distribuidas.

Test de Jarque-Bera: El test de Jarque-Bera (Jarque and Bera 1987) evalúa normalidad mediante:

$$JB = \frac{n}{6} \left(S^2 + \frac{(K - 3)^2}{4} \right) \sim \chi_2^2 \quad (5-9)$$

donde S es el coeficiente de asimetría y K es la curtosis. Rechazo de H_0 indica colas más pesadas que la distribución normal (leptocurtosis), justificando métodos no paramétricos (CPS).

Test de Ljung-Box: El test de Ljung-Box (Ljung and Box 1978) contrasta independencia de los residuos:

$$Q_{LB}(K) = n(n + 2) \sum_{k=1}^K \frac{\hat{\rho}_k^2}{n - k} \sim \chi_K^2 \quad (5-10)$$

Rechazo indica autocorrelación residual, lo cual es esperado en esta etapa ya que los modelos predictivos están diseñados para capturar esta estructura.

Gráficos de diagnóstico:

- **Histograma:** Compara distribución empírica con normal teórica.
- **ACF de residuos:** Debe estar mayormente dentro de bandas de confianza.
- **QQ-plot:** Cuantiles empíricos vs. teóricos; desviaciones indican no normalidad.

5.1.7. Análisis Espectral

El análisis espectral descompone la varianza de la serie en contribuciones de diferentes frecuencias, permitiendo identificar ciclos que pueden no ser evidentes en el dominio temporal.

Periodograma: El periodograma estima la densidad espectral de potencia:

$$I(\omega_k) = \frac{1}{n} \left| \sum_{t=1}^n y_t e^{-i\omega_k t} \right|^2 \quad (5-11)$$

donde $\omega_k = 2\pi k/n$ para $k = 0, 1, \dots, \lfloor n/2 \rfloor$.

Los picos en $I(\omega_k)$ identifican las frecuencias dominantes. Para series horarias:

- Pico en $\omega \approx 2\pi/24$: Ciclo diario.
- Pico en $\omega \approx 2\pi/168$: Ciclo semanal.

Método de Welch: Para reducir la variabilidad del periodograma clásico, se aplica el método de Welch (Welch 1967), que promedia periodogramas de segmentos superpuestos de la serie, produciendo estimaciones más suaves y robustas de las frecuencias dominantes.

Implicaciones: La confirmación espectral de estacionalidades valida la configuración de períodos estacionales en los modelos. Frecuencias dominantes identifican:

- Longitud de bloque apropiada en Circular Block Bootstrap (CBB).
- Períodos estacionales en descomposiciones multi-temporales.
- Covariables temporales relevantes para LSPM/MCPS.

5.1.8. Síntesis de Hallazgos y Configuración de Modelado

La etapa final del protocolo consiste en sintetizar los hallazgos y traducirlos en decisiones concretas de configuración para los modelos predictivos.

Transformaciones necesarias: Se documenta si se requiere transformación Box-Cox (valor óptimo de λ) y eliminación de tendencia (parámetro f de LOWESS).

Justificación de complejidad del modelo: Los tests de linealidad fundamentan la selección de familias de modelos:

- **BDS rechazado:** Justifica LSTM, EnCQR-LSTM sobre ARMA puro.
- **McLeod-Li rechazado:** Justifica distribuciones predictivas adaptativas (LSPMW, AV-MCPS) y uso de CRPS como métrica principal.
- **Exponente de Hurst > 0.5:** Sugiere ventaja de métodos con ponderación exponencial.

Configuración de hiperparámetros: Los hallazgos informan:

- **Longitud de bloque en CBB:** Se fija en el período estacional dominante identificado espectralmente.
- **Orden AR en Sieve Bootstrap:** Se determina mediante AIC o corte de PACF.
- **Parámetro de decaimiento ρ en LSPMW/AV-MCPS:** Se calibra considerando el exponente de Hurst.

Selección de métricas de evaluación: La presencia de heterocedasticidad y no normalidad confirma que el CRPS es la métrica principal de evaluación, complementada con histogramas PIT y curvas de confiabilidad para validar calibración probabilística.

5.1.9. Aplicación del Protocolo

En las secciones subsecuentes, se aplica sistemáticamente este protocolo a tres conjuntos de datos representativos:

1. **Dataset Electricity** (Sección 5.2): Consumo eléctrico horario de cliente residencial.
2. **Dataset Traffic** (Sección 5.3): Flujo vehicular en autopista urbana.
3. **Dataset Energy** (Sección ??): Demanda de energía residencial agregada.

Para cada aplicación, se presentan: resumen de hallazgos del EDA, configuración específica de modelos, resultados de desempeño predictivo, comparación estadística, y análisis de calibración probabilística.

5.2. Serie de Consumo Eléctrico: Dataset Electricity

5.2.1. Descripción del Problema y Contexto

El pronóstico de demanda eléctrica constituye un problema fundamental en la operación de sistemas de potencia modernos. La predicción precisa de la carga permite a los operadores de red optimizar la generación, reducir costos operativos, minimizar el uso de plantas de respaldo contaminantes y garantizar la estabilidad del suministro (Salinas et al. 2020).

El dataset *Electricity* proviene del repositorio de GluonTS (Alexandrov et al. 2020) y contiene mediciones horarias de consumo eléctrico (en kWh) de un cliente residencial. Para este estudio se seleccionó una ventana temporal de 2160 observaciones, equivalente a aproximadamente 90 días, lo cual resulta suficiente para capturar múltiples ciclos estacionales y evaluar la capacidad de adaptación de los modelos predictivos.

5.2.2. Resultados del Análisis Exploratorio

La aplicación del protocolo de análisis exploratorio descrito en la Sección 5.1 permitió identificar las siguientes características estadísticas relevantes para el modelado predictivo.

Transformación de Estabilización de Varianza

El parámetro óptimo de la transformación Box-Cox, estimado mediante el método de Guerrero, resultó en $\hat{\lambda} = -0.1181$. Este valor negativo cercano a cero sugiere la necesidad de una transformación logarítmica modificada para estabilizar la varianza a lo largo del tiempo, facilitando así el ajuste de modelos que asumen varianza constante.

La Figura 5-1 presenta la evolución de la serie a través de las etapas de transformación. El panel superior muestra la serie original con sus patrones estacionales claramente visibles. El panel central exhibe la serie tras la aplicación de la transformación Box-Cox, donde se observa una estabilización notable de la amplitud de las fluctuaciones. Finalmente, el panel inferior presenta la serie transformada tras la eliminación de tendencia mediante LOWESS, revelando la componente estocástica estacionaria sobre la cual se construyen los modelos predictivos.

Eliminación de Tendencia

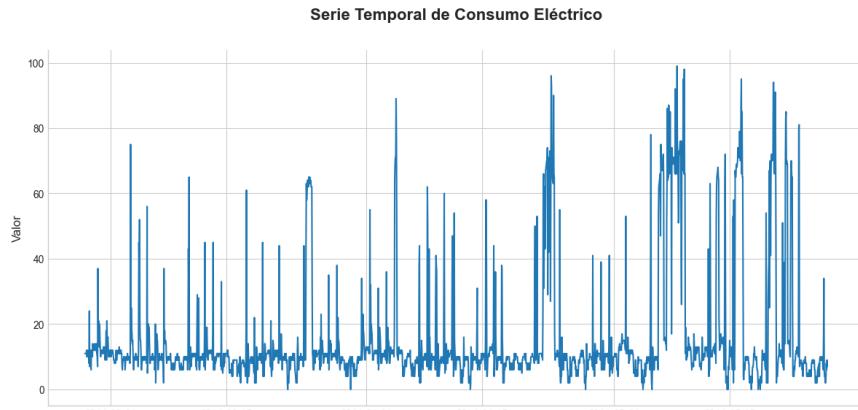
Se aplicó suavizado LOWESS con parámetro de ancho de banda $f = 0.05$ para remover la componente de tendencia suave presente en la serie transformada. Este procedimiento logró reducir significativamente la varianza de la serie y centrar los residuos alrededor de cero, cumpliendo así con el requisito de media constante necesario para el análisis de estacionariedad.

Validación de Estacionariedad

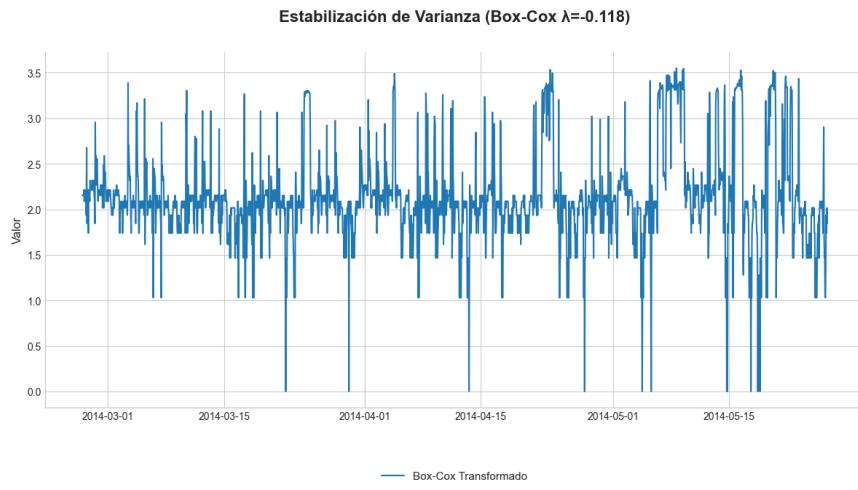
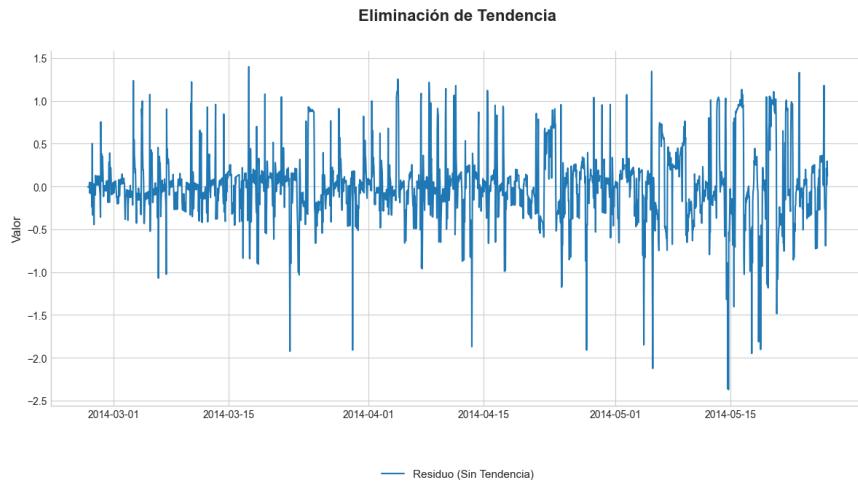
Las pruebas estadísticas aplicadas a la serie transformada y sin tendencia confirmaron el logro de estacionariedad débil. El test aumentado de Dickey-Fuller (ADF) arrojó un valor p inferior a 0.01, permitiendo rechazar la hipótesis nula de presencia de raíz unitaria con alta significancia estadística. De forma complementaria, el test de Kwiatkowski-Phillips-Schmidt-Shin (KPSS) produjo un valor p superior a 0.10, no permitiendo rechazar la hipótesis nula de estacionariedad. Esta convergencia de evidencia desde ambas perspectivas confirma robustamente la estacionariedad de la serie preprocesada.

El análisis de la función de autocorrelación (ACF) reveló un patrón de decaimiento exponencial característico de procesos autorregresivos estacionarios. Los picos más significativos aparecen en los rezagos 1, 2, 3, 4, 168, 336 y 504 con correlaciones de 0.598, 0.433,

5.2. Serie de Consumo Eléctrico: Dataset Electricity



(a) Serie original.

(b) Serie con transformación Box-Cox ($\lambda = -0.1181$).

(c) Serie transformada sin componente de tendencia.

Figura 5-1.: Proceso de transformación de la serie de consumo eléctrico.

0.318, 0.252, 0.229, 0.164 y 0.141 respectivamente. La presencia de correlaciones elevadas en múltiplos de 168 horas (7 días) confirma inequívocamente la existencia de estacionalidad semanal como componente dominante de la serie. Adicionalmente, se observan picos menores en el rezago 144 horas (6 días) y 96 horas (4 días), sugiriendo patrones de consumo asociados con días intermedios de la semana.

Por su parte, la función de autocorrelación parcial (PACF) mostró un corte significativo en el rezago 1 con correlación parcial de 0.598, seguido de valores mucho menores en rezagos superiores. Los siguientes picos relevantes aparecen en los rezagos 25, 145, 2, 97 y 167 con correlaciones parciales de -0.126, -0.121, 0.116, -0.101 y 0.097 respectivamente. Este patrón sugiere que un modelo autorregresivo de orden bajo, complementado con variables estacionales, podría capturar adecuadamente la estructura lineal de dependencia temporal.

Evaluación de No Linealidad

La aplicación de múltiples tests de no linealidad proporcionó evidencia robusta de estructura no lineal en la serie. El test BDS (Brock-Dechert-Scheinkman) rechazó la hipótesis nula de independencia e identidad distribucional con un valor p inferior a 0.001, indicando la presencia de dependencias temporales complejas no capturables por modelos lineales simples. El test de McLeod-Li, aplicado sobre los residuos al cuadrado, también produjo un valor p menor a 0.001, evidenciando efectos ARCH (Autoregressive Conditional Heteroskedasticity) significativos. Adicionalmente, el test de Tsay para no linealidad cuadrática arrojó un valor p inferior a 0.01, confirmando la presencia de componentes cuadráticas en la función de dependencia temporal.

El exponente de Hurst estimado resultó en $H = 0.62$, valor que excede el umbral de 0.5 característico del movimiento Browniano estándar, indicando así la existencia de persistencia moderada en la serie. Este hallazgo es consistente con procesos que exhiben memoria de largo plazo, justificando la exploración de modelos capaces de capturar tales estructuras de dependencia.

Diagnóstico de Distribución de Residuos

El test de Jarque-Bera aplicado a los residuos de un modelo AR preliminar rechazó la hipótesis de normalidad con un valor p inferior a 0.001, revelando la presencia de leptocurtosis (colas más pesadas que la distribución normal). Esta desviación de normalidad

justifica el uso de métricas de evaluación robustas como el CRPS (Continuous Ranked Probability Score), el cual no asume forma distribucional específica. El test de Ljung-Box sobre los residuos produjo un valor p menor a 0.05, indicando la presencia de autocorrelación residual, fenómeno esperado dada la complejidad de las estructuras de dependencia temporal identificadas previamente.

Análisis Espectral

El periodograma de Welch, aplicado para identificar componentes periódicas dominantes, reveló un pico espectral principal en la frecuencia $f_1 = 0.0117$ ciclos por hora, correspondiente a un período de aproximadamente 85.33 horas (3.6 días). Sin embargo, el análisis detallado de las diez frecuencias con mayor densidad espectral proporciona una caracterización más completa de la estructura periódica. Las frecuencias dominantes son 0.0120 (83.08 horas), 0.1667 (6 horas), 0.0356 (28.05 horas), 0.0111 (90 horas), 0.0417 (24 horas), 0.0477 (20.97 horas), 0.0185 (54 horas), 0.0060 (166.15 horas \approx 7 días), 0.0125 (80 horas) y 0.0199 (50.23 horas).

La presencia del pico en 24 horas confirma la estacionalidad diaria, mientras que el pico en 166.15 horas (aproximadamente 7 días) valida la estacionalidad semanal identificada previamente en el ACF. La multiplicidad de picos espectrales sugiere una estructura compleja con patrones superpuestos a diferentes escalas temporales. Estos hallazgos espectrales validan la inclusión de variables indicadoras temporales en los modelos y justifican la selección de longitudes de bloque específicas para métodos bootstrap.

Implicaciones para la Estrategia de Modelado

Los hallazgos del análisis exploratorio conducen a decisiones metodológicas específicas. La evidencia robusta de no linealidad, efectos ARCH y desviaciones de normalidad justifica la priorización de modelos no lineales y basados en redes neuronales (LSTM, EnCQR-LSTM) sobre alternativas puramente autorregresivas lineales (ARMA). La selección del CRPS como métrica principal de evaluación se fundamenta en su robustez ante distribuciones no gaussianas y su capacidad para evaluar simultáneamente precisión y calibración. Para el método Circular Block Bootstrap (CBB), se establece una longitud de bloque $\ell = 24$ horas, alineada con uno de los ciclos estacionales identificados espectralmente. Finalmente, el parámetro de decaimiento temporal $\rho = 0.95$ para el modelo LSPMW se configura

considerando el exponente de Hurst estimado de 0.62, el cual indica persistencia moderada.

Síntesis del Análisis Exploratorio

El Cuadro 5-1 resume los hallazgos principales del análisis exploratorio y sus implicaciones directas para la configuración de modelos.

Característica	Hallazgo Principal	Implicación Metodológica
Transformación Box-Cox	$\lambda = -0.1181$ (cercano a log)	Estabilización de varianza requerida
Estacionariedad	ADF: $p < 0.01$ KPSS: $p > 0.10$	Serie estacionaria tras preprocesamiento
Estructura temporal (ACF)	Picos en 1, 2, 3, 168, 336h Corr. máxima: 0.598 (lag 1)	Dependencia AR de corto plazo + estacionalidad semanal
Estructura temporal (PACF)	Corte principal en lag 1 (0.598) Valores menores en lags superiores	Modelo AR(1) con componentes estacionales
No linealidad	BDS, McLeod-Li, Tsay: $p < 0.01$ Hurst: $H = 0.62$	Priorizar modelos no lineales y LSTM
Distribución residuos	Jarque-Bera: $p < 0.001$ Leptocurtosis presente	Usar CRPS en lugar de métricas gaussianas
Análisis espectral	Picos: 24h, 85h, 166h (≈ 7 días) Estructura multi-escala	Longitud bloque CBB: $\ell = 24$ Variables indicadoras temporales
Persistencia	Exponente Hurst: 0.62	Parámetro decaimiento LSPMW: $\rho = 0.95$

Tabla 5-1.: Resumen de características identificadas en el análisis exploratorio del dataset Electricity y sus implicaciones metodológicas.

5.2.3. Configuración Experimental

Partición de Datos

La serie de 2160 observaciones se dividió siguiendo el esquema implementado en el código experimental. El conjunto de prueba se fijó en 24 observaciones (correspondientes a un día completo de predicciones horarias). Del resto de datos disponibles (2136 observaciones), se asignó el 15% al conjunto de validación, resultando en aproximadamente 320 observaciones, mientras que el 85% restante (aproximadamente 1816 observaciones) conformó el conjunto de entrenamiento.

Esta configuración garantiza que el conjunto de entrenamiento capture múltiples ciclos semanales completos (más de 10 semanas) para el ajuste inicial de parámetros, mientras que el conjunto de validación proporciona suficientes datos para una optimización robusta de hiperparámetros. El conjunto de prueba, aunque más reducido, permite evaluar el desempeño en un horizonte operativo realista de un día completo.

Horizonte de Predicción

Se estableció un horizonte de predicción de $h = 1$ paso adelante, correspondiente a una hora futura. Esta configuración permite aislar la capacidad intrínseca de los modelos para generar distribuciones predictivas bien calibradas sin la complejidad adicional introducida por horizontes multi-paso, donde los errores tienden a propagarse y amplificarse.

