“Evaluación de modelos para el pronóstico del precio externo del café”

Autor

Sergio Andrés Rojas Trujillo

Director

Jesús Otero

Magíster en Finanzas Cuantitativas

Facultad de Economía

Maestría en Finanzas Cuantitativas

Universidad del Rosario

Bogotá – Colombia

2026

Resumen

Esta tesis evalúa el desempeño predictivo de nueve modelos para el pronóstico mensual del precio externo del café colombiano, empleando un esquema de validación fuera de muestra con tres particiones temporales (de 18, 24 y 36 observaciones de prueba) y dos horizontes de pronóstico ($h = 1$ y $h = 3$ meses). Los modelos comparados incluyen tres benchmarks ingenuos, dos modelos estadísticos univariados (ARIMA, ETS), un modelo de vectores autorregresivos (VAR) con covariables macroeconómicas y climáticas, un modelo de bosques aleatorios y dos estrategias de combinación de pronósticos: promedio simple y ponderación por el inverso del RMSE acumulado. La muestra comprende 315 observaciones mensuales entre enero de 2000 y marzo de 2026. La variable objetivo se aproxima mediante el contrato continuo KC1 de Bloomberg, usado como proxy del precio externo del café. La métrica principal de comparación es el OWA (*Overall Weighted Average*), calculado en relación con el benchmark Naive 2, complementado por sMAPE, MASE, RMSE y MAE. La robustez de los resultados se evalúa mediante el test de Diebold-Mariano con corrección de muestra finita.

Para $h = 1$, el VAR y Naive 1 empatan en el primer lugar con OWA promedio de 0,937 y 0,938 respectivamente, seguidos por ARIMA (0,956) y ETS (0,957). Los cuatro modelos superan al benchmark en las tres particiones sin excepción. Para $h = 3$, el ordenamiento cambia: Naive 1 registra el mejor OWA (0,973), seguido por VAR (0,991) y ETS (0,992). En ambos horizontes, el Random Forest y las dos estrategias de combinación no superan al benchmark. El test de Diebold-Mariano no rechaza la igualdad predictiva al 5 %, resultado coherente con el limitado poder estadístico disponible ($n \leq 36$) y con la magnitud moderada de las mejoras observadas. Los resultados se interpretan como evidencia específica del esquema y la muestra analizados, sin generalización a otros contextos o períodos.

Palabras clave: pronóstico de series de tiempo, precio del café, ARIMA, ETS, VAR, bosques aleatorios, combinación de pronósticos, OWA, validación fuera de muestra.

Clasificación JEL: C22, C53, Q11, Q17.

Abstract

This thesis evaluates the predictive performance of nine models for monthly point forecasting of the external Colombian coffee price under a rolling-origin out-of-sample evaluation scheme with three temporal holdout partitions (of 18, 24, and 36 test observations) and two forecast horizons ($h = 1$ and $h = 3$ months ahead). The models under comparison include three naïve benchmarks, two classical univariate models (ARIMA, ETS), a vector autoregression (VAR) with macroeconomic and climatic covariates, a random forest, and two forecast combination strategies: a simple average and an inverse-RMSE weighted average. The sample covers 315 monthly observations from January 2000 to March 2026. The target variable is proxied by Bloomberg's continuous KC1 contract as a monthly benchmark for the external coffee price. The primary evaluation criterion is the OWA (*Overall Weighted Average*), computed relative to the Naive 2 benchmark and supplemented by sMAPE, MASE, RMSE, and MAE. Statistical robustness is assessed using the Diebold-Mariano test with the finite-sample correction of Harvey, Leybourne, and Newbold (1997).

For $h = 1$, VAR and Naive 1 tie for first place with average OWA of 0.937 and 0.938, respectively, followed by ARIMA (0.956) and ETS (0.957); all four models outperform the benchmark across all three partitions without exception. For $h = 3$, the ranking shifts: Naive 1 records the best OWA (0.973), followed by VAR (0.991) and ETS (0.992). At both horizons, the random forest and both combination strategies fail to outperform the Naive 2 benchmark. The Diebold-Mariano test does not reject equal predictive ability at the 5 % level, a result consistent with the limited statistical power available ($n \leq 36$) and the moderate magnitude of the observed improvements. Results are interpreted as evidence specific to the evaluated scheme and sample period, without generalization beyond the study context.