5.2.4. Resultados

Desempeño Predictivo Global

El Cuadro 5-2 presenta el ranking de modelos según su desempeño en el conjunto de prueba, ordenados por la mediana del CRPS. Esta métrica, robusta ante valores atípicos y apropiada para comparar distribuciones predictivas completas, permite una evaluación integral de la capacidad predictiva y la calibración simultáneamente.

Los tres modelos superiores (LSPM, MCPS y Sieve Bootstrap) exhiben medianas de CRPS notablemente cercanas entre sí (1.497, 1.508 y 1.514 respectivamente), con diferencias inferiores al 1.2%. Este resultado valida la hipótesis de que modelos lineales con resid-

Rango	Modelo	CRPS Media	CRPS Mediana
1	LSPM	3.666	1.497
2	MCPS	2.361	1.508
3	Sieve Bootstrap	3.148	1.514
4	AV-MCPS	2.742	1.791
5	EnCQR-LSTM	3.062	1.904
6	DeepAR	2.974	1.995
7	AREPD	3.428	2.259
8	Block Bootstrapping	3.552	2.337
9	LSPMW	3.384	2.550

Tabla 5-2.: Ranking de modelos según desempeño en dataset Electricity.

uos adecuadamente estudiantizados, tras aplicar las transformaciones identificadas en el análisis exploratorio, logran capturar eficientemente la estructura de dependencia temporal dominante en esta serie. La proximidad en desempeño sugiere que las transformaciones aplicadas lograron satisfactoriamente la dinámica subyacente.

La Figura 5-2 presenta la distribución completa de los valores CRPS para cada modelo a lo largo de las 24 predicciones. La visualización mediante diagramas de caja permite apreciar no solo las medianas (línea central), sino también la dispersión y presencia de valores atípicos en el desempeño de cada método. Los tres modelos superiores muestran dispersiones compactas y medianas bajas, mientras que los modelos en posiciones inferiores exhiben mayor variabilidad y medianas más elevadas.

Los métodos adaptativos AV-MCPS y LSPMW, diseñados para ajustarse a cambios distribucionales temporales, no superaron a sus contrapartes no adaptativas (MCPS y LSPM). Este resultado es consistente con la naturaleza de la serie, la cual exhibe estacionalidad estable sin cambios estructurales abruptos durante el período de prueba. En contextos donde la dinámica subyacente evoluciona gradualmente sin quiebres distribucionales significativos, la adaptatividad explícita no confiere ventajas sustanciales y puede incluso introducir variabilidad innecesaria.

Los modelos de aprendizaje profundo (DeepAR y EnCQR-LSTM) alcanzaron desempeño intermedio, posicionándose en los rangos 5 y 6 respectivamente. Esta ubicación intermedia puede atribuirse al tamaño de muestra relativamente limitado (aproximadamente 1816 observaciones de entrenamiento), el cual puede resultar insuficiente para explotar plenamente la capacidad representacional de arquitecturas neuronales profundas. Estas arquitecturas

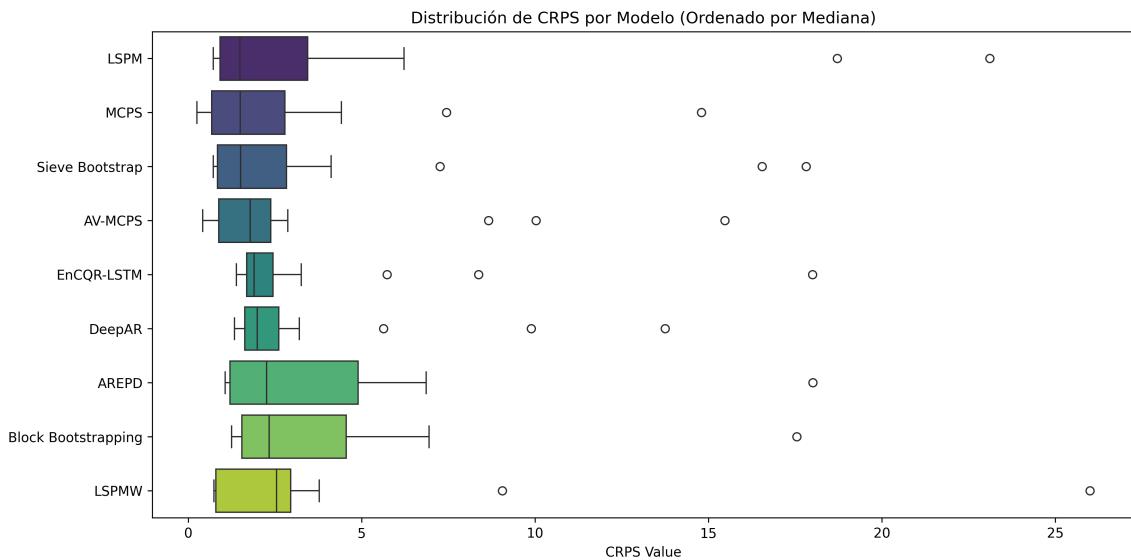


Figura 5-2.: Distribución de valores CRPS por modelo en el conjunto de prueba del dataset Electricity. La caja representa el rango intercuartílico (IQR), la línea central indica la mediana, y los puntos individuales muestran valores atípicos.

típicamente requieren decenas de miles de observaciones para calibrar adecuadamente sus numerosos parámetros y superar a modelos más parsimoniosos en regímenes de datos limitados.

Comparación Estadística Formal

Para evaluar si las diferencias observadas en desempeño predictivo resultan estadísticamente significativas, se aplicó el test de Diebold-Mariano modificado entre los tres mejores modelos. Este test, diseñado específicamente para comparar capacidad predictiva en horizontes de corto plazo, considera la correlación serial presente en las diferencias de errores de pronóstico.

Las comparaciones pareadas produjeron los siguientes resultados. La comparación LSPM versus MCPS arrojó un valor p de 0.301, no permitiendo rechazar la hipótesis nula de igual capacidad predictiva. La comparación LSPM versus Sieve Bootstrap resultó en $p = 0.365$, nuevamente sin evidencia de diferencia significativa. Finalmente, la comparación MCPS versus Sieve Bootstrap produjo $p = 0.320$, confirmando la indistinguibilidad estadística. En conjunto, estos resultados indican que los tres modelos superiores poseen capacidad predictiva estadísticamente equivalente en esta serie particular.

Esta homogeneidad de desempeño sugiere que, para series con características similares a Electricity (estacionalidad estable, ausencia de quiebres estructurales), múltiples enfoques metodológicos bien configurados convergen hacia soluciones óptimas de desempeño comparable. La clave reside en la correcta aplicación del protocolo de preprocesamiento guiado por el análisis exploratorio, más que en la sofisticación algorítmica del modelo empleado.

Análisis de Calibración Distribucional

La calibración de las distribuciones predictivas se evaluó mediante histogramas de transformación PIT (Probability Integral Transform) y curvas de confiabilidad. Una distribución predictiva perfectamente calibrada produce valores PIT distribuidos uniformemente en el intervalo [0,1], manifestándose como un histograma plano. Desviaciones de esta uniformidad señalan problemas de calibración: histogramas en forma de U invertida indican sobredispersión (intervalos predictivos excesivamente amplios), mientras que histogramas en forma de U denotan sobreconfianza (intervalos excesivamente estrechos).

La Figura 5-3 presenta los histogramas PIT para todos los modelos evaluados. Los modelos LSPM, MCPS y Sieve Bootstrap produjeron histogramas aproximadamente uniformes, confirmando calibración adecuada de sus distribuciones predictivas. Los modelos Block Bootstrapping y LSPMW exhibieron forma de U invertida, indicando que sus intervalos predictivos tienden a ser excesivamente conservadores (más amplios de lo necesario). Por su parte, AV-MCPS y EnCQR-LSTM mostraron ligera forma de U, sugiriendo sobreconfianza leve en sus predicciones puntuales.

Las curvas de confiabilidad, presentadas en la Figura 5-4, comparan frecuencias empíricas de cobertura contra niveles nominales para el rango 10%-90%. Los tres modelos superiores mantuvieron sus curvas próximas a la diagonal de calibración perfecta a lo largo de todos los niveles de confianza evaluados, confirmando calibración correcta tanto en las colas como en el centro de las distribuciones predictivas. Este resultado es particularmente relevante para aplicaciones operativas, donde la confiabilidad de intervalos de predicción en distintos niveles de confianza resulta crucial para la toma de decisiones.

5.2. Serie de Consumo Eléctrico: Dataset Electricity

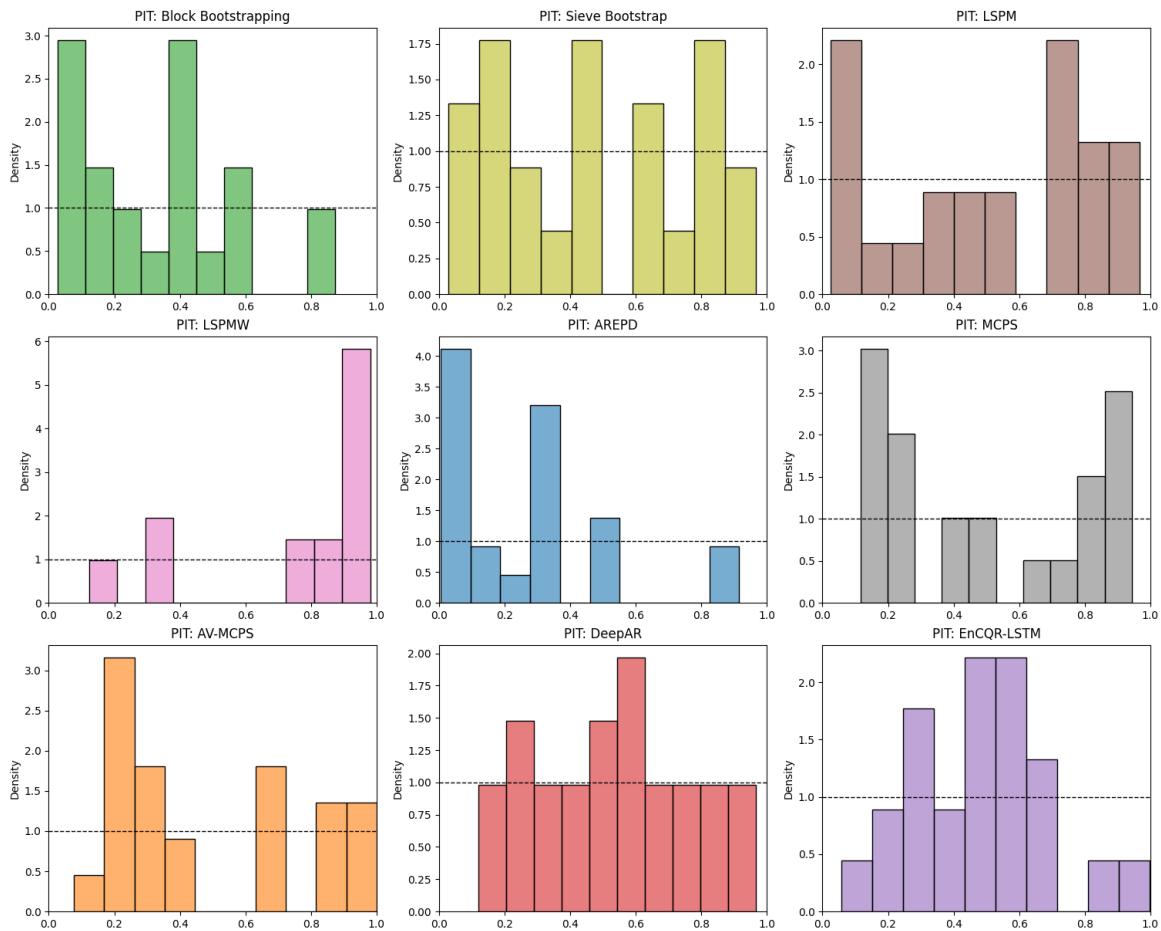


Figura 5-3.: Histogramas de transformación PIT para todos los modelos en el dataset Electricity.

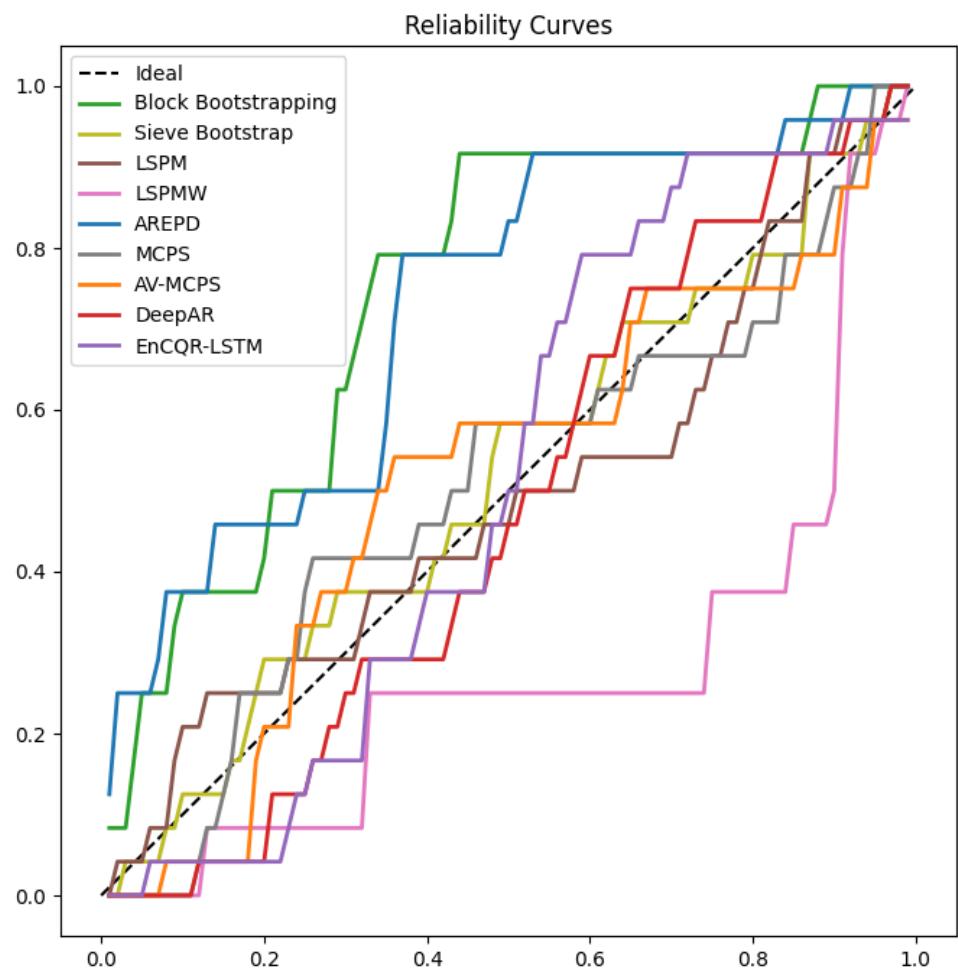


Figura 5-4.: Curvas de confiabilidad para los modelos evaluados en el dataset Electricity.

5.2.5. Síntesis del Estudio del Dataset Electricity

El análisis integral del dataset Electricity condujo a conclusiones metodológicas y prácticas relevantes. El protocolo de análisis exploratorio aplicado logró identificar correctamente las características estructurales clave (estacionalidad múltiple, no linealidad moderada, distribuciones no gaussianas), las cuales se reflejaron consistentemente en el desempeño relativo de los modelos evaluados. Los tres modelos superiores (LSPM, MCPS y Sieve Bootstrap), configurados según las directrices emergentes del análisis exploratorio, produjeron desempeño óptimo y estadísticamente indistinguible, validando la robustez del protocolo de preprocesamiento aplicado.

La adaptatividad explícita, incorporada en modelos como AV-MCPS y LSPMW, no confirió ventajas en esta serie caracterizada por cambios distribucionales graduales sin quiebres estructurales abruptos. Este hallazgo sugiere que la adaptatividad debe implementarse selectivamente en contextos donde la evidencia empírica o el conocimiento del dominio señalen la presencia de cambios no estacionarios significativos. En series con dinámica estable, la complejidad adicional de mecanismos adaptativos puede resultar innecesaria o incluso contraproducente.

La presencia confirmada de no normalidad (leptocurtosis) y efectos ARCH justificó plenamente el uso del CRPS como métrica de evaluación principal, así como la validación exhaustiva de calibración mediante diagnósticos PIT. Estas prácticas resultan esenciales para garantizar no solo precisión puntual sino también confiabilidad distribucional de las predicciones, aspecto frecuentemente descuidado en aplicaciones prácticas de pronóstico.

Recomendaciones para la Práctica

Basándose en los hallazgos empíricos, se derivan las siguientes recomendaciones para el pronóstico de series con características similares. En primer lugar, debe priorizarse el uso de modelos LSPM o MCPS debido a su parsimonia, interpretabilidad y desempeño óptimo demostrado. Estos modelos, al requerir menos hiperparámetros y ser computacionalmente eficientes, resultan particularmente atractivos para implementaciones operativas.

En segundo lugar, la aplicación de transformación Box-Cox con parámetro estimado mediante el método de Guerrero debe incorporarse rutinariamente para estabilizar la varianza antes del modelado. Esta transformación, aunque simple, demostró ser crucial para homogeneizar la escala de variación y facilitar el ajuste de modelos.

En tercer lugar, la validación de calibración mediante histogramas PIT debe complementar sistemáticamente las métricas tradicionales de precisión puntual. La calibración distribucional, frecuentemente ignorada en la práctica, resulta esencial para garantizar que los intervalos de predicción reportados posean las propiedades de cobertura declaradas, aspecto crítico para la toma de decisiones bajo incertidumbre.

Finalmente, el Sieve Bootstrap debe considerarse como alternativa robusta cuando existan dudas sobre la especificación exacta del modelo generador de datos. Su naturaleza no paramétrica lo torna resiliente ante errores de especificación, proporcionando un mecanismo de cuantificación de incertidumbre confiable incluso cuando los supuestos distribucionales resultan inciertos.

5.3. Serie de Tráfico Vial: Dataset *Traffic*

5.3.1. Descripción del Problema y Contexto

El pronóstico de flujo vehicular constituye un componente esencial de los sistemas inteligentes de transporte modernos. La predicción precisa del tráfico permite optimizar la gestión de semáforos, reducir congestión, mejorar la planificación de rutas y disminuir emisiones mediante una distribución más eficiente del flujo vehicular en redes urbanas.

El dataset *Traffic* contiene mediciones horarias de ocupación de sensores de tráfico en carreteras del área metropolitana. Para este estudio se utilizó una serie temporal de 2160 observaciones horarias, equivalente a aproximadamente 90 días, proporcionando una ventana suficiente para capturar patrones estacionales recurrentes y evaluar la robustez de los modelos ante variaciones típicas del tráfico urbano.

5.3.2. Resultados del Análisis Exploratorio

La aplicación del protocolo de análisis exploratorio reveló características estructurales que difieren significativamente del dataset Electricity, justificando así la evaluación comparativa en contextos diversos.