Keywords: time series forecasting, coffee price, ARIMA, ETS, VAR, random forest, forecast combination, OWA, out-of-sample evaluation.

JEL Classification: C22, C53, Q11, Q17.

Índice

1. Introducción

El café ocupa un lugar singular en la estructura productiva y exportadora de Colombia. Como tercer productor mundial y uno de los principales exportadores de café suave lavado, el país concentra en este cultivo una fracción significativa de sus divisas agrícolas, del empleo rural y de la actividad económica de amplias regiones (Aduteye et al., 2023). El precio externo del café relevante para Colombia se forma en estrecha conexión con el mercado de futuros arábica de la ICE en Nueva York. En esta tesis se aproxima mediante el contrato continuo KC1 de Bloomberg, utilizado como proxy replicable del precio externo mensual. Esta elección no equipara el precio del futuro con el precio físico de exportación observado; más bien, adopta el benchmark internacional que concentra la señal de precios del mercado externo. La Organización Internacional del Café (ICO) documenta, para el grupo *Colombian Milds*, una relación lineal muy estrecha entre los precios físicos y los futuros de Nueva York, con $R^2 = 0{,}9212$ en la regresión spot–futuros (International Coffee Organization, 2011), magnitud suficiente para tratar a KC1 como una aproximación empíricamente fundada del precio externo relevante para Colombia. Esta variable constituye, por tanto, una referencia de primer orden para las finanzas públicas sectoriales, para las decisiones de inversión de los productores y para las instituciones de financiamiento y cobertura que operan en la cadena de valor del grano (Useche Mahecha, 2024). La relevancia económica del sector ha motivado, desde décadas atrás, distintos esfuerzos de modelación orientados a analizar su producción, consumo y comercio internacional (Cepeda C., 1981). Sin embargo, el problema de pronosticar el precio externo con precisión fuera de muestra —es decir, en un horizonte genuinamente desconocido en el momento de elaborar la predicción— conserva hoy una dificultad estadística considerable.

Esa dificultad tiene raíces teóricas bien documentadas. El precio mensual del café comparte con otros productos básicos las propiedades que hacen difícil su pronóstico: alta volatilidad idiosincrásica, episodios prolongados de tendencia seguidos de reversiones abruptas, estacionalidad débil o variable, e integración de orden uno en la mayoría de los períodos (Kwas & Rubaszek, 2021). En un mercado donde los participantes procesan continuamente información sobre condiciones climáticas en Brasil y Vietnam, sobre la dinámica del dólar, sobre las tasas de interés de largo plazo y sobre la actividad especulativa en mercados de futuros, la pregunta de si es posible construir modelos que superen de manera consistente al paseo aleatorio en horizontes de uno a

tres meses es empíricamente abierta. La literatura sobre productos básicos ha mostrado que incluso modelos estructurales elaborados no superan sistemáticamente una previsión de no-cambio cuando son evaluados en tiempo real y bajo condiciones estrictamente fuera de muestra (Kwas & Rubaszek, 2021). Este patrón —comparable al documentado en series de tipos de cambio desde los trabajos seminales de la década de 1980— impone un estándar de comparación exigente: cualquier modelo que aspire a ser útil para la gestión del riesgo en este mercado debe primero demostrar que supera a alternativas triviales antes de justificar su complejidad adicional.

La literatura empírica sobre pronóstico del precio del café ha crecido en años recientes, impulsada en parte por la proliferación de métodos de aprendizaje automático. Los trabajos disponibles cubren un espectro amplio de enfoques: Naveena et al. (2017) aplican modelos ARIMA e híbridos para el café arábica en India; Deina et al. (2022) proponen máquinas de aprendizaje extremo (ELM) y las comparan con ARIMA en horizontes cortos; Cardozo Rueda (2022) exploran redes neuronales para el caso colombiano; Y. Chen (2024) y Hwase y Fofanah (2021) aplican múltiples algoritmos de aprendizaje automático para el mercado del café con énfasis en futuros y precios de bolsa. Para el mercado colombiano en particular, Díaz-Pinzón (2025) construye modelos de predicción y proyección del precio interno del café pergamino seco, y Pinto-Rodriguez et al. (2025) presentan una revisión sistemática de métodos para la estimación de precios de café de especialidad. Más allá del café, en el espacio más amplio del pronóstico de productos básicos, Ly et al. (2021) aplican redes LSTM al pronóstico de algodón y petróleo y concluyen que los métodos de aprendizaje automático no superan sistemáticamente a ARIMA fuera de muestra, resultado coherente con la evidencia de Hamouda et al. (2025) para productos básicos energéticos, donde la ventaja de los modelos de inteligencia artificial sobre los estadísticos depende fuertemente del contexto empírico.

Una lectura transversal de esta literatura revela, no obstante, una brecha metodológica relevante. La mayoría de los trabajos existentes comparan uno o dos enfoques seleccionados —habitualmente el modelo propuesto frente a ARIMA— sin incluir el espectro completo de alternativas que va desde los benchmarks ingenuos hasta las estrategias de combinación de pronósticos. Adicionalmente, son pocos los estudios que adoptan protocolos de validación fuera de muestra comparables al *rolling-origin expanding window* con múltiples particiones temporales, que es el estándar recomendado en la literatura metodológica para producir estimaciones robustas de la

precisión predictiva (Hyndman & Athanasopoulos, 2021; Tashman, 2000). Sin ese protocolo, la evidencia sobre qué modelo es "mejor" puede ser sensible al período de evaluación elegido y difícilmente comparable entre estudios. La ausencia de un ejercicio comparativo amplio, reproducible y documentado que evalúe simultáneamente benchmarks ingenuos, modelos univariados clásicos, un modelo multivariado con covariables, un modelo de aprendizaje automático y combinaciones de pronósticos —bajo el mismo protocolo de validación y con las mismas métricas— para el precio externo del café colombiano es, en síntesis, la motivación central de esta tesis.

Con ese propósito, el estudio evalúa y compara el desempeño predictivo de nueve modelos para el pronóstico mensual del precio externo del café colombiano, sobre una serie mensual del contrato continuo KC1 como proxy del precio externo en el período enero de 2000 a marzo de 2026 (315 observaciones mensuales). Los modelos incluyen tres benchmarks ingenuos (Naive 1, Naive S, Naive 2), dos modelos univariados clásicos (ARIMA y ETS), un modelo de vectores autorregresivos con covariables macroeconómicas y climáticas (VAR), un modelo de bosques aleatorios (Random Forest) y dos estrategias de combinación de pronósticos (simple y ponderada). La pregunta que organiza el análisis es directa: ¿qué modelo ofrece el mejor desempeño predictivo fuera de muestra en horizontes de $h = 1$ y $h = 3$ meses, medido por el OWA relativo al benchmark Naive 2, bajo un esquema de validación con ventana expandible?

Las contribuciones de la tesis son de naturaleza metodológica y empírica. En primer lugar, el diseño de evaluación adopta la convención de las competencias M de pronóstico (Makridakis et al., 2020): el OWA —promedio entre el cociente sMAPE y el cociente MASE respecto al benchmark— se emplea como métrica primaria por ser invariante a la escala y directamente interpretable como superioridad relativa al paseo aleatorio estacional. El MASE, propuesto por Hyndman y Koehler (2006) como medida escalada y robusta, y el sMAPE completan el cuadro de indicadores. En segundo lugar, el estudio incorpora tres ventanas de evaluación independientes (18, 24 y 36 meses de observaciones fuera de muestra), siguiendo la recomendación de Tashman (2000) de desensibilizar la evidencia frente a eventos particulares de cualquier subperíodo. La consistencia del ordenamiento de los modelos a través de estas tres ventanas constituye el principal criterio de robustez del estudio. En tercer lugar, la comparación se complementa con el test de Diebold y Mariano (1995), corregido por Harvey et al. (1997) para muestras finitas, que permite evaluar la significancia estadística de las diferencias de precisión observadas. Finalmente, el pipe-

line computacional está implementado en Python en módulos separados y documentados, lo que garantiza que todos los resultados son completamente reproducibles a partir del código y los datos disponibles.