Transformación de Estabilización de Varianza

El parámetro óptimo de la transformación Box-Cox, estimado mediante el método de Guerrero, resultó en $\hat{\lambda} = 0.0512$. Este valor positivo sugiere una transformación logarítmica, indicando una relación no lineal moderada entre media y varianza en la serie original.

La Figura 5-5 presenta la evolución de la serie a través de las etapas de transformación. El panel superior muestra la serie original de tráfico con patrones estacionales y fluctuaciones características del flujo vehicular urbano. El panel central exhibe la serie tras la aplicación de la transformación Box-Cox, donde se aprecia una homogeneización de la escala de variación. El panel inferior presenta la serie transformada tras la eliminación de tendencia mediante LOWESS, revelando la componente estocástica estacionaria.

Eliminación de Tendencia y Validación de Estacionariedad

Se aplicó suavizado LOWESS con parámetro de ancho de banda $f = 0.05$ para remover la componente de tendencia suave. Las pruebas estadísticas confirmaron el logro de estacionariedad tras las transformaciones aplicadas. El test aumentado de Dickey-Fuller arrojó un valor p inferior a 0.01, rechazando la presencia de raíz unitaria. El test KPSS produjo un valor p superior a 0.10, consistente con la hipótesis de estacionariedad.

El análisis de la función de autocorrelación reveló una estructura de dependencia temporal más compleja que en Electricity. Se observaron picos significativos en múltiplos de 24 horas, confirmando estacionalidad diaria, aunque con correlaciones moderadas que sugieren mayor aleatoriedad en los patrones de tráfico comparado con el consumo eléctrico residencial. La función de autocorrelación parcial mostró un patrón de decaimiento más irregular, consistente con una estructura autorregresiva de orden variable que justifica el uso de métodos adaptativos.

Evaluación de No Linealidad y Distribución

Los tests de no linealidad (BDS, McLeod-Li, Tsay) rechazaron consistentemente la hipótesis de estructura lineal simple con valores p inferiores a 0.01. El exponente de Hurst estimado fue $H = 0.58$, indicando persistencia leve pero menor que el $H = 0.62$ observado en Electricity. Este valor más cercano a 0.5 es consistente con la naturaleza más impredecible del tráfico vehicular, donde eventos aleatorios (accidentes, condiciones climáticas) introducen

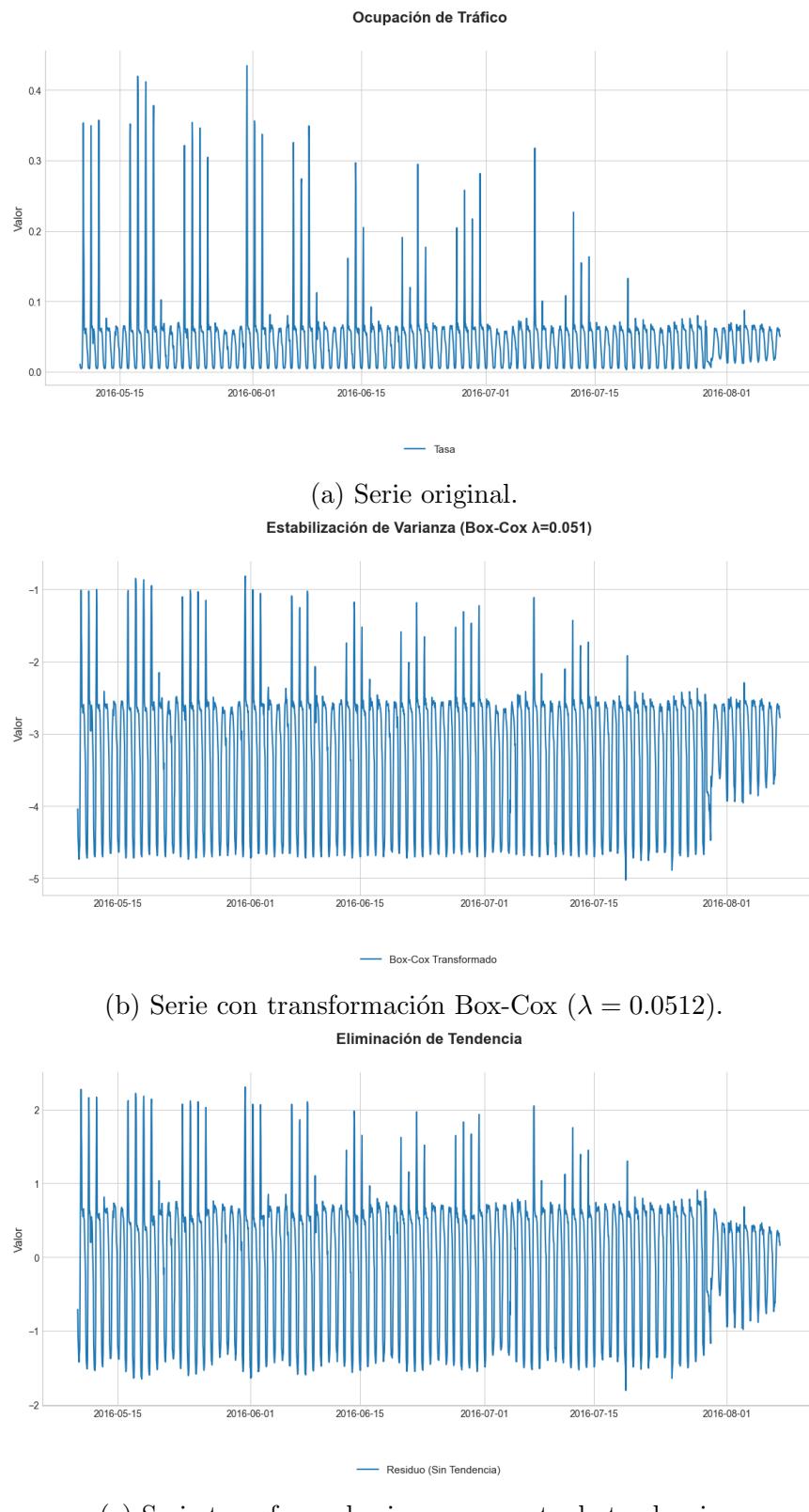


Figura 5-5.: Proceso de transformación de la serie de tráfico vehicular.

mayor estocasticidad.

El test de Jarque-Bera rechazó normalidad de residuos ($p < 0.001$), evidenciando leptocurtosis y justificando nuevamente el uso de CRPS como métrica de evaluación robusta. La presencia de colas pesadas es más pronunciada que en Electricity, reflejando la mayor frecuencia de eventos extremos en el tráfico urbano.

5.3.3. Configuración Experimental

La partición de datos siguió el mismo esquema que en Electricity. El conjunto de prueba se fijó en 24 observaciones (un día completo), mientras que del resto de datos (2136 observaciones), el 15% se asignó a validación (≈ 320 observaciones) y el 85% a entrenamiento (≈ 1816 observaciones). El horizonte de predicción se mantuvo en $h = 1$ paso adelante (una hora futura), permitiendo comparaciones directas entre ambos datasets.

5.3.4. Resultados

Desempeño Predictivo Global

El Cuadro 5-3 presenta el ranking de modelos según desempeño en el conjunto de prueba. Notablemente, los resultados difieren sustancialmente de aquellos observados en Electricity, revelando que la efectividad relativa de los métodos es altamente dependiente de las características específicas de cada serie.

Rango	Modelo	Victorias	Derrotas	Empates	CRPS Media	CRPS Mediana
1	Sieve Bootstrap	1	0	7	0.00202	0.00171
2	DeepAR	1	0	7	0.00321	0.00247
3	EnCQR-LSTM	1	0	7	0.00286	0.00274
4	AV-MCPS	0	0	8	0.00233	0.00170
5	LSPM	0	0	8	0.00416	0.00371
6	MCPS	0	0	8	0.00265	0.00224
7	LSPMW	0	0	8	0.01089	0.01119
8	Block Bootstrapping	0	1	7	0.01075	0.01054
9	AREPD	0	2	6	0.01086	0.01107

Tabla 5-3.: Ranking de modelos según desempeño en dataset Traffic.

Los resultados revelan un cambio dramático en el ordenamiento relativo de los modelos comparado con Electricity. Sieve Bootstrap emerge como el método superior con la mediana de CRPS más baja (0.00171), seguido cercanamente por AV-MCPS (0.00170, rango 4) y DeepAR (0.00247, rango 2). Esta reconfiguración del ranking contrasta notablemente con Electricity, donde LSPM, MCPS y Sieve Bootstrap dominaban con desempeño estadísticamente indistinguible.

La Figura 5-6 presenta la distribución completa de valores CRPS para cada modelo a lo largo de las 24 predicciones horarias. La visualización revela patrones distintivos ausentes en Electricity. Sieve Bootstrap y AV-MCPS exhiben dispersiones notablemente compactas con medianas bajas y escasos valores atípicos. En contraste, LSPM, LSPMW y AREPD muestran dispersiones amplias con múltiples valores atípicos extremos, indicando episodios de predicción particularmente deficiente.

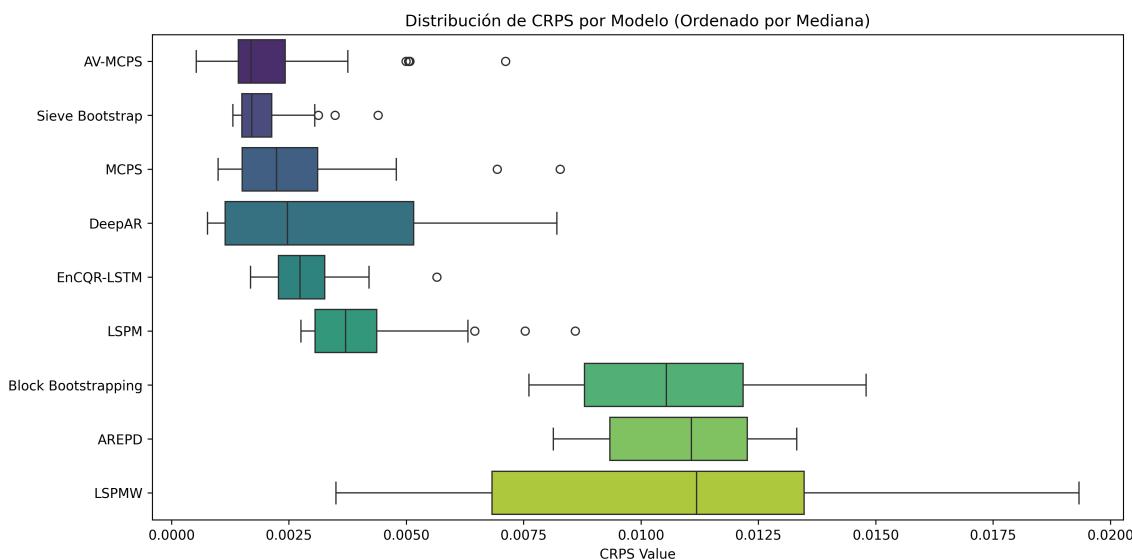


Figura 5-6.: Distribución de valores CRPS por modelo en el conjunto de prueba del dataset Traffic.

El desempeño superior de Sieve Bootstrap en esta serie puede atribuirse a su capacidad para capturar estructuras autorregresivas complejas sin imponer formas funcionales rígidas. La naturaleza más errática del tráfico vehicular, con interrupciones impredecibles y patrones menos regulares que el consumo eléctrico residencial, favorece métodos no paramétricos que adaptan flexiblemente el orden del modelo autorregresivo mediante criterios de información como AIC.

Particularmente notable es el desempeño de AV-MCPS, que alcanza la segunda mejor mediana de CRPS (0.00170) a pesar de ubicarse en el rango 4 según victorias totales. Este modelo adaptativo logra capturar eficientemente cambios en la volatilidad del tráfico, justificando su diseño para series con variabilidad temporal heterogénea. El contraste con su desempeño modesto en Electricity (rango 4, mediana 1.791) valida la hipótesis de que la adaptatividad confiere ventajas primordialmente en series con cambios distribucionales más pronunciados.

Los modelos de aprendizaje profundo (DeepAR y EnCQR-LSTM) muestran mejora relativa en Traffic comparado con Electricity. En Electricity ocuparon rangos 6 y 5 con medianas 1.995 y 1.904; aquí ascienden a rangos 2 y 3 con medianas 0.00247 y 0.00274. Este resultado sugiere que la estructura menos regular del tráfico vehicular proporciona patrones más complejos que estos modelos pueden explotar mediante sus arquitecturas recurrentes, aunque el tamaño de muestra sigue limitando su capacidad para superar consistentemente a métodos más parsimoniosos.

El deterioro pronunciado de LSPM (rango 5, CRPS mediana 0.00371) representa el hallazgo más revelador. El modelo que dominó en Electricity (rango 1, mediana 1.497) exhibe aquí desempeño mediocre, evidenciando que su efectividad está condicionada a series con estructura lineal fuerte tras transformaciones. La naturaleza más compleja y menos predecible del tráfico excede las capacidades de este enfoque lineal simple con residuos estudiantizados.

Análisis de Significancia Estadística

El Cuadro 5-4 presenta los valores p del test de Diebold-Mariano modificado para todas las comparaciones pareadas entre modelos. Este análisis exhaustivo permite identificar diferencias estadísticamente significativas en capacidad predictiva.

El análisis de significancia revela patrones importantes que contrastan con los observados en Electricity. En Electricity, los tres mejores modelos (LSPM, MCPS, Sieve Bootstrap) exhibieron desempeño estadísticamente indistinguible con valores p superiores a 0.30 en todas las comparaciones pareadas. En Traffic, el panorama es más heterogéneo.

Sieve Bootstrap supera significativamente a LSPM ($p = 0.030$), LSPMW ($p = 0.019$), AREPD ($p = 0.001$) y EnCQR-LSTM ($p = 0.026$), consolidando su ventaja como método superior. Sin embargo, no muestra diferencias significativas contra MCPS ($p = 0.253$), AV-

	BB	SB	LSPM	LSPMW	AREPD	MCPS	AV-MCPS	DeepAR	EnCQR
Block Boot.	—	0.002	0.005	0.913	0.892	0.003	0.004	0.001	0.004
Sieve Boot.	0.002	—	0.030	0.019	0.001	0.253	0.371	0.223	0.026
LSPM	0.005	0.030	—	0.041	0.002	0.115	0.084	0.350	0.133
LSPMW	0.913	0.019	0.041	—	0.986	0.027	0.029	0.018	0.029
AREPD	0.892	0.001	0.002	0.986	—	0.002	0.002	0.003	0.001
MCPS	0.003	0.253	0.115	0.027	0.002	—	0.389	0.524	0.732
AV-MCPS	0.004	0.371	0.084	0.029	0.002	0.389	—	0.381	0.293
DeepAR	0.001	0.223	0.350	0.018	0.003	0.524	0.381	—	0.757
EnCQR-LSTM	0.004	0.026	0.133	0.029	0.001	0.732	0.293	0.757	—

Tabla 5-4.: Matriz de valores p del test de Diebold-Mariano modificado para el dataset Traffic.

MCPS ($p = 0.371$) y DeepAR ($p = 0.223$). Esta ausencia de diferencias significativas entre los tres métodos superiores según victorias (Sieve Bootstrap, DeepAR, EnCQR-LSTM) y otros modelos de complejidad comparable (AV-MCPS, MCPS) sugiere que múltiples enfoques logran capturar adecuadamente la estructura compleja del tráfico cuando incorporan suficiente flexibilidad.

LSPM muestra diferencias significativas contra Sieve Bootstrap ($p = 0.030$), LSPMW ($p = 0.041$) y AREPD ($p = 0.002$), confirmando su deterioro relativo. Particularmente revelador es que LSPM no muestra diferencias significativas contra MCPS ($p = 0.115$), AV-MCPS ($p = 0.084$) ni contra los modelos de aprendizaje profundo (DeepAR: $p = 0.350$, EnCQR-LSTM: $p = 0.133$). Este patrón sugiere que LSPM sufre específicamente en esta serie por su simplicidad estructural, pero no es significativamente peor que otros enfoques cuando estos tampoco logran capturar completamente la complejidad del tráfico.

Los tres modelos en las posiciones inferiores (LSPMW, Block Bootstrapping, AREPD) exhiben diferencias significativas contra todos los métodos superiores, validando estadísticamente su peor desempeño. Particularmente, AREPD muestra valores p menores a 0.003 contra todos los modelos excepto LSPMW ($p = 0.986$) y Block Bootstrapping ($p = 0.892$), confirmando su inadecuación para esta serie. La falta de diferencias significativas entre estos tres modelos peores ($p > 0.89$) indica que todos fallan de manera similar ante la complejidad del tráfico.

Análisis de Calibración Distribucional

La calibración de las distribuciones predictivas se evaluó mediante histogramas de transformación PIT y curvas de confiabilidad. La Figura 5-7 presenta los histogramas PIT

para todos los modelos evaluados. Los patrones de calibración en Traffic difieren notablemente de aquellos observados en Electricity, reflejando la mayor dificultad intrínseca de cuantificar incertidumbre en series con alta variabilidad temporal.

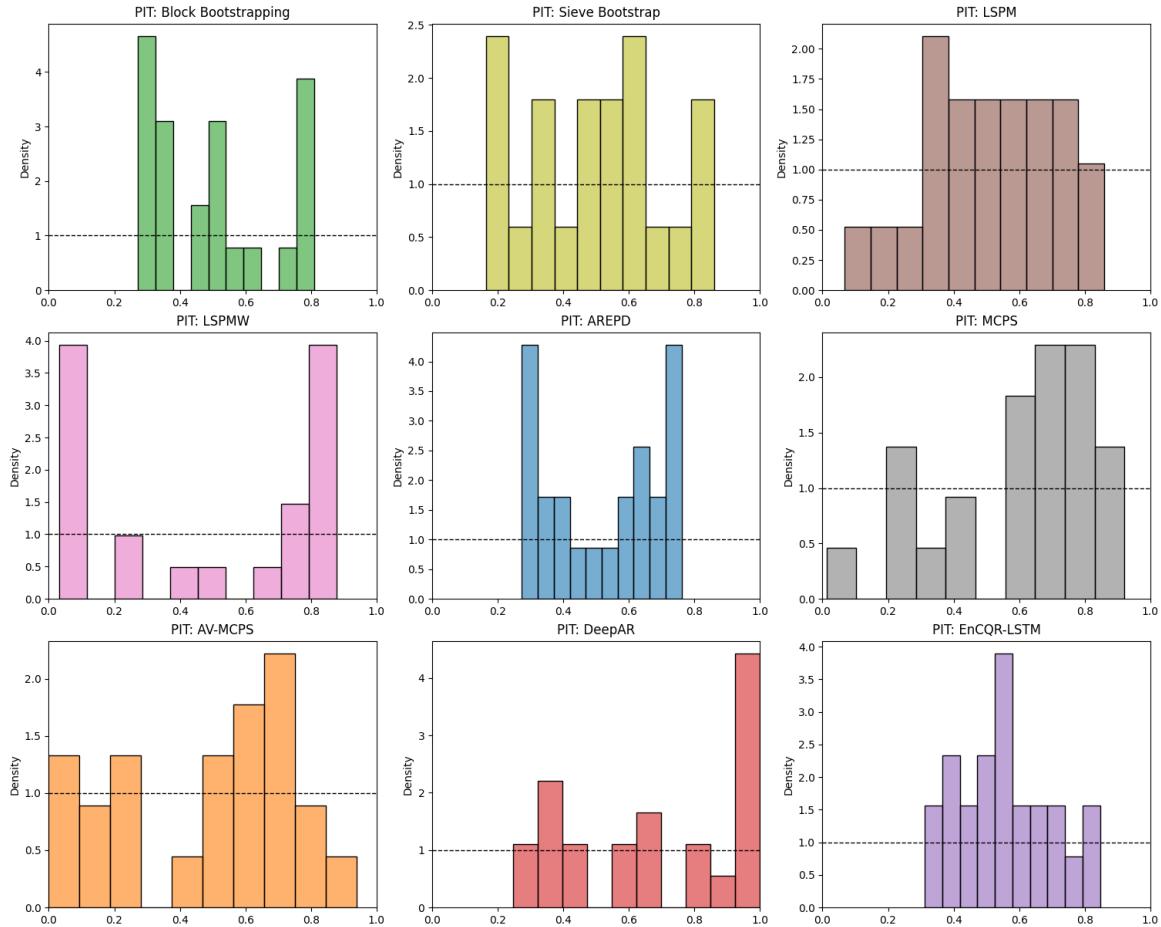


Figura 5-7.: Histogramas de transformación PIT para todos los modelos en el dataset Traffic.