El resto del documento se organiza de la siguiente manera. La sección 2 presenta la revisión de literatura sobre dinámica del mercado del café, pronóstico de productos básicos, y las cuatro familias de modelos evaluadas. La sección 3 describe el diseño metodológico: métricas de evaluación, esquema de validación fuera de muestra y protocolo comparativo. La sección 4 detalla las fuentes de datos, el proceso de construcción del dataset maestro mensual y las características descriptivas de las series. La sección 5 documenta la implementación computacional del pipeline y la especificación de cada modelo. La sección 6 presenta y discute los resultados empíricos, incluyendo el análisis de significancia estadística mediante el test de Diebold-Mariano. La sección 7 sintetiza los hallazgos, responde la pregunta de investigación, discute las limitaciones del estudio y propone líneas de trabajo futuro.

2. Revisión de literatura

2.1. Dinámica del mercado cafetero y el desafío del pronóstico

El café es uno de los bienes agrícolas más intensamente transados en los mercados internacionales, y su precio en bolsa constituye una señal de primer orden para productores, exportadores e instituciones de financiamiento a lo largo de toda la cadena de valor (Aduteye et al., 2023). El índice de precios compuesto publicado por la Organización Internacional del Café (ICO) —referencia estándar en estudios académicos y de política sectorial— resume la interacción de fuerzas heterogéneas que operan simultáneamente sobre la cotización: por el lado de la oferta, las condiciones climáticas en Brasil y Vietnam, la disponibilidad de inventarios globales y los ciclos de producción plurianuales; por el lado de la demanda, la expansión del consumo en economías emergentes y las preferencias cambiantes de los mercados de especialidad; y, en la dimensión financiera, el comportamiento del dólar estadounidense, las tasas de interés de largo plazo y la actividad especulativa de operadores no comerciales en los mercados de futuros de la Bolsa Intercontinental de Nueva York (ICE). Esta multiplicidad de determinantes hace que la serie mensual del precio del

café exhiba propiedades características de otros productos básicos: alta volatilidad idiosincrásica, episodios prolongados de tendencia seguidos de reversiones abruptas y una estacionalidad débil o variable, rasgos documentados en la literatura comparada de productos básicos desde la década de 1980 (Kwas & Rubaszek, 2021).

Para Colombia, tercer productor mundial de café y exportador dominante en el segmento de suave lavado, la relevancia macroeconómica del precio externo trasciende el ámbito sectorial: afecta directamente los ingresos de los productores rurales, el comportamiento de las exportaciones agrícolas y las decisiones de política agropecuaria nacional (Useche Mahecha, 2024). La modelación del precio del café colombiano tiene antecedentes que se remontan a los primeros intentos de formalización econométrica del sector (Cepeda C., 1981), pero la pregunta sobre qué modelo produce pronósticos más precisos fuera de muestra permanece empíricamente abierta. Esta apertura no es accidental: responde a la naturaleza estadística de la serie. La evidencia disponible para productos básicos en general indica que incluso modelos estructurales bien especificados no superan sistemáticamente una previsión de no-cambio cuando son evaluados en tiempo real (Kwas & Rubaszek, 2021), un patrón que González Casimiro (2009) sitúa en el marco más amplio de la dificultad de predecir series económicas con alta componente estocástica.