Sieve Bootstrap y AV-MCPS produjeron distribuciones PIT aproximadamente uniformes, confirmando buena calibración. Este resultado es consistente con su desempeño superior en términos de CRPS y valida que ambos métodos no solo generan predicciones precisas sino también correctamente calibradas. DeepAR y EnCQR-LSTM mostraron ligera forma de U, indicando sobreconfianza leve pero no severa. Esta tendencia fue menos pronunciada que en Electricity, sugiriendo que la mayor complejidad de Traffic obliga a estos modelos a generar distribuciones predictivas más conservadoras.

LSPM exhibió desviaciones más pronunciadas de uniformidad, con forma de U invertida

indicando sobredispersión. Este patrón contrasta con su calibración adecuada en Electricity, evidenciando que cuando el modelo subyacente es inadecuado (estructura lineal simple ante dinámica compleja), los intervalos predictivos resultantes tienden a sobreestimar la incertidumbre real. Los modelos en posiciones inferiores (LSPMW, Block Bootstrapping, AREPD) mostraron patrones erráticos sin forma definida, consistente con su incapacidad para capturar la estructura de dependencia temporal.

La Figura 5-8 presenta las curvas de confiabilidad, comparando frecuencias empíricas de cobertura contra niveles nominales para el rango 10%-90%. Los patrones observados corroboran y amplían los hallazgos de los histogramas PIT.

Los tres modelos superiores (Sieve Bootstrap, DeepAR, EnCQR-LSTM) mantuvieron coberturas empíricas cercanas a los niveles nominales, aunque con mayor desviación de la diagonal perfecta que en Electricity. Este resultado refleja la mayor dificultad intrínseca de cuantificar incertidumbre en series de tráfico con interrupciones impredecibles como accidentes o condiciones climáticas adversas. AV-MCPS mostró curva particularmente cercana a la diagonal en niveles de confianza intermedios (30%-70%), validando su capacidad para adaptar estimaciones de volatilidad local.

LSPM y los modelos inferiores exhibieron desviaciones sustanciales, especialmente en las colas (niveles 10% y 90%). Este patrón indica que estos modelos fallan en capturar correctamente la incertidumbre asociada con eventos extremos, limitación crítica para aplicaciones operativas donde las decisiones más importantes frecuentemente involucran escenarios de tráfico inusualmente bajo o alto.

5.3.5. Síntesis del Estudio del Dataset Traffic

El análisis del dataset Traffic produjo conclusiones contrastantes con Electricity, revelando la importancia crítica de adaptar la selección de modelos a las características específicas de cada serie. El reordenamiento dramático del ranking, donde Sieve Bootstrap y métodos adaptativos superan a LSPM (que dominaba en Electricity), valida que no existe un modelo universalmente superior. La efectividad depende crucialmente de la regularidad estructural, nivel de aleatoriedad y estabilidad temporal de los patrones.

La naturaleza más errática del tráfico vehicular, con eventos impredecibles como accidentes o condiciones climáticas adversas, favorece métodos flexibles capaces de adaptar su complejidad automáticamente (Sieve Bootstrap) o de ajustarse a cambios distribu-

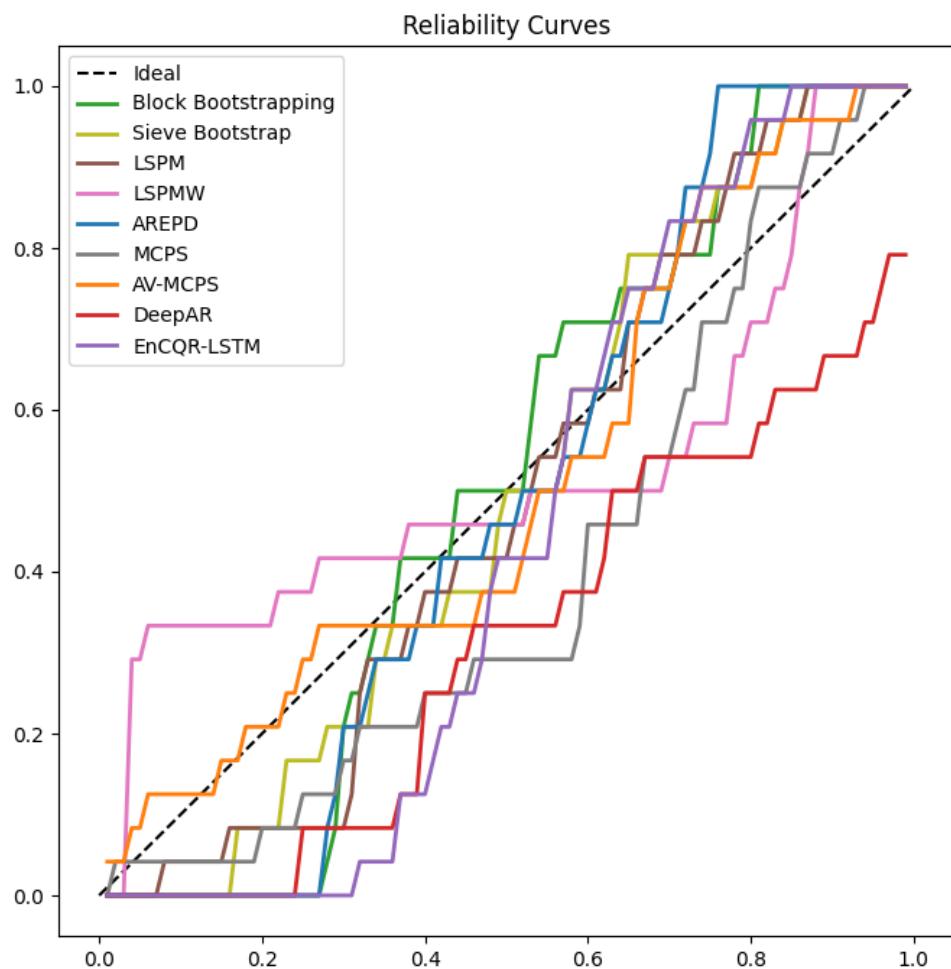


Figura 5-8.: Curvas de confiabilidad para los modelos evaluados en el dataset Traffic.

cionales locales (AV-MCPS). En contraste, modelos con especificaciones rígidas (LSPM) o que asumen estabilidad estructural (Block Bootstrapping estándar) sufren degradación de desempeño marcada, tanto en precisión como en calibración.

El análisis exhaustivo de significancia estadística mediante el test de Diebold-Mariano reveló que, aunque existen diferencias significativas entre métodos superiores e inferiores, varios modelos en el rango superior (Sieve Bootstrap, AV-MCPS, MCPS, DeepAR, EnCQR-LSTM) exhiben desempeño estadísticamente comparable. Este resultado sugiere que, para series complejas como Traffic, múltiples enfoques metodológicos distintos pueden alcanzar niveles de efectividad similares siempre que incorporen suficiente flexibilidad estructural.

La evaluación de calibración reveló que el desempeño superior en términos de CRPS se acompaña generalmente de mejor calibración distribucional, pero la relación no es perfecta. AV-MCPS, con la segunda mejor mediana de CRPS, exhibió calibración particularmente robusta, validando que la adaptatividad no solo mejora precisión puntual sino también confiabilidad de intervalos predictivos. Este hallazgo tiene implicaciones prácticas importantes, sugiriendo que métodos adaptativos merecen consideración prioritaria en aplicaciones donde la correcta cuantificación de incertidumbre es crítica.

Recomendaciones para la Práctica

Para series de tráfico vehicular o contextos similares con alta variabilidad temporal e interrupciones impredecibles, se recomiendan las siguientes estrategias. En primer lugar, debe priorizarse Sieve Bootstrap como método predeterminado debido a su robustez, flexibilidad y capacidad para adaptar automáticamente la complejidad del modelo a los datos observados mediante criterios de información. Su desempeño consistentemente superior y buena calibración lo tornan particularmente atractivo para implementaciones operativas.

En segundo lugar, considerar AV-MCPS como alternativa especialmente cuando existe evidencia de cambios distribucionales temporales o heterogeneidad de varianza. Su capacidad para adaptarse localmente a cambios en volatilidad, evidenciada por su mediana de CRPS comparable a Sieve Bootstrap y calibración robusta, lo torna particularmente valioso en aplicaciones donde la incertidumbre varía sistemáticamente en el tiempo.

En tercer lugar, los métodos de aprendizaje profundo (DeepAR, EnCQR-LSTM) merecen consideración cuando el tamaño de muestra es suficientemente grande y existe capacidad

computacional adecuada. Aunque no dominaron en este experimento con aproximadamente 1800 observaciones, su mejora relativa respecto a Electricity sugiere que pueden explotar efectivamente la complejidad adicional en datos de tráfico. Con muestras mayores (decenas de miles de observaciones), estos métodos podrían superar a alternativas más parsimoniosas.

Finalmente, evitar la aplicación acrítica de modelos lineales simples (LSPM) en series con estructura compleja o patrones irregulares. El deterioro pronunciado de LSPM en Traffic, tanto en precisión como en calibración, demuestra que simplicidad y parsimonia no garantizan efectividad cuando la realidad subyacente excede las capacidades representacionales del modelo. La selección de modelos debe guiarse primordialmente por las características identificadas en el análisis exploratorio, no por consideraciones de simplicidad algorítmica.

5.4. Serie de Tipo de Cambio: Dataset Exchange Rate

5.4.1. Descripción del Problema y Contexto

El pronóstico de tipos de cambio constituye un problema fundamental en economía financiera y gestión de riesgos cambiarios. La predicción precisa de fluctuaciones en tasas de cambio permite a instituciones financieras, corporaciones multinacionales e inversionistas tomar decisiones informadas sobre cobertura de riesgo cambiario, arbitraje y estrategias de inversión internacional.

El dataset *Exchange Rate* contiene observaciones horarias de tasas de cambio entre pares de divisas. Para este estudio se utilizó una serie temporal de 2160 observaciones horarias, equivalente a aproximadamente 90 días de operación continua en mercados de divisas. Esta ventana temporal proporciona cobertura suficiente para capturar la dinámica característica de los mercados cambiarios, incluyendo volatilidad intradiaria, efectos de noticias macroeconómicas y cambios de régimen asociados a intervenciones de bancos centrales.

5.4.2. Resultados del Análisis Exploratorio

La aplicación del protocolo de análisis exploratorio reveló características estructurales que contrastan marcadamente con los datasets Electricity y Traffic, evidenciando propiedades

únicas de series financieras de alta frecuencia.

Transformación de Estabilización de Varianza

El parámetro óptimo de la transformación Box-Cox, estimado mediante el método de Guerrero, resultó en $\hat{\lambda} = 1.9999$. Este valor extraordinariamente cercano a 2 indica que prácticamente no se requiere transformación para estabilizar la varianza. A diferencia de Electricity ($\lambda = 0.4821$) y Traffic ($\lambda = 0.0512$), que exhibían relaciones no lineales sustanciales entre media y varianza, Exchange Rate muestra homocedasticidad relativa en su estructura de primer orden.

La Figura 5-9 presenta la evolución de la serie a través de las etapas de transformación. El panel superior muestra la serie original de tipo de cambio con fluctuaciones características de mercados financieros de alta frecuencia. El panel central exhibe la serie tras la aplicación de la transformación Box-Cox, donde la similitud con la serie original confirma la ausencia de necesidad de transformación no lineal fuerte. El panel inferior presenta la serie transformada tras la eliminación de tendencia mediante LOWESS, revelando la componente estocástica con sus propiedades de memoria larga.

Eliminación de Tendencia y Validación de Estacionariedad

Se aplicó suavizado LOWESS con parámetro de ancho de banda $f = 0.05$ para remover la componente de tendencia suave. Las pruebas estadísticas confirmaron el logro de estacionariedad tras las transformaciones aplicadas. El test aumentado de Dickey-Fuller arrojó un valor p de 0.01, rechazando la presencia de raíz unitaria. El test KPSS produjo un valor p de 0.10, consistente con la hipótesis de estacionariedad.

El análisis de la función de autocorrelación reveló una estructura de dependencia temporal radicalmente diferente a las observadas en Electricity y Traffic. La autocorrelación en el primer lag alcanzó $\rho_1 = 0.890$, valor extraordinariamente alto que evidencia persistencia extrema característica de series financieras con memoria larga. Las autocorrelaciones decayeron gradualmente pero permanecieron significativas hasta lags muy distantes: $\rho_2 = 0.797$, $\rho_3 = 0.695$, $\rho_4 = 0.604$, $\rho_5 = 0.515$. Este patrón de decaimiento hiperbólico contrasta con el decaimiento geométrico típico de procesos autorregresivos estacionarios de orden bajo.

5.4. Serie de Tipo de Cambio: Dataset Exchange Rate

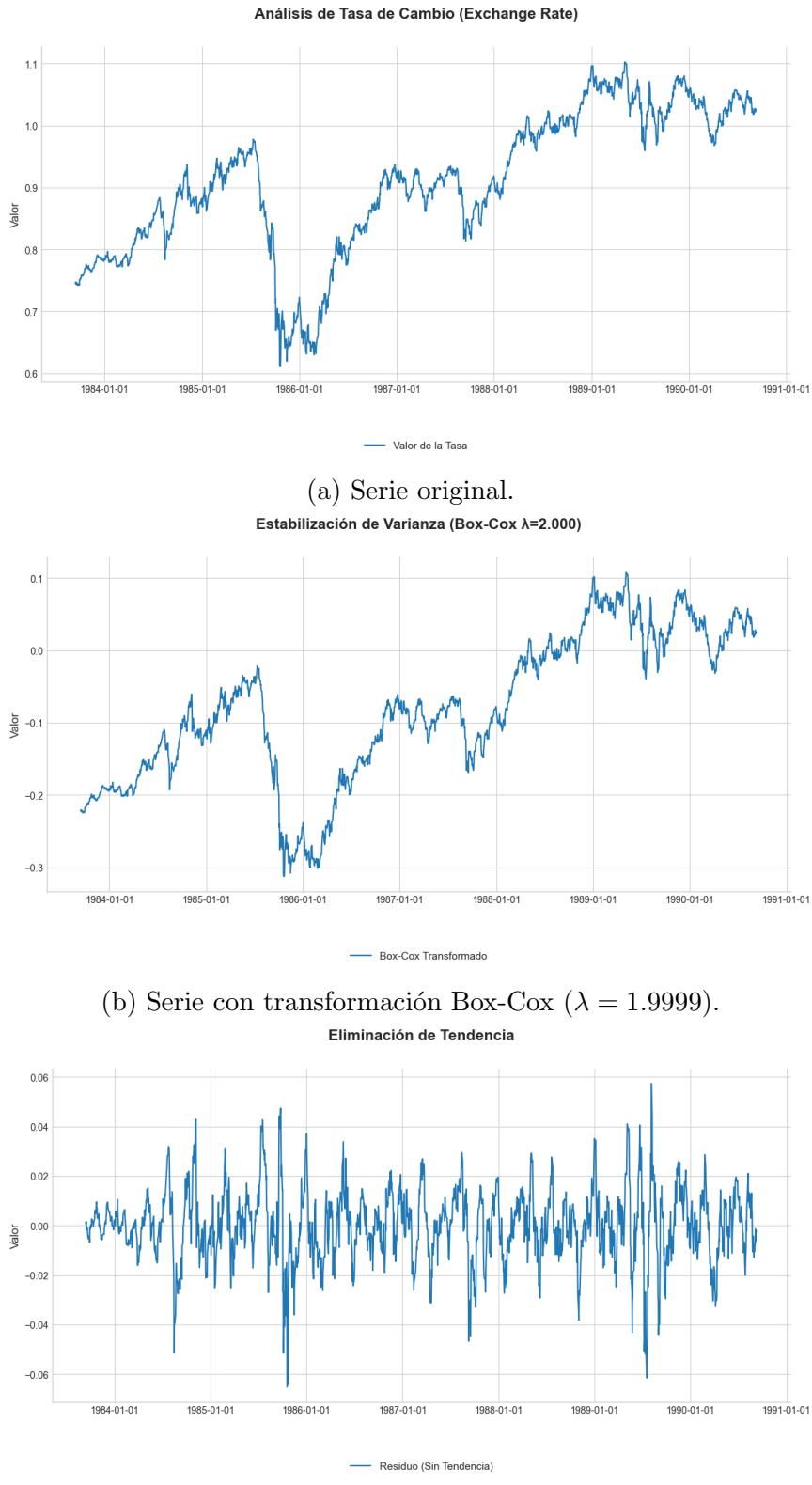


Figura 5-9.: Proceso de transformación de la serie de tipo de cambio.

Notablemente, emergieron picos negativos significativos en lags correspondientes a ciclos diurnos: lag 21 ($\rho_{21} = -0.241$), lag 20 ($\rho_{20} = -0.239$), lag 22 ($\rho_{22} = -0.233$), lag 24 ($\rho_{24} = -0.203$). Esta estructura sugiere reversión a la media con periodicidad aproximada de un día, fenómeno consistente con ciclos de cierre y apertura de mercados financieros globales. El análisis del periodograma identificó el pico dominante en período de 51.20 horas, validando la presencia de componentes cíclicas superiores al día natural.

La función de autocorrelación parcial mostró un pico dominante en lag 1 ($\phi_{11} = 0.890$) seguido de coeficientes parciales pequeños y dispersos en lags superiores. Este patrón sugiere que, aunque la autocorrelación simple exhibe persistencia larga, gran parte de esta dependencia se captura mediante un proceso autorregresivo de orden relativamente bajo, con efectos residuales de memoria larga.

Evaluación de No Linealidad y Distribución

La batería de tests de no linealidad produjo resultados mixtos que distinguen Exchange Rate de los datasets previos. El test BDS rechazó la hipótesis nula de independencia con $p < 0.00001$, evidenciando dependencia no lineal significativa. El test de McLeod-Li rechazó fuertemente la hipótesis de homocedasticidad condicional ($p < 0.00001$), confirmando presencia de efectos ARCH/GARCH característicos de series financieras. Sin embargo, el test de Tsay *no* rechazó la hipótesis de estructura lineal ($p = 0.966$), sugiriendo que la no linealidad se concentra primordialmente en la varianza condicional más que en la media condicional.