La literatura aplicada sobre pronóstico del precio del café ha crecido en años recientes, impulsada en parte por la proliferación de métodos de aprendizaje automático, aunque con heterogeneidad metodológica notable. Un antecedente directo de este trabajo es Milas et al. (2004), quienes evalúan modelos de corrección de error lineales y no lineales para el pronóstico de los precios spot de distintos tipos de café y concluyen que la ganancia sobre el paseo aleatorio es difícil de sostener fuera de muestra, documentando así tempranamente la resistencia del benchmark Naive en este mercado. Naveena et al. (2017) aplican modelos ARIMA e híbridos para el café arábica en India y concluyen que los modelos híbridos no superan sistemáticamente a los univariados en todos los horizontes evaluados. Deina et al. (2022) proponen máquinas de aprendizaje extremo (ELM) como alternativa a ARIMA y redes neuronales para el pronóstico de precios de café en Brasil, obteniendo mejoras en horizontes cortos pero sin evaluar la estabilidad de los resultados en múltiples ventanas fuera de muestra. Cardozo Rueda (2022) exploran redes neuronales artificiales para el caso colombiano, y Y. Chen (2024) y Hwase y Fofanah (2021) aplican múltiples algoritmos de aprendizaje automático —incluyendo bosques aleatorios y gradient boosting— para el merca-

do del café con énfasis en futuros y precios de bolsa. Para el mercado colombiano en particular, Díaz-Pinzón (2025) construye modelos de predicción del precio interno del café pergamino seco con proyecciones hasta 2028, y Pinto-Rodriguez et al. (2025) presentan una revisión sistemática de métodos para la estimación de precios de café de especialidad que identifica R^2, AIC y MSE como las métricas más frecuentes en la literatura, evidenciando la ausencia de un protocolo unificado de evaluación predictiva fuera de muestra. En el espacio más amplio del pronóstico de productos básicos agrícolas, Kipkoech et al. (2023) aplican ARIMA al pronóstico de arroz y trigo en Kenia con resultados comparables a los de los modelos más elaborados. Una lectura transversal de estos trabajos revela que la mayoría compara uno o dos enfoques sin incluir el espectro completo desde benchmarks ingenuos hasta combinaciones de pronósticos, y que pocos adoptan protocolos de validación fuera de muestra con múltiples particiones temporales que permitan evaluar la robustez del ordenamiento observado.

2.2. Evolución metodológica: de los modelos estadísticos al aprendizaje automático

El pronóstico de series temporales parte del supuesto de que los patrones históricos de una variable contienen información relevante para anticipar su comportamiento futuro (González Casimiro, 2009). Sobre esa premisa común se han desarrollado familias metodológicas con lógicas internas distintas, cuyas fortalezas y limitaciones relativas son especialmente relevantes en el contexto de series de productos básicos que exhiben integración de orden uno.

Los modelos ARIMA, sistematizados por Box y Jenkins (1976), constituyen el punto de partida natural de cualquier ejercicio comparativo de pronóstico univariado. Su fortaleza reside en la parsimonia: con un número reducido de parámetros, el ARIMA(p,d,q) captura la autocorrelación residual de la serie diferenciada y produce pronósticos asintóticamente óptimos bajo linealidad y gaussianidad. La extensión estacional SARIMA permite modelar patrones periódicos en series subanualeas. Para series de productos básicos con $d = 1$, el ARIMA impone una estructura coherente con la no estacionariedad observada y tiende a producir pronósticos conservadores —cercanos al paseo aleatorio con deriva— que resultan difíciles de superar de forma consistente en horizontes cortos (Hyndman & Athanasopoulos, 2021). Esta propiedad lo convierte en un

benchmark metodológico de hecho en evaluaciones comparativas, más que en un simple punto de partida. La selección automática de especificación por AIC, tal como se implementa en esta tesis, ha demostrado ser competitiva frente a procedimientos de selección más laboriosos (Hyndman & Athanasopoulos, 2021).

La familia ETS (*Error, Trend, Seasonal*), desarrollada dentro del marco unificado de modelos de espacio de estados (Hyndman & Athanasopoulos, 2021), agrupa las distintas variantes del suavizamiento exponencial bajo una estructura probabilística común que facilita la selección automática de especificación. Su ventaja relativa sobre ARIMA no es universal: en series sin tendencia determinista clara —como la del precio del café en el período analizado—, la selección automática tiende a favorecer especificaciones de suavizamiento simple, metodológicamente equivalentes a dar mayor peso a las observaciones recientes que a las lejanas sin imponer una estructura paramétrica rígida. La competencia M4 (Makridakis et al., 2020), con 100.000 series de distinta naturaleza, mostró que los métodos ETS y sus extensiones son sistemáticamente competitivos y difíciles de superar en horizontes cortos, lo que justifica su inclusión como modelo de referencia en esta tesis.

El modelo VAR, introducido por Sims (1980) como alternativa irrestricta a los modelos de ecuaciones simultáneas, añade una dimensión que ARIMA y ETS no contemplan: la posibilidad de capturar la correlación dinámica entre la variable objetivo y covariables potencialmente informativas. La referencia metodológica estándar para su estimación y análisis es Lütkepohl (2005). En series integradas de orden uno, el VAR se estima típicamente en primeras diferencias, y los pronósticos en nivel se recuperan mediante acumulación. La pregunta relevante para esta tesis es si la correlación dinámica entre el precio del café y variables como los precios del petróleo (WTI, Brent) y el índice climático ONI es suficientemente estable para generar ganancias predictivas netas fuera de muestra. La evidencia para productos básicos en general es mixta: la ganancia predictiva del VAR sobre modelos univariados depende de la fuerza y estabilidad de las correlaciones disponibles (Kwas & Rubaszek, 2021), y los beneficios se concentran principalmente en horizontes cortos donde la señal informativa de las covariables no ha sido diluida por la acumulación de incertidumbre.

Frente a los modelos paramétricos, los métodos de aprendizaje automático reformulan el pronóstico como un problema de regresión supervisada sobre una matriz de características cons-

truida a partir de rezagos de la variable objetivo y covariables exógenas. Los bosques aleatorios, propuestos por Breiman (2001), son un método de ensamble de árboles de decisión que introduce aleatoriedad en la selección de características a nivel de nodo, lo que reduce la correlación entre los árboles del ensamble y mejora la generalización fuera de muestra respecto a un árbol individual. Su atractivo teórico en el contexto del pronóstico reside en la capacidad de capturar interacciones no lineales entre predictores sin necesidad de especificar la forma funcional a priori. Para pronósticos multi-paso ($h > 1$), la implementación puede seguir estrategias directas o recursivas, con diferentes compromisos entre sesgo y varianza (Bojer, 2022). Sin embargo, esta flexibilidad tiene un costo que es especialmente pronunciado en series de productos básicos: cuando la señal predictiva es débil —como ocurre en series cercanas al paseo aleatorio—, la alta varianza propia de métodos no paramétricos no es compensada por reducciones de sesgo, y el resultado fuera de muestra puede ser inferior al de modelos parsimoniosos. Medeiros et al. (2021) documentaron que el aprendizaje automático mejora el pronóstico de variables macroeconómicas en entornos con muchos predictores informativos, pero los mismos autores advierten que la ventaja no está garantizada en contextos con señal débil. Ly et al. (2021) aplicaron redes LSTM —conceptualmente análogas al RF en su enfoque no paramétrico— al pronóstico de algodón y petróleo y concluyeron que no superan sistemáticamente a ARIMA fuera de muestra, aunque la combinación de ambos enfoques puede reducir el error. En el ámbito de productos básicos energéticos, Hamouda et al. (2025) muestran que modelos de inteligencia artificial con factores externos producen mejoras sobre métodos estadísticos, pero su ventaja depende críticamente de la disponibilidad de covariables con poder informativo real. Para productos básicos agrícolas con conjuntos de covariables acotados, la evidencia de Z. Chen et al. (2021) sugiere que esa condición no siempre se cumple. La consecuencia metodológica es directa: la inclusión del Random Forest en este estudio no persigue demostrar la superioridad del aprendizaje automático en este contexto específico —para lo que no existen garantías teóricas— sino documentar empíricamente su desempeño relativo bajo condiciones estándar de aplicación, contribuyendo así a delimitar las condiciones bajo las cuales aporta valor.