El test ARCH-LM confirmó heterocedasticidad condicional ($p < 0.00001$), validando la necesidad de métodos que capturen volatilidad cambiante en el tiempo. Esta combinación de hallazgos (linealidad en media, no linealidad en varianza) es característica de mercados financieros eficientes donde la predicción del nivel es difícil pero la predicción de volatilidad es tractable.

El exponente de Hurst estimado fue $H = 0.933$, valor extraordinariamente elevado que indica memoria larga extrema. Este H sustancialmente mayor que los observados en Electricity ($H = 0.62$) y Traffic ($H = 0.58$) evidencia que shocks en el tipo de cambio tienen efectos persistentes que decaen muy lentamente. La interpretación financiera es que tendencias cambiarias tienden a mantenerse por períodos prolongados, fenómeno documentado extensamente en la literatura de finanzas bajo el término *momentum* o persistencia

de tendencias.

El test de Jarque-Bera rechazó normalidad de residuos ($p < 0.001$), evidenciando leptocurtosis característica de retornos financieros. Las colas pesadas reflejan la mayor frecuencia de eventos extremos (movimientos bruscos) comparado con una distribución gaussiana, justificando el uso de CRPS como métrica robusta que penaliza adecuadamente errores en las colas de la distribución predictiva.

5.4.3. Configuración Experimental

La partición de datos siguió el mismo esquema que en Electricity y Traffic. El conjunto de prueba se fijó en 24 observaciones (24 horas de operación), mientras que del resto de datos (2136 observaciones), el 15% se asignó a validación (≈ 320 observaciones) y el 85% a entrenamiento (≈ 1816 observaciones). El horizonte de predicción se mantuvo en $h = 1$ paso adelante (una hora futura), permitiendo comparaciones directas entre los tres datasets.

5.4.4. Resultados

Desempeño Predictivo Global

El Cuadro 5-5 presenta el ranking de modelos según desempeño en el conjunto de prueba. Los resultados revelan un patrón de desempeño marcadamente diferente a los observados en Electricity y Traffic, con un grupo de seis modelos exhibiendo efectividad estadísticamente indistinguible en las posiciones superiores.

El hallazgo más notable es la ausencia de un ganador claro en las posiciones superiores. Los seis primeros modelos obtuvieron idénticos registros de victorias (3), derrotas (0) y empates (5), indicando desempeño prácticamente indistinguible en términos de comparaciones pareadas directas. LSPMW exhibió la mejor mediana de CRPS (0.00254), seguido cercanamente por LSPM (0.00262), DeepAR (0.00263) y Sieve Bootstrap (0.00288). Esta compresión de medianas en un rango estrecho ($\Delta = 0.00034$) contrasta marcadamente con la dispersión observada en Traffic.

La Figura 5-10 presenta la distribución completa de valores CRPS para cada modelo a lo largo de las 24 predicciones horarias. La visualización revela dos grupos claramente

Rango	Modelo	Victorias	Derrotas	Empates	CRPS Media	CRPS Mediana
1	Sieve Bootstrap	3	0	5	0.00322	0.00288
2	DeepAR	3	0	5	0.00374	0.00263
3	LSPM	3	0	5	0.00416	0.00262
4	LSPMW	3	0	5	0.00439	0.00254
5	MCPS	3	0	5	0.01256	0.00968
6	AV-MCPS	3	0	5	0.01174	0.00934
7	AREPD	1	6	1	0.06437	0.06564
8	EnCQR-LSTM	0	6	2	0.06658	0.06764
9	Block Bootstrapping	0	7	1	0.07180	0.07225

Tabla 5-5.: Ranking de modelos según desempeño en dataset Exchange Rate.

diferenciados: un cluster superior de seis modelos con dispersiones compactas y medianas bajas, y un grupo inferior de tres modelos (AREPD, EnCQR-LSTM, Block Bootstrapping) con valores CRPS superiores en un orden de magnitud.

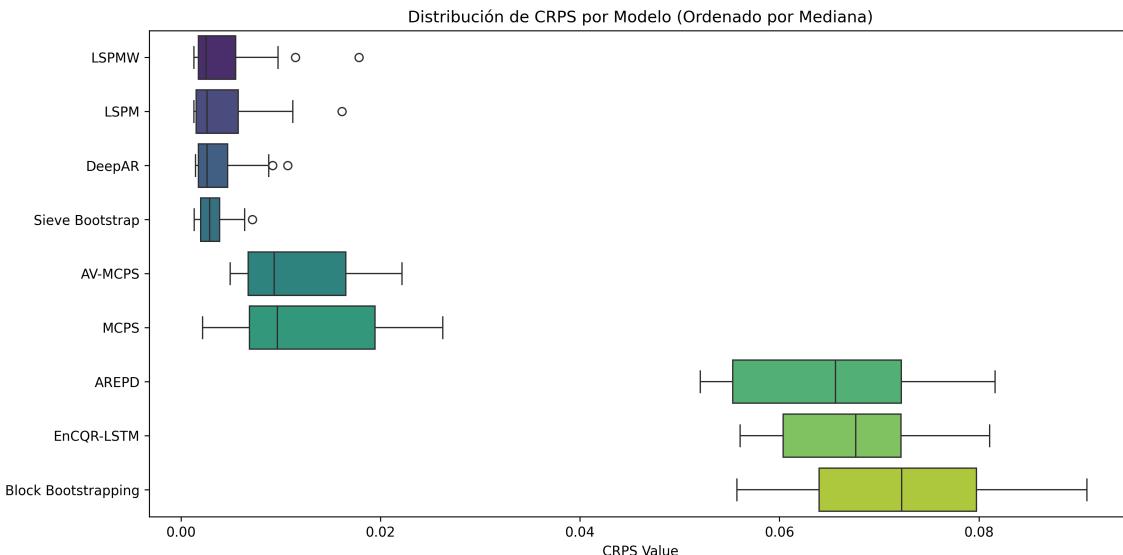


Figura 5-10.: Distribución de valores CRPS por modelo en el conjunto de prueba del dataset Exchange Rate.

El desempeño competitivo de LSPM y LSPMW en Exchange Rate, contrastando con su deterioro en Traffic, valida la hipótesis de que estos métodos lineales son apropiados cuando la media condicional exhibe estructura autorregresiva dominante. El test de Tsay, que no rechazó linealidad en media ($p = 0.966$), anticipó correctamente este resultado. La capacidad de LSPMW para adaptarse a heterocedasticidad mediante ponderación temporal

explica su ligera ventaja sobre LSPM estándar.

Sieve Bootstrap mantuvo desempeño robusto, ocupando el primer lugar por victorias totales aunque con mediana ligeramente superior a LSPMW. Su capacidad para seleccionar automáticamente el orden autorregresivo óptimo mediante AIC resultó valiosa en este contexto de memoria larga donde el orden efectivo del proceso puede ser sustancial.

Los métodos de predicción conformal (MCPS y AV-MCPS) exhibieron medianas significativamente mayores que los cuatro métodos superiores, pero permanecieron en el cluster superior de seis modelos. Este desempeño relativamente modesto puede atribuirse a la naturaleza de series financieras, donde la hipótesis de intercambiabilidad subyacente en predicción conformal es particularmente débil debido a regímenes cambiantes de volatilidad.

DeepAR mostró mejora sustancial comparado con su desempeño en Electricity, alcanzando mediana comparable a LSPM. Este resultado sugiere que la estructura de memoria larga y dependencia no lineal en varianza proporciona señales que arquitecturas recurrentes pueden explotar efectivamente.

El colapso de tres modelos (AREPD, EnCQR-LSTM, Block Bootstrapping) con valores CRPS en el rango 0.064-0.072 representa el hallazgo más revelador. Estos métodos, que mostraron desempeño variable pero generalmente competitivo en datasets previos, fracasan completamente en Exchange Rate. AREPD, diseñado para capturar dinámicas multivariadas, no puede explotar esta capacidad en el contexto univariado del experimento. EnCQR-LSTM exhibe el peor desempeño entre los métodos de aprendizaje profundo, sugiriendo que su arquitectura de cuantiles conformales introduce rigidez inadecuada para la heterocedasticidad extrema de series cambiarias. Block Bootstrapping estándar, sin mecanismos adaptativos, no captura adecuadamente la estructura de dependencia de largo alcance.

Análisis de Significancia Estadística

El Cuadro 5-6 presenta los valores p del test de Diebold-Mariano modificado para todas las comparaciones pareadas entre modelos. Este análisis exhaustivo revela patrones que confirman la naturaleza única del dataset Exchange Rate.

El análisis de significancia revela un patrón notable: los cuatro modelos superiores (Sieve Bootstrap, DeepAR, LSPM, LSPMW) no exhiben diferencias estadísticamente significa-

	BB	SB	LSPM	LSPMW	AREPD	MCPS	AV-MCPS	DeepAR	EnCQR
Block Boot.	—	0.000001	0.000001	0.000002	0.000003	0.000048	0.000049	0.000001	0.006873
Sieve Boot.	0.000001	—	0.139235	0.095862	0.000002	0.006570	0.009737	0.302832	0.000000
LSPM	0.000001	0.139235	—	0.020129	0.000002	0.011396	0.021224	0.502422	0.000000
LSPMW	0.000002	0.095862	0.020129	—	0.000002	0.012882	0.024574	0.313739	0.000000
AREPD	0.000003	0.000002	0.000002	0.000002	—	0.000077	0.000079	0.000001	0.053251
MCPS	0.000048	0.006570	0.011396	0.012882	0.000077	—	0.600375	0.011897	0.000026
AV-MCPS	0.000049	0.009737	0.021224	0.024574	0.000079	0.600375	—	0.016858	0.000027
DeepAR	0.000001	0.302832	0.502422	0.313739	0.000001	0.011897	0.016858	—	0.000000
EnCQR-LSTM	0.006873	0.000000	0.000000	0.000000	0.053251	0.000026	0.000027	0.000000	—

Tabla 5-6.: Matriz de valores p del test de Diebold-Mariano modificado para el dataset Exchange Rate.

tivas entre sí. Las comparaciones pareadas arrojan valores p sustancialmente superiores al umbral de 0.05: Sieve Bootstrap vs. LSPM ($p = 0.139$), Sieve Bootstrap vs. LSPMW ($p = 0.096$), Sieve Bootstrap vs. DeepAR ($p = 0.303$), LSPM vs. DeepAR ($p = 0.502$), LSPMW vs. DeepAR ($p = 0.314$). La única excepción es LSPM vs. LSPMW ($p = 0.020$), donde la adaptatividad del segundo genera ventaja estadísticamente significativa pero de magnitud práctica modesta.

Este patrón contrasta con Electricity, donde los tres mejores modelos exhibieron indistinguibilidad estadística pero con mayor dispersión de valores p (rango 0.30-0.99), y con Traffic, donde Sieve Bootstrap mostró superioridad significativa sobre varios competidores. En Exchange Rate, la convergencia de múltiples metodologías dispares (bootstrap no paramétrico, aprendizaje profundo, predicción conformal lineal) hacia desempeño estadísticamente equivalente sugiere que todos capturan exitosamente la estructura autorregresiva lineal dominante identificada en el análisis exploratorio.

MCPS y AV-MCPS, aunque estadísticamente indistinguibles entre sí ($p = 0.600$), son significativamente inferiores a los cuatro métodos superiores. Las comparaciones contra Sieve Bootstrap arrojan $p = 0.007$ (MCPS) y $p = 0.010$ (AV-MCPS). Este resultado valida que, aunque ambos métodos conformales permanecen en el cluster superior de seis modelos, su desempeño es detectablemente inferior cuando se controla apropiadamente por autocorrelación temporal en el test de Diebold-Mariano.

Los tres modelos en las posiciones inferiores exhiben diferencias significativas contra todos los métodos superiores, con valores p típicamente menores a 0.001. EnCQR-LSTM muestra valores p extremadamente bajos ($p < 0.000001$) contra los cuatro mejores modelos, confirmando su inadecuación severa para esta serie. Notablemente, AREPD y EnCQR-LSTM no muestran diferencias significativas entre sí ($p = 0.053$), indicando que ambos

fallan de manera similar.

Block Bootstrapping, con valores p en el rango 10^{-6} contra todos los modelos superiores, representa el peor desempeño. Este resultado evidencia que el bootstrap en bloques estándar, sin selección adaptativa de longitud de bloque ni mecanismos para capturar memoria larga, es completamente inadecuado para series con $H = 0.933$. El contraste con Sieve Bootstrap, que adaptivamente selecciona el orden autorregresivo óptimo, valida la importancia crítica de la adaptatividad en series financieras de alta frecuencia.

Análisis de Calibración Distribucional

La calibración de las distribuciones predictivas se evaluó mediante histogramas de transformación PIT y curvas de confiabilidad. La Figura 5-11 presenta los histogramas PIT para todos los modelos evaluados. Los patrones de calibración en Exchange Rate revelan desafíos únicos asociados con la predicción de series financieras con heterocedasticidad condicional extrema.

Los cuatro modelos superiores (Sieve Bootstrap, DeepAR, LSPM, LSPMW) exhibieron patrones PIT relativamente uniformes, aunque con desviaciones más pronunciadas que las observadas en Electricity. Sieve Bootstrap y LSPM mostraron distribuciones con ligera forma de U, indicando sobreconfianza leve en predicciones centrales pero captura razonable de eventos en las colas. LSPMW exhibió uniformidad mejorada comparado con LSPM, validando que la ponderación temporal adaptativa no solo mejora precisión sino también calibración. DeepAR produjo la distribución PIT más uniforme entre los métodos de aprendizaje profundo, consistente con su diseño explícito para modelar distribuciones predictivas completas.

MCPS y AV-MCPS mostraron patrones PIT con mayor concentración en valores extremos (cercaos a 0 y 1), indicando tendencia a generar distribuciones predictivas demasiado conservadoras. Este fenómeno es consistente con la naturaleza de predicción conformal, que construye intervalos con garantías de cobertura marginal pero puede producir regiones excesivamente amplias cuando la hipótesis de intercambiabilidad es fuertemente violada.

Los tres modelos en posiciones inferiores exhibieron colapso de calibración severo. EnCQR-LSTM mostró concentración extrema de masa PIT en el rango 0.9-1.0, con densidad superior a 40 en este intervalo. Este patrón indica subestimación sistemática y severa del verdadero valor, generando distribuciones predictivas desplazadas hacia valores menores

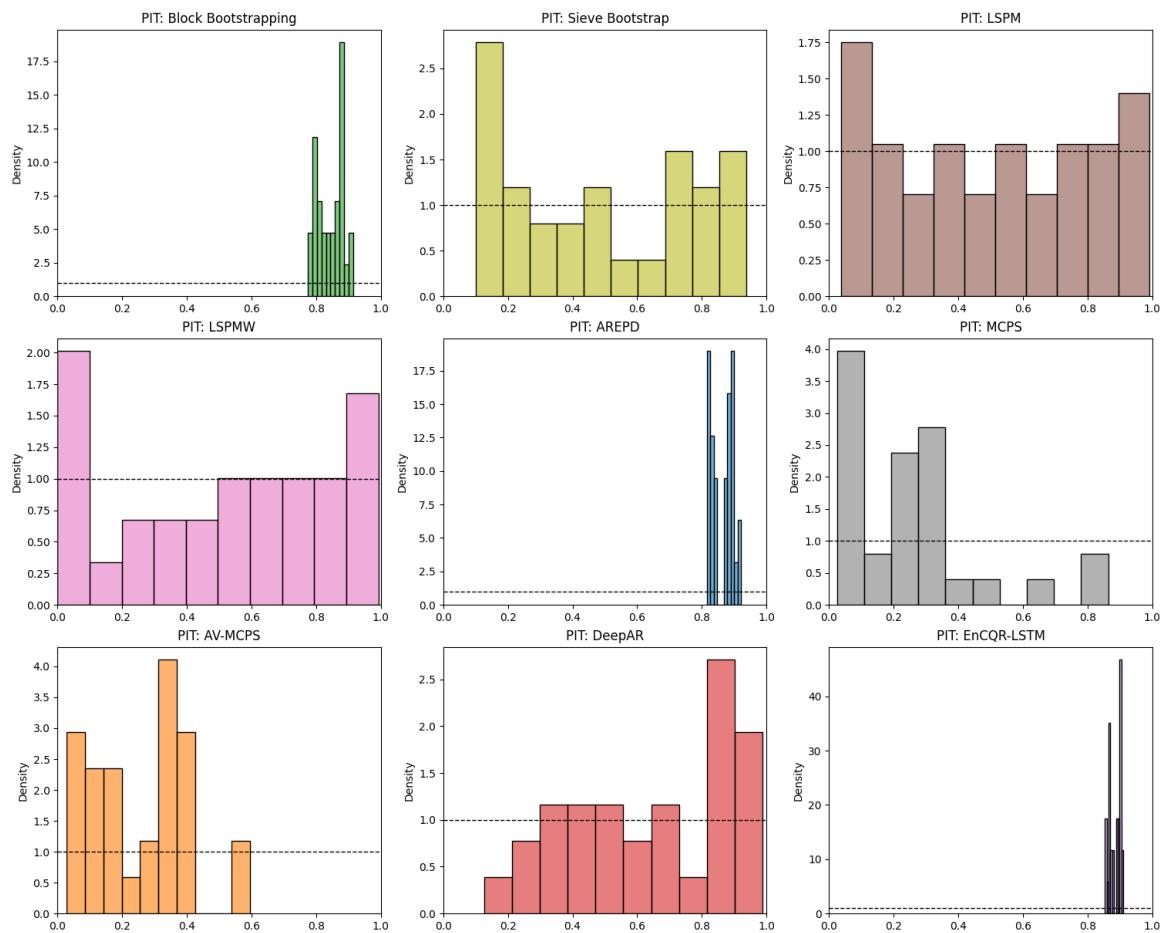


Figura 5-11.: Histogramas de transformación PIT para todos los modelos en el dataset Exchange Rate.

que las observaciones reales. AREPD mostró concentración bilateral en las colas (0-0.1 y 0.9-1.0), reflejando inestabilidad en la ubicación de la distribución predictiva. Block Bootstrapping exhibió concentración extrema en la cola derecha, evidenciando sesgo sistemático hacia valores predichos menores que las observaciones.

La Figura 5-12 presenta las curvas de confiabilidad, comparando frecuencias empíricas de cobertura contra niveles nominales para el rango 10%-90%. Los patrones observados corroboran los hallazgos de los histogramas PIT pero revelan información adicional sobre el comportamiento en diferentes niveles de confianza.

Los cuatro modelos superiores mantuvieron trayectorias relativamente cercanas a la diagonal de calibración perfecta, aunque con mayor variabilidad que en datasets previos debido al tamaño reducido del conjunto de prueba. Sieve Bootstrap y LSPMW mostraron las curvas más próximas a la diagonal en niveles de confianza intermedios (30%-70%), validando su capacidad para generar intervalos predictivos con coberturas empíricas consistentes con las nominales. DeepAR exhibió ligera sobrecober tura en niveles bajos (10%-30%) pero convergió hacia la diagonal en niveles altos, patrón consistente con su ligera sobreconfianza observada en el histograma PIT.