El contraste entre la riqueza teórica de estas metodologías y su desempeño empírico en condiciones de evaluación rigurosa es, precisamente, lo que hace necesario el tipo de ejercicio que esta tesis implementa. La estacionalidad, cuando existe, puede ser difícil de especificar cuando sus patrones no son constantes ni crecientes en el tiempo (Madrigal Espinoza, 2011), lo que añade

una fuente de incertidumbre que afecta diferencialmente a los modelos según cómo traten ese componente. Ningún modelo domina a los demás en todas las condiciones posibles, y la evaluación fuera de muestra bajo protocolos comparables es el único mecanismo disponible para establecer qué enfoque funciona mejor en un contexto empírico específico.

2.3. Combinación de pronósticos, evaluación comparativa y posicionamiento

La combinación de pronósticos tiene una raíz histórica anterior a su formalización econométrica. Wallis (2014) reexamina la competencia de estimación de peso descrita por Francis Galton en 1907 y muestra que, tras corregir inconsistencias menores en los datos, el promedio de 787 estimaciones individuales coincidió exactamente con el peso observado del buey. El episodio constituye un antecedente temprano tanto de las competencias de pronóstico como de la "sabiduría de las multitudes": la agregación de juicios individuales puede producir una señal más precisa que muchas estimaciones consideradas de forma aislada. La contribución moderna de Bates y Granger (1969) fue formalizar esa intuición en términos econométricos, demostrando que, bajo condiciones generales, una combinación lineal ponderada puede reducir el error cuadrático medio respecto a cualquier pronóstico componente. La intuición es que la combinación diversifica los errores de especificación de los modelos individuales: si dos modelos cometen errores negativamente correlacionados, su promedio tiende a cancelar parte del ruido. La literatura posterior confirmó esta lógica y, al mismo tiempo, produjo uno de los hallazgos más robustos de la econometría aplicada: el promedio simple es difícil de superar por esquemas de ponderación más sofisticados, resultado conocido como el *forecast combination puzzle* (Timmermann, 2006). Wang et al. (2023), en una revisión de más de cincuenta años de literatura sobre el tema, confirman que la ventaja del promedio simple sobre métodos más elaborados persiste en una amplia variedad de contextos.

Sin embargo, la validez de ese resultado depende críticamente de la calidad del conjunto de modelos combinados. Genre et al. (2013) muestran que la inclusión de pronósticos de baja calidad en la combinación puede degradar su desempeño hasta el punto de anular las ventajas de la diversificación. La combinación ponderada por el inverso del RMSE acumulado —esquema implementado en esta tesis— responde precisamente a esta preocupación: asigna mayor peso a los

modelos con menor error histórico, atenuando la influencia de los componentes de peor desempeño. La evidencia empírica sobre la efectividad de este mecanismo en aplicaciones latinoamericanas de pronóstico macroeconómico sugiere que la combinación no siempre supera al mejor modelo individual cuando el conjunto de partida incluye modelos de calidad muy heterogénea (Humérez Quiroz, 2012). Este resultado anticipa una tensión que esta tesis documenta directamente: si el pool de modelos incluye tanto `VAR` (OWA $\approx$ 0,937) como `NaiveS` (OWA $\approx$ 3,56), la combinación simple estará dominada por el peor componente, y la ponderada deberá asumir toda la carga de corrección.

El marco de evaluación que permite articular comparaciones de este tipo de forma rigurosa proviene de las competencias M de pronóstico. La competencia M4 (Makridakis et al., 2020), con 100.000 series y 61 métodos, estableció el OWA (*Overall Weighted Average*) —el promedio entre el cociente sMAPE y el cociente MASE respecto a un benchmark de referencia— como criterio primario de clasificación. La elección del OWA no es arbitraria: al expresar el desempeño de cada modelo en relación con el benchmark, produce una medida directamente interpretable como superioridad o inferioridad relativa, independientemente de la escala o las unidades de la serie. El MASE, propuesto por Hyndman y Koehler (2006) como medida escalada y robusta frente a series con valores cercanos a cero, y el sMAPE, como indicador de error porcentual simétrico, son preferibles al MAPE convencional, cuya degeneración en situaciones comunes ha sido documentada por los mismos autores. La convención de las competencias M provee así un lenguaje común que hace posible la comparación con una comunidad de práctica internacional y que esta tesis adopta explícitamente.