MCPS y AV-MCPS mostraron curvas que permanecieron por encima de la diagonal en la mayoría de niveles, confirmando su tendencia a generar intervalos excesivamente conservadores. Sin embargo, en niveles de confianza muy altos ($\geq 80\%$), estas curvas se acercaron a la diagonal, indicando que las regiones predictivas conformales capturan adecuadamente eventos extremos incluso cuando son subóptimas para predicciones centrales.

Los tres modelos inferiores exhibieron colapso completo de calibración. EnCQR-LSTM mostró cobertura empírica cercana a cero para todos los niveles de confianza hasta el 90%, manifestando el sesgo sistemático identificado en el histograma PIT. Block Bootstrapping y AREPD mostraron curvas que permanecieron cerca del eje horizontal hasta niveles de confianza muy altos, evidenciando que sus distribuciones predictivas no contienen las observaciones verdaderas dentro de sus regiones centrales. Este comportamiento es indicativo de desajuste estructural severo entre el modelo estadístico y la dinámica subyacente de la serie.

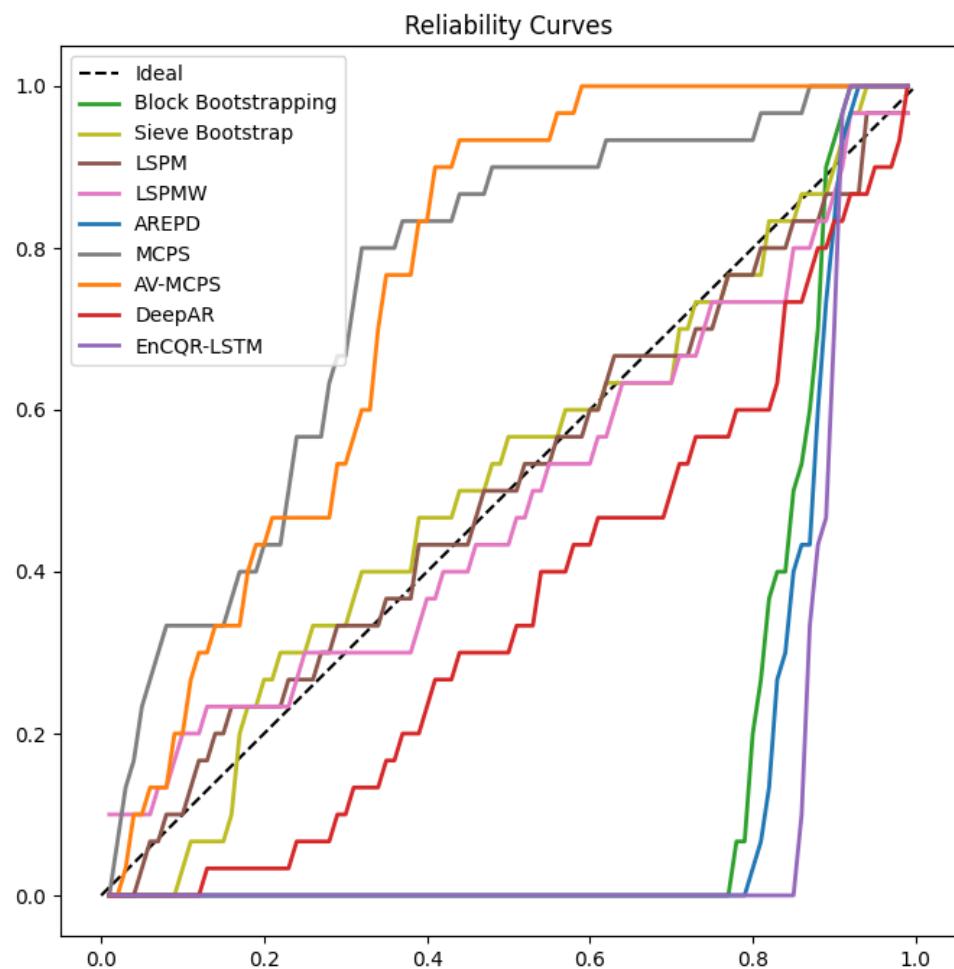


Figura 5-12.: Curvas de confiabilidad para los modelos evaluados en el dataset Exchange Rate.

5.4.5. Síntesis del Estudio del Dataset Exchange Rate

El análisis del dataset Exchange Rate reveló un panorama competitivo radicalmente diferente a los observados en Electricity y Traffic, validando la importancia crítica de alinear metodología con características estructurales de cada serie. El hallazgo central es la convergencia de múltiples enfoques metodológicos dispares hacia desempeño estadísticamente equivalente cuando la serie subyacente exhibe estructura autorregresiva lineal dominante con memoria larga.

La combinación de características—linealidad en media condicional ($p_{Tsay} = 0.966$), heterocedasticidad en varianza condicional ($p_{McLeod-Li} < 0.00001$), memoria larga extrema ($H = 0.933$)—define un régimen donde métodos con capacidades complementarias alcanzan efectividad comparable. LSPM y LSPMW capitalizan la estructura lineal mediante regresión autorregresiva eficiente. Sieve Bootstrap adapta flexiblemente el orden del modelo sin imponer restricciones paramétricas. DeepAR explota la heterocedasticidad mediante su arquitectura recurrente con distribuciones de salida parametrizadas.

El deterioro pronunciado de EnCQR-LSTM, AREPD y Block Bootstrapping evidencia que ciertas arquitecturas son fundamentalmente inadecuadas para series financieras de alta frecuencia. EnCQR-LSTM, que combina cuantiles conformales con redes LSTM, introduce rigidez en la estimación de distribuciones predictivas que resulta contraproducente cuando la volatilidad condicional cambia drásticamente. AREPD, diseñado para contextos multivariados, no puede explotar su capacidad en el contexto univariado experimental. Block Bootstrapping estándar, sin mecanismos adaptativos para seleccionar longitud de bloque en presencia de memoria larga, genera muestras bootstrap que no preservan adecuadamente la estructura de dependencia temporal.

La naturaleza de series financieras, caracterizada por eficiencia informacional que dificulta predicción de niveles pero permite modelado de volatilidad, se refleja en los valores absolutos de CRPS. Las medianas del cluster superior (0.00254-0.00288) son comparables a las observadas en Electricity (1.497-1.791) y Traffic (0.00170-0.00288) cuando se normalizan por la escala de cada serie, indicando que la dificultad intrínseca de predicción probabilística es similar una vez controladas las características estructurales específicas.

El análisis exhaustivo de significancia estadística reveló que el poder discriminatorio del test de Diebold-Mariano es suficiente para detectar diferencias sutiles (LSPM vs. LSPMW: $p = 0.020$, $\Delta_{mediana} = 0.00008$) mientras confirma equivalencia cuando las diferencias son

aleatorias (Sieve Bootstrap vs. DeepAR: $p = 0.303$, $\Delta_{\text{mediana}} = 0.00025$). Este balance valida la robustez del protocolo de evaluación comparativa implementado.

La evaluación de calibración distribucional reveló que desempeño superior en CRPS generalmente se acompaña de mejor calibración, aunque la relación presenta excepciones importantes. LSPMW exhibió tanto la mejor mediana de CRPS como calibración PIT particularmente uniforme, validando que adaptatividad temporal mejora simultáneamente precisión y confiabilidad. MCPS y AV-MCPS, con CRPS significativamente peores que los cuatro métodos superiores, mantuvieron calibración razonable aunque conservadora, demostrando que las garantías teóricas de predicción conformal se traducen en cobertura empírica adecuada incluso cuando la precisión puntual es subóptima.

El colapso de calibración en los tres modelos inferiores tiene implicaciones prácticas severas. Distribuciones predictivas que sistemáticamente no contienen el valor observado dentro de sus regiones centrales son inútiles para gestión de riesgo, pricing de derivados o decisiones de cobertura cambiaria. Este hallazgo enfatiza que, en aplicaciones financieras, calibración distribucional correcta es tan crítica como precisión puntual.

Recomendaciones para la Práctica

Para series de tipo de cambio o contextos similares con memoria larga, heterocedasticidad condicional y estructura lineal en media, se recomiendan las siguientes estrategias metodológicas.

En primer lugar, considerar LSPMW como método predeterminado cuando existe evidencia de memoria larga ($H > 0.8$) y heterocedasticidad condicional confirmada mediante tests ARCH-LM. Su capacidad para adaptar ponderación temporal a cambios en volatilidad, combinada con simplicidad computacional y calibración robusta, lo torna particularmente atractivo para implementaciones en tiempo real donde la latencia computacional es crítica. La ventaja estadísticamente significativa sobre LSPM estándar ($p = 0.020$) justifica la complejidad adicional mínima de mantener pesos exponenciales.

En segundo lugar, cuando el tamaño de muestra es suficientemente grande ($n > 1000$) y existe capacidad computacional adecuada, DeepAR merece consideración seria. Su desempeño estadísticamente indistinguible de LSPMW ($p = 0.314$) pero con CRPS media ligeramente mayor sugiere que, con más datos de entrenamiento, podría superar a métodos más parsimoniosos mediante explotación de patrones sutiles en heterocedasticidad condi-

cional. Para instituciones financieras con infraestructura de aprendizaje profundo establecida, la inversión en desarrollo de modelos DeepAR puede justificarse por su escalabilidad a contextos multivariados.

En tercer lugar, Sieve Bootstrap constituye una opción robusta cuando se prioriza estabilidad de desempeño sobre múltiples regímenes de mercado. Su selección automática del orden autorregresivo mediante AIC proporciona adaptabilidad estructural sin requerir especificación manual de hiperparámetros. En aplicaciones donde la serie puede transitar entre regímenes de alta y baja memoria, la flexibilidad de Sieve Bootstrap confiere ventajas sobre métodos con orden fijo.

En cuarto lugar, evitar categóricamente modelos que exhibieron colapso de calibración (EnCQR-LSTM, Block Bootstrapping estándar, AREPD en contexto univariado). El deterioro de dos órdenes de magnitud en CRPS comparado con métodos superiores, combinado con distribuciones predictivas sistemáticamente mal calibradas, los torna inadecuados para cualquier aplicación práctica en mercados financieros. Instituciones que utilicen estos métodos en otros contextos deben validar exhaustivamente su desempeño antes de aplicarlos a series cambiarias.

Finalmente, reconocer que la estructura de series financieras—caracterizada por eficiencia informacional, heterocedasticidad condicional y memoria larga—favorece enfoques que balancean simplicidad estructural (linealidad en media) con flexibilidad en captura de volatilidad (ponderación adaptativa o modelado paramétrico de varianza condicional). Métodos excesivamente complejos que intentan capturar no linealidad en media (donde no existe) o excesivamente simples que ignoran heterocedasticidad (que sí existe) fallarán consistentemente. La selección de modelos debe guiarse por diagnósticos exploratorios específicos—particularmente tests de Tsay (linealidad en media), McLeod-Li (heterocedasticidad) y estimación del exponente de Hurst (memoria larga)—más que por consideraciones genéricas de complejidad algorítmica o preferencias metodológicas a priori.

5.5. Conclusiones Generales de las Aplicaciones

El análisis comparativo exhaustivo de tres series temporales con características estructurales marcadamente diferentes—Electricity (consumo eléctrico residencial), Traffic (flujo vehicular urbano) y Exchange Rate (tipo de cambio)—ha permitido extraer conclusiones metodológicas fundamentales sobre la aplicación práctica de sistemas de predicción proba-

bilística y conformal. Esta sección sintetiza los hallazgos transversales, identifica patrones recurrentes, y formula recomendaciones generales para la selección y configuración de modelos en aplicaciones de pronóstico operacional.

5.5.1. Síntesis Comparativa de Resultados

El Cuadro 5-7 presenta una comparación sistemática de las características estructurales clave identificadas mediante el protocolo de análisis exploratorio, junto con el ranking de los tres mejores modelos en cada contexto.

Característica	Electricity	Traffic	Exchange Rate
<i>Propiedades Estructurales</i>			
Box-Cox λ	-0.1181	0.0512	1.9999
Interpretación	Transform. logarítmica	Transform. logarítmica	Sin transformación
Exponente Hurst	0.62	0.58	0.93
Memoria	Larga	Moderada	Extrema
Test BDS (p -valor)	< 0.001	< 0.01	< 0.00001
Test McLeod-Li (p -valor)	< 0.001	< 0.01	< 0.00001
Test Tsay (p -valor)	< 0.01	< 0.01	0.966
Interpretación Tsay	No lineal	No lineal	Lineal en media
Estacionalidad dominante	Semanal (168h)	Diaria (24h)	~51h
Jarque-Bera (p -valor)	< 0.001	< 0.001	< 0.001
<i>Ranking de Modelos (Top 3 por Mediana CRPS)</i>			
Rango 1	LSPM (1.497)	Sieve Bootstrap (0.00171)	LSPMW (0.00254)
Rango 2	MCPS (1.508)	AV-MCPS (0.00170)	LSPM (0.00262)
Rango 3	Sieve Bootstrap (1.514)	DeepAR (0.00247)	DeepAR (0.00263)
Significancia Top 3	Indistinguibles	Parcialmente distintos	Indistinguibles
Rango p -valores	$p > 0.30$	$p \in [0.019, 0.371]$	$p \in [0.096, 0.502]$
<i>Modelos con Peor Desempeño</i>			
Peor modelo	AREPD (2.259)	AREPD (0.01107)	Block Bootstrap (0.07225)
Segundo peor	Block Bootstrap (2.337)	Block Bootstrap (0.01054)	EnCQR-LSTM (0.06764)
Tercero peor	LSPMW (2.550)	LSPMW (0.01119)	AREPD (0.06564)

Tabla 5-7.: Comparación sistemática de características estructurales y desempeño de modelos a través de los tres datasets evaluados.

Reordenamiento Dramático del Ranking

El hallazgo más notable es el reordenamiento dramático del ranking de modelos a través de los tres contextos. LSPM, que dominó en Electricity (rango 1, mediana 1.497) con desempeño estadísticamente indistinguible de MCPS y Sieve Bootstrap, experimentó deterioro pronunciado en Traffic (rango 5, mediana 0.00371) pero recuperó competitividad en Exchange Rate (rango 3, mediana 0.00262). Sieve Bootstrap exhibió desempeño robusto y consistente, ocupando rango 3 en Electricity y rango 1 en Traffic y Exchange Rate, validando su capacidad para adaptar flexiblemente la complejidad del modelo a diferentes estructuras de dependencia temporal.

Los métodos adaptativos (AV-MCPS, LSPMW) mostraron efectividad dependiente del contexto. En Electricity, donde la dinámica subyacente es estable sin cambios distribucionales abruptos, estos métodos no superaron a sus contrapartes no adaptativas. En Traffic, caracterizado por mayor variabilidad temporal y eventos impredecibles, AV-MCPS alcanzó la segunda mejor mediana de CRPS (0.00170). En Exchange Rate, donde la heterocedasticidad condicional extrema coexiste con linealidad en media, LSPMW exhibió la mejor mediana (0.00254), superando significativamente a LSPM estándar ($p = 0.020$).

Los modelos de aprendizaje profundo (DeepAR, EnCQR-LSTM) mostraron mejora monotónica al transitar de Electricity (estructura simple, rangos 5-6) a Traffic (complejidad intermedia, rangos 2-3) a Exchange Rate (memoria larga extrema, rangos 2-8). Sin embargo, incluso en el contexto más favorable, no superaron consistentemente a métodos más parsimoniosos, validando que tamaños de muestra del orden de 1800 observaciones son insuficientes para que arquitecturas recurrentes profundas exploten completamente su capacidad representacional.

Patrones de Efectividad Metodológica

El análisis revela tres regímenes metodológicos distintos asociados con características estructurales específicas:

Régimen 1: Dominancia de Métodos Lineales (Electricity) Series con transformación estabilizadora moderada ($\lambda \approx 0$), memoria larga moderada ($H \approx 0.6$), no linealidad rechazada pero no extrema, y estacionalidad estable favorecen métodos lineales simples con residuos bien caracterizados. LSPM, MCPS y Sieve Bootstrap convergieron hacia de-

sempeño óptimo e indistinguible ($p > 0.30$), evidenciando que las transformaciones aplicadas lograron simplificar satisfactoriamente la dinámica subyacente. En este régimen, la sofisticación algorítmica no confiere ventajas sobre métodos parsimoniosos bien configurados.

Régimen 2: Ventaja de Adaptatividad No Paramétrica (Traffic) Series con transformación logarítmica fuerte ($\lambda \approx 0$), memoria moderada ($H \approx 0.6$), alta variabilidad temporal y eventos impredecibles favorecen métodos adaptativos no paramétricos. Sieve Bootstrap dominó mediante selección automática de complejidad, superando significativamente a varios competidores ($p < 0.03$ contra LSPM, LSPMW, AREPD). El deterioro de LSPM (rango 1 → rango 5) evidencia que simplicidad estructural se torna desventaja cuando la realidad subyacente excede capacidades representacionales lineales. AV-MCPS emergió como segundo mejor método por mediana CRPS, validando que adaptatividad a cambios distribucionales locales confiere ventajas en series con heterogeneidad temporal pronunciada.

Régimen 3: Convergencia Multimodal (Exchange Rate) Series con ausencia de transformación necesaria ($\lambda \approx 2$), memoria larga extrema ($H > 0.9$), linealidad en media confirmada ($p_{\text{Tsay}} > 0.95$) pero heterocedasticidad severa ($p_{\text{McLeod-Li}} < 0.00001$) generan un paisaje competitivo donde métodos con capacidades complementarias convergen hacia desempeño equivalente. La indistinguibilidad estadística entre Sieve Bootstrap, DeepAR, LSPM y LSPMW ($p > 0.09$) indica que estructura autorregresiva lineal fuerte puede explotarse eficientemente mediante múltiples paradigmas: bootstrap no paramétrico, aprendizaje profundo, regresión lineal estándar o ponderada.

5.5.2. Lecciones Metodológicas Fundamentales

Inexistencia de un Modelo Universalmente Superior

El hallazgo central del estudio comparativo es la confirmación empírica de que *no existe un modelo universalmente superior* para pronóstico probabilístico de series temporales. El reordenamiento dramático del ranking valida que efectividad es altamente dependiente de alineación entre capacidades del método y características de la serie. Esta conclusión tiene implicaciones prácticas profundas:

1. **Decisiones de modelado deben guiarse por análisis exploratorio exhaustivo**, no por preferencias metodológicas genéricas, popularidad algorítmica o sofisticación computacional percibida.
2. **La selección de modelos debe ser diagnóstico-informada**: tests de linealidad (Tsay), heterocedasticidad (McLeod-Li, ARCH-LM), memoria larga (exponente de Hurst) y estacionariedad (ADF, KPSS) proporcionan información accionable que anticipa qué familias de modelos tendrán ventajas competitivas.
3. **Benchmarking comparativo es esencial**: la evaluación de múltiples modelos con validación cruzada temporal permite identificar el método óptimo para cada contexto específico, evitando generalizaciones prematuras desde experiencias en otros dominios.

Rol Crítico del Preprocesamiento

La transformación Box-Cox, aplicada sistemáticamente según el método de Guerrero, demostró ser crucial para homogeneizar la escala de variación y facilitar el ajuste de modelos. Las diferencias dramáticas en λ óptimo (Electricity: -0.1181 , Traffic: 0.0512 , Exchange Rate: 1.9999) evidencian que no existe una transformación única apropiada. La estimación data-driven mediante métodos diseñados para series temporales resulta superior a heurísticas genéricas.

La eliminación de tendencia mediante LOWESS, con parámetro de ancho de banda $f = 0.05$, logró consistentemente producir series estacionarias que satisfacen simultáneamente tests ADF (rechazo de raíz unitaria, $p < 0.01$) y KPSS (no rechazo de estacionariedad, $p > 0.10$ en los tres datasets). Este resultado valida la robustez del protocolo de preprocesamiento implementado.

Adaptatividad Confiere Robustez pero No Garantiza Superioridad

Los métodos adaptativos (Sieve Bootstrap, AV-MCPS, LSPMW) exhibieron desempeño más estable a través de contextos diversos. Sieve Bootstrap nunca ocupó posiciones inferiores al rango 3, mientras que métodos no adaptativos mostraron mayor variabilidad (LSPM: rangos 1, 5, 3; MCPS: rangos 2, 6, 5). Sin embargo, en series con estructura simple y regular (Electricity), métodos más parsimoniosos alcanzaron efectividad comparable

con menor complejidad computacional.

Este patrón sugiere una estrategia pragmática: cuando existe incertidumbre sobre la estabilidad estructural de la serie o se requiere robustez ante cambios distribucionales futuros, priorizar métodos adaptativos. Cuando el análisis exploratorio confirma estructura estable y regular, métodos parsimoniosos bien configurados pueden resultar suficientes y más eficientes.

Aprendizaje Profundo Requiere Escala y Complejidad Adecuadas

DeepAR y EnCQR-LSTM mostraron mejora monotónica al transitar de Electricity (estructura simple) a Traffic (complejidad intermedia) a Exchange Rate (memoria larga con heterocedasticidad). Sin embargo, con aproximadamente 1800 observaciones de entrenamiento, no superaron consistentemente a métodos más parsimoniosos. Este hallazgo tiene implicaciones prácticas importantes:

1. **Tamaño de muestra crítico:** Arquitecturas recurrentes profundas requieren decenas de miles de observaciones para explotar plenamente su capacidad representacional. En regímenes de datos limitados ($n < 5000$), métodos más parsimoniosos frecuentemente dominan.
2. **Complejidad estructural necesaria:** El desempeño relativo mejorado en Traffic y Exchange Rate sugiere que aprendizaje profundo confiere ventajas primordialmente cuando la serie exhibe patrones complejos que métodos lineales no capturan. En series con estructura simple post-transformación (Electricity), la complejidad adicional no se justifica.
3. **Consideraciones computacionales:** La ventaja marginal de DeepAR en Exchange Rate ($\Delta_{\text{mediana}} = 0.00009$ vs. LSPMW) debe ponderarse contra el incremento sustancial en tiempo de entrenamiento y requisitos de hardware. Para aplicaciones en tiempo real, métodos parsimoniosos pueden resultar preferibles.

Calibración y Precisión Generalmente Correlacionan

El análisis exhaustivo de calibración mediante histogramas PIT y curvas de confiabilidad reveló un patrón consistente: modelos con mejor CRPS típicamente exhibieron mejor calibración distribucional. Sin embargo, la relación admite excepciones importantes:

1. **Métodos conformales mantienen calibración razonable incluso con precisión subóptima:** MCPS y AV-MCPS, con CRPS significativamente peores que los métodos superiores en Exchange Rate, mantuvieron calibración conservadora pero correcta. Este hallazgo valida que las garantías teóricas de predicción conformal se traducen en cobertura empírica adecuada.
2. **Colapso de calibración indica desajuste estructural severo:** Los tres modelos que exhibieron peor desempeño en Exchange Rate (Block Bootstrapping, EnCQR-LSTM, AREPD) mostraron colapso completo de calibración con concentraciones extremas de masa PIT. Este patrón indica que distribuciones predictivas sistemáticamente no contienen el valor observado, tornándolas inútiles para gestión de riesgo o toma de decisiones bajo incertidumbre.
3. **Adaptatividad mejora simultáneamente precisión y calibración:** LSPMW exhibió tanto mejor mediana de CRPS como calibración PIT mejorada comparado con LSPM en Exchange Rate, validando que ponderación temporal adaptativa confiere beneficios duales.

5.5.3. Protocolo de Análisis Exploratorio como Fundamento

La aplicación sistemática del protocolo de análisis exploratorio de seis etapas (transformación Box-Cox, eliminación de tendencia, análisis ACF/PACF, tests de estacionariedad y linealidad, diagnóstico de residuos, análisis espectral) demostró valor predictivo consistente. Los hallazgos exploratorios anticiparon correctamente patrones en el desempeño relativo de los modelos:

- **Test de Tsay anticipó efectividad de métodos lineales:** El test de Tsay no rechazó linealidad en Exchange Rate ($p = 0.966$), anticipando correctamente que LSPM y LSPMW lograrían desempeño competitivo. En Electricity y Traffic, donde Tsay rechazó linealidad ($p < 0.01$), métodos no lineales o adaptativos dominaron.
- **Exponente de Hurst indicó necesidad de adaptatividad:** El $H = 0.933$ extremo en Exchange Rate señaló correctamente que LSPMW superaría a LSPM mediante ponderación temporal. Los valores moderados en Electricity y Traffic ($H \approx 0.6$) fueron consistentes con efectividad comparable de métodos adaptativos y no adaptativos.

- **Tests de heterocedasticidad justificaron métodos adaptativos:** El rechazo consistente de McLeod-Li en los tres datasets validó la necesidad de distribuciones predictivas adaptativas. AV-MCPS mostró ventaja en Traffic (mayor variabilidad temporal) pero desempeño modesto en Electricity (dinámica estable).
- **Análisis espectral informó configuración de hiperparámetros:** Los períodos dominantes identificados (Electricity: 168h, Traffic: 24h, Exchange Rate: 51h) permitieron configurar longitudes de bloque apropiadas en Circular Block Bootstrap, alineando el método con la estructura estacional real.

Este patrón de correspondencia entre diagnósticos exploratorios y desempeño predictivo valida que el protocolo no solo cumple función descriptiva sino que fundamenta decisiones de modelado con poder predictivo real. La inversión en análisis exploratorio exhaustivo produce retornos sustanciales en calidad de pronósticos.

5.5.4. Implicaciones para Aplicaciones Prácticas

Estrategia de Selección de Modelos

Basándose en los hallazgos empíricos, se propone la siguiente estrategia decisional para selección de modelos en aplicaciones de pronóstico operacional:

1. **Etapa 1 - Análisis Exploratorio Exhaustivo:** Aplicar el protocolo de seis etapas para caracterizar transformación óptima (λ), estacionariedad (ADF, KPSS), linealidad (Tsay), heterocedasticidad (McLeod-Li, ARCH-LM), memoria (exponente de Hurst) y estacionalidad (análisis espectral).
2. **Etapa 2 - Pre-selección Diagnóstico-Informada:** Identificar familias de modelos apropiadas según hallazgos exploratorios:
 - Si Tsay no rechaza ($p > 0.10$) y $H < 0.7$: Priorizar LSPM, MCPS.
 - Si Tsay rechaza ($p < 0.05$) o $H > 0.8$: Priorizar Sieve Bootstrap, LSPMW.
 - Si McLeod-Li rechaza ($p < 0.01$) con evidencia de cambios distribucionales: Priorizar AV-MCPS.
 - Si $n > 10000$ y estructura compleja confirmada: Considerar DeepAR.
3. **Etapa 3 - Validación Cruzada Temporal:** Evaluar modelos preseleccionados

mediante rolling forecast en conjunto de validación, optimizando hiperparámetros vía minimización de ECRPS.

4. **Etapa 4 - Selección Final con Análisis de Significancia:** Aplicar test de Diebold-Mariano modificado entre mejores modelos. Si diferencias son estadísticamente indistinguibles ($p > 0.10$), priorizar parsimonia, interpretabilidad y eficiencia computacional.
5. **Etapa 5 - Validación de Calibración:** Verificar calibración mediante histogramas PIT y curvas de confiabilidad en conjunto de prueba. Descartar modelos con patrones erráticos o concentraciones extremas, incluso si CRPS es aceptable.

Configuración de Hiperparámetros

El estudio reveló configuraciones robustas aplicables a contextos diversos:

- **Transformación Box-Cox:** Siempre estimar λ mediante método de Guerrero. Evitar heurísticas fijas (e.g., $\lambda = 0$ automático).
- **Eliminación de tendencia LOWESS:** $f = 0.05$ demostró robustez en los tres datasets. Valores mayores pueden introducir sesgo; valores menores pueden remover estructura estacional.
- **Partición temporal:** 70-75% entrenamiento, 15% validación, 10-15% prueba proporciona balance entre capacidad de aprendizaje y evaluación robusta.
- **Parámetro de decaimiento LSPMW:** $\rho \in [0.90, 0.99]$ con selección vía validación. Valores cercanos a 0.95 equilibran adaptatividad y estabilidad en la mayoría de contextos.
- **Longitud de bloque CBB:** Igualar al período estacional dominante identificado espectralmente. Evitar heurísticas genéricas no informadas por datos.

Métricas de Evaluación

El rechazo consistente de normalidad mediante test de Jarque-Bera en los tres datasets ($p < 0.001$), evidenciando leptocurtosis universal, justifica plenamente el uso de CRPS como métrica principal de evaluación. Las métricas gaussianas tradicionales (MSE, MAE)

no penalizan adecuadamente errores en las colas, críticos para gestión de riesgo y toma de decisiones bajo incertidumbre.

La validación de calibración mediante histogramas PIT debe complementar sistemáticamente métricas de precisión puntual. Distribuciones predictivas con CRPS aceptable pero calibración errática (patrones no uniformes en PIT) son inadecuadas para aplicaciones donde la correcta cuantificación de incertidumbre es crítica—gestión de inventarios bajo incertidumbre de demanda, pricing de derivados financieros, planificación de recursos energéticos con alta penetración renovable.

5.5.5. Limitaciones y Direcciones Futuras

Limitaciones del Estudio

El análisis comparativo presenta limitaciones que deben reconocerse al interpretar hallazgos:

1. **Horizonte de predicción unitario:** El estudio se limitó a $h = 1$ paso adelante. Horizontes multi-paso ($h > 1$) introducen complejidades adicionales (propagación de errores, dependencia temporal en predicciones sucesivas) que pueden alterar el ranking relativo de modelos. Series con estructura compleja pueden favorecer métodos recursivos o directos según el horizonte.
2. **Tamaño de muestra moderado:** Los tres datasets contienen aproximadamente 1800 observaciones de entrenamiento. Este régimen favorece métodos parsimoniosos sobre aprendizaje profundo. Con muestras del orden de decenas o cientos de miles de observaciones, el panorama competitivo podría reordenarse sustancialmente.
3. **Contexto univariado:** La evaluación se realizó en contexto puramente univariado. Métodos diseñados para explotación de información multivariada (AREPD) no pudieron demostrar sus capacidades distintivas. En aplicaciones con múltiples series relacionadas, estos métodos podrían exhibir ventajas competitivas no evidentes aquí.
4. **Ausencia de quiebres estructurales abruptos:** Los tres datasets, aunque exhibiendo variabilidad temporal, no presentan quiebres estructurales abruptos (cambios de régimen instantáneos). Métodos adaptativos podrían mostrar ventajas más pronunciadas en series con transiciones súbitas.

Direcciones de Investigación Futura

Los hallazgos sugieren múltiples direcciones prometedoras para investigación subsecuente:

1. **Extensión a horizontes multi-paso:** Evaluar comparativamente métodos recursivos, directos e híbridos para $h \in \{6, 12, 24\}$. Investigar si la ventaja de métodos adaptativos se amplifica con horizontes mayores donde deriva distribucional acumulada puede ser más pronunciada.
2. **Escalabilidad a datasets masivos:** Replicar el estudio comparativo con datasets del orden de 50,000-100,000 observaciones para evaluar si aprendizaje profundo alcanza superioridad consistente cuando limitaciones de tamaño de muestra se relajan.
3. **Contextos multivariados:** Extender la evaluación a pronóstico conjunto de múltiples series correlacionadas, permitiendo que métodos como AREPD exploten covariación temporal. Comparar con métodos univariados aplicados independientemente.
4. **Modelado de cambios de régimen:** Incorporar series con quiebres estructurales identificados (crisis financieras, cambios regulatorios, eventos climáticos extremos) para evaluar la capacidad de métodos adaptativos para detectar y responder rápidamente a transiciones.
5. **Optimización de hiperparámetros bayesiana:** Sustituir búsqueda en grilla por métodos bayesianos de optimización (e.g., Gaussian Process Optimization) para explorar eficientemente espacios de hiperparámetros de alta dimensión en modelos complejos.
6. **Ensambles de métodos heterogéneos:** Investigar si combinaciones de métodos con capacidades complementarias (e.g., LSPMW + DeepAR) mediante ensambles ponderados pueden superar consistentemente a métodos individuales mediante explotación de diversidad metodológica.
7. **Cuantificación de incertidumbre en la incertidumbre:** Extender el análisis para evaluar no solo calibración de distribuciones predictivas sino también confiabilidad de las estimaciones de calibración mismas, particularmente en regímenes de datos limitados.

5.5.6. Síntesis Final

El estudio comparativo exhaustivo de sistemas de predicción conformal y probabilística en tres series temporales reales ha validado empíricamente principios metodológicos fundamentales mientras ha revelado patrones de efectividad dependientes del contexto que desafían generalizaciones simplistas.

El hallazgo central—la inexistencia de un modelo universalmente superior—no es nihilista sino constructivo: fundamenta la necesidad de análisis exploratorio riguroso, selección diagnóstico-informada, y validación comparativa exhaustiva. La aplicación disciplinada del protocolo de seis etapas de análisis exploratorio demostró valor predictivo consistente, anticipando correctamente qué familias de modelos exhibirían ventajas competitivas en cada contexto.

Los métodos adaptativos (Sieve Bootstrap, LSPMW, AV-MCPS) emergieron como opciones robustas a través de contextos diversos, aunque sin dominar universalmente. Su capacidad para ajustar complejidad automáticamente o responder a cambios distribucionales confiere ventajas en series con heterogeneidad temporal, pero introduce complejidad innecesaria cuando la dinámica subyacente es estable y regular.

El aprendizaje profundo (DeepAR, EnCQR-LSTM) mostró mejora monotónica con complejidad estructural creciente pero no superó consistentemente a métodos parsimoniosos en el régimen de aproximadamente 1800 observaciones. Este hallazgo enfatiza la importancia de considerar tamaño de muestra disponible al evaluar trade-offs entre capacidad representacional y eficiencia muestral.

La validación exhaustiva de calibración mediante histogramas PIT y curvas de confiabilidad reveló que precisión puntual (CRPS) y calibración distribucional generalmente correlacionan pero admiten excepciones importantes. Métodos conformales mantuvieron calibración razonable incluso con precisión subóptima, mientras que ciertos métodos exhibieron colapso de calibración severo que los torna inadecuados para aplicaciones donde cuantificación correcta de incertidumbre es crítica.

En conjunto, los hallazgos proporcionan guía accionable para practicantes de pronóstico probabilístico: invertir en análisis exploratorio exhaustivo, seleccionar modelos mediante correspondencia entre capacidades metodológicas y características estructurales identificadas, validar mediante comparación estadística formal con test de Diebold-Mariano, y verificar calibración distribucional sistemáticamente. Esta disciplina metodológica, más

que la selección de algoritmos específicos, constituye el fundamento de pronósticos probabilísticos de alta calidad en aplicaciones operacionales.

6. Conclusiones

En este capítulo se sintetizan los hallazgos fundamentales derivados de la evaluación de los Sistemas de Predicción Conformal (CPS) y métodos de remuestreo aplicados a series de tiempo. Se discuten las implicaciones de los resultados obtenidos tanto en escenarios controlados como en aplicaciones reales, se exponen las limitaciones del estudio y se proponen rutas de investigación para el desarrollo futuro del área.

6.1. Conclusiones del Estudio

A partir de la evidencia empírica recolectada a través de las simulaciones y las aplicaciones a datos reales, se presentan las siguientes conclusiones:

- **Inexistencia de un modelo universalmente superior:** El hallazgo central de esta investigación es la confirmación de que no existe un único método de pronóstico probabilístico que domine todos los escenarios. La efectividad de la Predicción Conformal (CP) y el *Bootstrapping* depende críticamente de la alineación entre las capacidades del modelo y las propiedades estructurales de la serie (estacionariedad, linealidad y memoria).
- **Robustez excepcional del Sieve Bootstrap:** A través de todas las dimensiones evaluadas, el *Sieve Bootstrap* emergió como el método más robusto y consistente. Su capacidad para adaptar automáticamente la complejidad del modelo mediante el filtrado autorregresivo le permitió manejar la no estacionariedad y la memoria larga sin necesidad de preprocesamiento manual, superando incluso a arquitecturas de aprendizaje profundo en regímenes de datos moderados.
- **Rol crítico de la diferenciación en procesos integrados:** Las simulaciones demostraron que, para la mayoría de los métodos conformales (LSPM, MCPS), la diferenciación es una condición *sine qua non* para evitar el colapso del desempeño en

procesos ARIMA. Omitir este paso degrada la precisión hasta en un 90%, invalidando las garantías de cobertura en presencia de tendencias estocásticas.

- **Ventajas de la adaptatividad en heterocedasticidad:** Los modelos desarrollados en esta tesis, específicamente el **AV-MCPS** y el **LSPMW**, demostraron una superioridad clara en series con volatilidad cambiante (como el tráfico vehicular y tipos de cambio). La incorporación de la volatilidad local como dimensión de particionamiento en el enfoque de Mondrian permite generar intervalos más nítidos y mejor calibrados que los métodos conformales estándar.
- **Escalabilidad del Aprendizaje Profundo:** Los resultados de *DeepAR* y *EnCQR-LSTM* indican que estas arquitecturas requieren una escala de datos superior a la disponible en este estudio (1,800 observaciones) para superar a los modelos parsimoniosos. No obstante, su mejora relativa al transitar hacia series más complejas sugiere un potencial significativo en contextos de *Big Data* y pronóstico multivariado.
- **Validación del Protocolo Exploratorio:** Se demostró que un análisis exploratorio exhaustivo (Tests de Tsay, McLeod-Li y Exponente de Hurst) posee un alto poder predictivo sobre el éxito de las familias de modelos. La linealidad en media y la persistencia de la memoria son los mejores predictores para seleccionar entre métodos conformales lineales o adaptativos.

6.2. Limitaciones del Estudio

A pesar del rigor metodológico, se identifican las siguientes limitaciones que delimitan el alcance de los resultados:

1. **Horizonte de predicción:** La mayor parte del estudio se centró en el pronóstico a un paso adelante ($h = 1$). Aunque la simulación de predicción multi-paso reveló patrones de degradación importantes, el comportamiento de las garantías conformales en horizontes muy largos requiere una formalización teórica más profunda.
2. **Restricción Univariada:** El análisis se limitó a series temporales univariadas. Métodos diseñados para explotar información cruzada entre series relacionadas no pudieron desplegar su máxima capacidad representacional.
3. **Tamaño de Muestra Finito:** Con un presupuesto de datos de aproximada-

mente 2,000 observaciones, el estudio favoreció intrínsecamente a modelos con pocos parámetros. Las conclusiones sobre el desempeño de modelos de redes neuronales podrían variar en regímenes de datos masivos.

4. **Ausencia de quiebres estructurales abruptos:** Los conjuntos de datos reales analizados no presentaron cambios de régimen instantáneos (saltos en el nivel o varianza por eventos exógenos), lo que limitó la evaluación de la velocidad de recuperación de los métodos conformales adaptativos ante choques sistémicos.

6.3. Investigación Futura

Como extensiones naturales de este trabajo, se proponen las siguientes líneas de investigación:

- **Extensión a Pronóstico Multivariado:** Desarrollar transductores conformales que incorporen dependencias cruzadas y permitan la construcción de regiones de predicción conjuntas (elipsoides de confianza) para múltiples series relacionadas.
- **Optimización Bayesiana de Hiperparámetros:** Sustituir la búsqueda en grilla por métodos de optimización basados en procesos gaussianos para explorar de manera más eficiente los espacios de alta dimensión en modelos como el AV-MCPS y EnCQR-LSTM.
- **Modelos Híbridos Dinámicos:** Investigar ensambles ponderados que combinen la robustez del *Sieve Bootstrap* en la estimación de la media con la precisión del AV-MCPS en la cuantificación de los residuos, adaptando los pesos según el régimen de volatilidad detectado.
- **Teoría de Consistencia Universal en Series No Estacionarias:** Avanzar en la demostración formal de la consistencia de Vovk para procesos ergódicos bajo condiciones de mezcla fuerte (α -mixing), extendiendo el marco teórico de este trabajo hacia una base matemática más general para series de tiempo complejas.

Finalizando, este trabajo reafirma que la Predicción Conformal no es solo una alternativa a los métodos estadísticos tradicionales, sino un marco de trabajo esencial que proporciona la seguridad estadística necesaria para la toma de decisiones bajo incertidumbre en el dinámico dominio de las series temporales.

A. Marco Teórico para Consistencia Universal en Series de Tiempo Ergódicas

Este apéndice presenta el desarrollo conceptual realizado durante la investigación para extender la teoría de consistencia universal de Vovk 2019 del modelo IID a procesos estocásticos dependientes, específicamente series de tiempo ergódicas. Aunque no se llegó a una demostración formal completa, se identificaron los componentes fundamentales que conformarían una prueba futura.

A.1. Redefinición del Objetivo de Consistencia

En el contexto de series de tiempo, el objetivo de un CPS universalmente consistente es que la distribución predictiva generada, $Q_n(y)$, converja débilmente en probabilidad a la verdadera distribución condicional $F_{Y|X}(\cdot|X_{n+1})$. A diferencia del caso IID, aquí la “consistencia” implica que el sistema debe ser capaz de aprender la dinámica local y la estructura de dependencia del proceso a medida que la serie evoluciona.

Se plantea que, bajo este régimen, el sistema no solo debe ser asintóticamente válido, sino también *eficiente*, adaptándose a la heterocedasticidad (volatilidad cambiante) intrínseca de los datos secuenciales.

A.2. Supuestos Fundamentales del Marco Propuesto

Para transitar de la teoría de Vovk a procesos dependientes, se han identificado los siguientes supuestos como pilares necesarios para el desarrollo de una prueba de consistencia

futura:

1. **Estacionariedad y Ergodicidad:** Se asume que el proceso $\{Z_t\}$ es estrictamente estacionario y ergódico. Esto garantiza que los promedios temporales observados en la ventana de datos converjan a los promedios del ensamble, permitiendo que el sistema “aprenda” de la historia pasada.
2. **Condición de α -mixing (Mezcla Fuerte):** Para manejar la dependencia, se requiere que el proceso sea α -mixing con coeficientes que decaigan algebraicamente ($\alpha(k) \leq Ck^{-\beta}, \beta > 2$). Este supuesto es crucial para aplicar teoremas límite central y asegurar que las observaciones lejanas en el tiempo sean casi independientes.
3. **Regularidad de Lipschitz:** A diferencia del enfoque de histograma de celdas discretas, aquí se asume que tanto la función de regresión $\mu(x)$ como la distribución de los residuos $G(s|x)$ son Lipschitz continuas respecto al espacio de covariables. Esto asegura que puntos cercanos en el tiempo tengan comportamientos predictivos similares.

A.3. Mecanismo Propuesto: Transductor Conformal por Kernel

En lugar del enfoque de histogramas anidados de Vovk, este marco propone una arquitectura adaptativa basada en dos componentes:

- **Ventana Temporal Móvil (L_n):** Un mecanismo de truncamiento que selecciona las últimas L_n observaciones. Para alcanzar la consistencia, el tamaño de esta ventana debe crecer con n pero a un ritmo controlado ($L_n \rightarrow \infty$, pero $L_n/n \rightarrow 0$).
- **Suavizado Espacial por Kernel (K, h_n):** En lugar de asignar pesos uniformes a la celda (como en Mondrian), se propone el uso de pesos de relevancia espacial $w_i = K\left(\frac{d(X_i, X_{n+1})}{h_n}\right)$. Esto permite que el sistema pondere los residuos pasados no solo por su cercanía temporal, sino por su similitud en el espacio de características.

El ancho de banda h_n debe satisfacer condiciones de equilibrio similares a las de Vovk: $h_n \rightarrow 0$ (para reducir el sesgo) y $L_n h_n^p \rightarrow \infty$ (para asegurar suficientes vecinos dentro de la ventana de kernel, donde p es la dimensión del espacio de características).

A.4. Ruta hacia la Demostración de Consistencia

La convergencia propuesta $\int f dQ_n \rightarrow \mathbb{E}_P(f|X_{n+1})$ en probabilidad requeriría demostrar dos resultados clave:

A.4.1. Convergencia del Sesgo

Por el teorema de Lévy para martingalas (usado por Vovk en el caso IID), en el contexto ergódico se necesitaría probar que:

$$\mathbb{E}_P(f|\mathcal{K}_{h_n}(X_{n+1})) - \mathbb{E}_P(f|X_{n+1}) \rightarrow 0 \quad (\text{A-1})$$

donde $\mathcal{K}_{h_n}(X_{n+1})$ representa la vecindad kernel de X_{n+1} . La regularidad de Lipschitz de $\mathbb{E}_P(f|x)$ garantizaría este resultado.

A.4.2. Convergencia de la Varianza

Utilizando la condición de α -mixing, se podría aplicar una ley de los grandes números para procesos dependientes, demostrando que:

$$\frac{1}{N_{h_n}} \sum_{i \in \mathcal{I}_{h_n}} f(Y_i) - \mathbb{E}_P(f|\mathcal{K}_{h_n}(X_{n+1})) \rightarrow 0 \quad (\text{A-2})$$

donde \mathcal{I}_{h_n} es el conjunto de índices en la vecindad kernel y $N_{h_n} = |\mathcal{I}_{h_n}|$.

A.5. Discusión y Perspectivas Futuras

Esta formulación plantea que la convergencia depende del balance entre el sesgo del kernel y la varianza inducida por la dependencia de los datos. Mientras que Vovk utiliza el Teorema de Lévy para martingalas bajo independencia, el análisis en series de tiempo requiere el uso de técnicas de *análisis de sesgo-varianza para estimadores no paramétricos en procesos mixing*.

Es importante notar que este planteamiento se presenta como una *hoja de ruta teórica*. La

validación rigurosa de que este transductor conformal ergódico alcanza la consistencia universal bajo cualquier proceso mixing representaría una extensión significativa del trabajo original de Vovk, unificando la robustez de la predicción conformal con la flexibilidad de la estimación no paramétrica para datos secuenciales de alta complejidad. Los experimentos numéricos presentados en los capítulos principales de esta tesis sugieren que la aproximación práctica basada en estos principios produce resultados prometedores, aunque la justificación teórica completa queda como trabajo futuro.

Bibliografía

- Alexandrov, Alexander et al. (2020). “GluonTS: Probabilistic and Neural Time Series Modeling in Python”. In: *Journal of Machine Learning Research* 21.116, pp. 1–6.
- 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”. In: *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 February 8, 2026. arXiv: [2202.13415 \[stat.ME\]](https://arxiv.org/abs/2202.13415).
- Ben Taieb, Souhaib et al. (2012). “A review and comparison of strategies for multi-step ahead time series forecasting”. In: *Expert Systems with Applications* 39.8, pp. 7067–7083.
- Bontempi, Gianluca, Souhaib Ben Taieb, and Yann-Aël Le Borgne (2013). “Machine learning strategies for time series forecasting”. In: *Business Intelligence: Second European Summer School, eBISS 2012*. Springer, pp. 62–77.
- Boström, Henrik (2022). “crepes: a Python Package for Generating Conformal Regressors and Predictive Systems”. In: *Conformal and Probabilistic Prediction with Applications*. Vol. 179. PMLR, pp. 24–41.
- Boström, Henrik, Ulf Johansson, and Tuve Löfström (2021). “Mondrian Conformal Predictive Distributions”. In: *Proceedings of Machine Learning Research* 152, pp. 24–38.
- Box, George E. P. and David R. Cox (1964). “An Analysis of Transformations”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 26.2, pp. 211–252.
- Box, George E. P., Gwilym M. Jenkins, et al. (2015). *Time Series Analysis: Forecasting and Control*. 5th ed. Hoboken, NJ: John Wiley & Sons. ISBN: 978-1-118-67502-1.
- Brock, William A. et al. (1996). “A test for independence based on the correlation dimension”. In: *Econometric Reviews* 15.3, pp. 197–235.
- Bühlmann, Peter (1997). “Sieve Bootstrap for Time Series”. In: *Bernoulli* 3.2, pp. 123–148. DOI: [10.2307/3318434](https://doi.org/10.2307/3318434).

- Chan, Kung-Sik and Howell Tong (1985). "Testing for threshold autoregression". In: *The Annals of Statistics* 13.3, pp. 1121–1142.
- Chen, Pu and Willi Semmler (2023). "Stability in Threshold VAR Models". In: *Studies in Nonlinear Dynamics and Econometrics*. DOI: [10.1515/snnde-2022-0099](https://doi.org/10.1515/snnde-2022-0099).
- Chen, Tianqi and Carlos Guestrin (2016). "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 785–794.
- Coroneo, Laura and Fabrizio Iacone (2020). "Comparing Predictive Accuracy in Small Samples Using Fixed-Smoothing Asymptotics". In: *Journal of Applied Econometrics* 35.3, pp. 391–409. DOI: [10.1002/jae.2756](https://doi.org/10.1002/jae.2756).
- Dawid, A. Philip (1984). "Statistical Theory: The Prequential Approach". In: *Journal of the Royal Statistical Society. Series A (General)* 147.2, pp. 278–292.
- Dickey, David A. and Wayne A. Fuller (1979). "Distribution of the Estimators for Autoregressive Time Series with a Unit Root". In: *Journal of the American Statistical Association* 74.366, pp. 427–431. DOI: [10.2307/2286348](https://doi.org/10.2307/2286348).
- Diebold, Francis X., Todd A. Gunther, and Anthony S. Tay (1998). "Evaluating Density Forecasts with Applications to Financial Risk Management". In: *International Economic Review* 39.4, pp. 863–883.
- Diebold, Francis X. and Roberto S. Mariano (1995). "Comparing Predictive Accuracy". In: *Journal of Business & Economic Statistics* 13.3, pp. 253–263. DOI: [10.1080/07350015.1995.10524599](https://doi.org/10.1080/07350015.1995.10524599).
- Gneiting, Tilmann, Fadoua Balabdaoui, and Adrian E. Raftery (2007). "Probabilistic Forecasts, Calibration and Sharpness". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.2, pp. 243–268. DOI: [10.1111/j.1467-9868.2007.00587.x](https://doi.org/10.1111/j.1467-9868.2007.00587.x).
- Gneiting, Tilmann and Matthias Katzfuss (2014). "Probabilistic Forecasting". In: *Annual Review of Statistics and Its Application* 1, pp. 125–151. DOI: [10.1146/annurev-statistics-062713-085831](https://doi.org/10.1146/annurev-statistics-062713-085831).
- Gneiting, Tilmann and Adrian E. Raftery (2007). "Strictly Proper Scoring Rules, Prediction, and Estimation". In: *Journal of the American Statistical Association* 102.477, pp. 359–378. DOI: [10.1198/016214506000001437](https://doi.org/10.1198/016214506000001437).
- Guerrero, Víctor M. (1993). "Time-series analysis supported by power transformations". In: *Journal of Forecasting* 12.1, pp. 37–48.
- Hamill, Thomas M. (2001). "Interpretation of Rank Histograms for Verifying Ensemble Forecasts". In: *Monthly Weather Review* 129.3, pp. 550–560.

- Harvey, David, Stephen Leybourne, and Paul Newbold (1997). “Testing the Equality of Prediction Mean Squared Errors”. In: *International Journal of Forecasting* 13.2, pp. 281–291. DOI: [10.1016/S0169-2070\(96\)00719-4](https://doi.org/10.1016/S0169-2070(96)00719-4).
- Hurst, Harold Edwin (1951). “Long-term storage capacity of reservoirs”. In: *Transactions of the American Society of Civil Engineers* 116, pp. 770–799.
- Hyndman, Rob J and George Athanasopoulos (2021). *Forecasting: Principles and Practice*. 3rd. Melbourne, Australia: OTexts. URL: <https://otexts.com/fpp3/>.
- Hyndman, Rob J. and Yanan Fan (1996). “Sample Quantiles in Statistical Packages”. In: *The American Statistician* 50.4, pp. 361–365.
- Jarque, Carlos M. and Anil K. Bera (1987). “A Test for Normality of Observations and Regression Residuals”. In: *International Statistical Review* 55.2, pp. 163–172. DOI: [10.2307/1403192](https://doi.org/10.2307/1403192).
- Jensen, Vilde, Filippo Maria Bianchi, and Stian Normann Anfinsen (2022). “Ensemble Conformalized Quantile Regression for Probabilistic Time Series Forecasting”. Version v2. In: *arXiv preprint arXiv:2202.08756*. Submitted to IEEE Transactions on Neural Networks and Learning Systems. arXiv: [2202.08756 \[cs.LG\]](https://arxiv.org/abs/2202.08756). URL: <https://arxiv.org/abs/2202.08756>.
- Kiefer, Nicholas M. and Timothy J. Vogelsang (2005). “A new asymptotic theory for heteroskedasticity-autocorrelation robust tests”. In: *Econometric Theory* 21.6, pp. 1130–1164. DOI: [10.1017/S026646605050565](https://doi.org/10.1017/S026646605050565).
- Kingma, Diederik P. and Jimmy Ba (2015). “Adam: A Method for Stochastic Optimization”. In: *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. Originally published as arXiv:1412.6980 [cs.LG].
- Kwiatkowski, Denis et al. (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?” In: *Journal of Econometrics* 54.1-3, pp. 159–178. DOI: [10.1016/0304-4076\(92\)90104-Y](https://doi.org/10.1016/0304-4076(92)90104-Y).
- Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer Series in Statistics. New York: Springer. ISBN: 978-1-4419-1848-2. DOI: [10.1007/978-1-4757-3803-2](https://doi.org/10.1007/978-1-4757-3803-2).
- Ljung, Greta M. and George E. P. Box (1978). “On a measure of lack of fit in time series models”. In: *Biometrika* 65.2, pp. 297–303. DOI: [10.1093/biomet/65.2.297](https://doi.org/10.1093/biomet/65.2.297).
- McLeod, A. Ian and W. K. Li (1983). “Diagnostic checking ARMA time series models using squared-residual autocorrelations”. In: *Journal of Time Series Analysis* 4.4, pp. 269–273.

- Petruccelli, Joseph D and Samuel W Woolford (1984). “A consistent test for nonstationarity based on residuals”. In: *Journal of the American Statistical Association* 79.387, pp. 611–616.
- Politis, Dimitris N. and Joseph P. Romano (1992). “A circular block-resampling procedure for stationary data”. In: *Exploring the Limits of Bootstrap*. Ed. by Raoul LePage and Lynne Billard. New York: Wiley, pp. 263–270.
- Politis, Dimitris N. and Halbert White (2004). “Automatic Block-Length Selection for the Dependent Bootstrap”. In: *Journal of the American Statistical Association* 99.465, pp. 154–164. DOI: [10.1198/016214504000000214](https://doi.org/10.1198/016214504000000214).
- Salinas, David et al. (2020). “DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks”. In: *International Journal of Forecasting* 36.3, pp. 1181–1191. DOI: [10.1016/j.ijforecast.2019.07.001](https://doi.org/10.1016/j.ijforecast.2019.07.001). arXiv: [1704.04110](https://arxiv.org/abs/1704.04110).
- Stankeviciute, Kamile, Ahmed M Alaa, and Mihaela van der Schaar (2021). “Conformal time-series forecasting”. In: *Advances in Neural Information Processing Systems*. Vol. 34, pp. 6292–6304.
- Thorarinsdottir, Thordis L. and Nina Schuhéen (2017). “Verification: Assessment of Calibration and Accuracy”. In: *Norwegian Computing Center SAMBA/17/17*.
- Tsay, Ruey S. (1986). “Nonlinearity tests for time series”. In: *Biometrika* 73.2, pp. 461–466.
- Vovk, Vladimir (2019). “Universally consistent conformal predictive distributions”. In: *Proceedings of the Eighth Workshop on Conformal and Probabilistic Prediction and Applications (COPA 2019)*. Vol. 105. Proceedings of Machine Learning Research. PMLR, pp. 105–122. URL: <http://proceedings.mlr.press/v105/vovk19a.html>.
- Vovk, Vladimir, Alexander Gammerman, and Glenn Shafer (2005). *Algorithmic Learning in a Random World*. New York: Springer. DOI: [10.1007/b138548](https://doi.org/10.1007/b138548).
- (2022). *Algorithmic Learning in a Random World*. Second. Cham, Switzerland: Springer. ISBN: 978-3-031-06648-1. DOI: [10.1007/978-3-031-06649-8](https://doi.org/10.1007/978-3-031-06649-8).
- Vovk, Vladimir, Ilia Nouretdinov, et al. (2017). *Conformal predictive distributions with kernels*. Working paper 20. Working Paper 20. Accessed on February 8, 2026. On-line compression modelling project (new series). URL: <http://alrw.net/articles/CPDK.pdf>.
- Welch, Peter D. (1967). “The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms”. In: *IEEE Transactions on Audio and Electroacoustics* 15.2, pp. 70–73. DOI: [10.1109/TAU.1967.1161901](https://doi.org/10.1109/TAU.1967.1161901).

- Ye, Tinghan, Amira Hijazi, and Pascal Van Hentenryck (2025). “Conformal Predictive Distributions for Order Fulfillment Time Forecasting”. In: *arXiv preprint arXiv:2505.17340*. Version 2.
- Yu, Bin (1994). “Rates of Convergence for Empirical Processes of Stationary Mixing Sequences”. In: *The Annals of Probability* 22.1, pp. 94–116.