Un hallazgo sistemático de las competencias M es que los métodos simples —incluyendo el promedio aritmético y los benchmarks ingenuos— tienden a ser muy competitivos en horizontes cortos, y que los enfoques más elaborados rara vez los superan de manera consistente sobre muestras amplias y heterogéneas de series (Makridakis et al., 2020). Es importante matizar este resultado: está establecido sobre colecciones de miles de series de distinta naturaleza, frecuencia y dominio, y no se traslada automáticamente al caso de una única serie mensual de precio de un producto básico. Castle et al. (2021), en un análisis de los principios derivados de las competencias de pronóstico, subrayan que el valor informativo de la evaluación fuera de muestra depende tanto del protocolo adoptado como de la naturaleza de la serie analizada; en presencia de no estaciona-

riedad y choques intermitentes, privilegiar la evidencia fuera de muestra sobre el ajuste dentro de la muestra es una prescripción metodológica de primer orden. En esa dirección, Tashman (2000) argumenta que el uso de múltiples períodos de prueba es preferible a una única partición, porque reduce la sensibilidad de los resultados a eventos particulares de cualquier subperíodo. La evaluación con múltiples ventanas temporales que esta tesis implementa —18, 24 y 36 meses de observaciones de prueba— sigue directamente esta recomendación.

Tres brechas de la literatura justifican el diseño de esta tesis. Primera, la mayoría de los estudios de pronóstico del precio del café evalúan uno o dos enfoques sin cubrir el espectro completo desde benchmarks ingenuos hasta combinaciones de pronósticos, lo que impide establecer qué mejoras son genuinas y cuáles son artefactos de la ausencia de comparación con alternativas simples. Segunda, son escasos los trabajos que adoptan esquemas de validación fuera de muestra con múltiples particiones temporales comparables al *rolling-origin expanding window*, que es el estándar metodológico para producir estimaciones robustas de la precisión predictiva (Hyndman & Athanasopoulos, 2021; Tashman, 2000). Tercera, la adopción de métricas basadas en la convención M4 —OWA, sMAPE, MASE— es prácticamente ausente en la literatura aplicada al café, lo que dificulta la comparación directa con la evidencia más amplia sobre pronóstico de series de productos básicos. Frente a estas brechas, la contribución de esta tesis es metodológica y empírica: ofrecer una evaluación exhaustiva, reproducible y alineada con estándares internacionales de qué modelos funcionan mejor para el pronóstico mensual del precio externo del café colombiano, bajo el protocolo más riguroso que los datos disponibles permiten implementar.

3. Metodología

3.1. Diseño empírico y modelos candidatos

El objetivo de esta investigación no es explicar causalmente la formación del precio externo del café, sino evaluar la capacidad predictiva de distintos enfoques metodológicos cuando todos compiten bajo las mismas condiciones de información y protocolo de validación. El diseño sigue la lógica de las *forecasting competitions* (Makridakis et al., 2020), adaptada al caso de una serie objetivo mensual: nueve modelos se evalúan sobre la misma muestra, con el mismo esquema de

entrenamiento y prueba, y sus resultados se comparan mediante métricas uniformes. Esta estructura de "laboratorio controlado" es la única que permite atribuir las diferencias de precisión al método y no al diseño de evaluación.

La variable objetivo es el precio externo mensual del café, denotada y_t, aproximada por el contrato continuo KC1 de Bloomberg como proxy del benchmark internacional del café arábica. El período analizado abarca enero de 2000 a marzo de 2026, con 315 observaciones mensuales. Esta elección no confunde el precio del futuro con el precio físico de exportación; su propósito es adoptar una referencia internacional líquida, trazable y empíricamente próxima al mercado físico relevante para Colombia. La ICO muestra que, para el grupo *Colombian Milds*, la relación spot–futuros frente al mercado de Nueva York alcanza $R^2 = 0{,}9212$ (International Coffee Organization, 2011), de modo que la mayor parte de la variación del precio físico se explica por el benchmark de futuros. La frecuencia mensual es la apropiada para armonizar la serie objetivo con las covariables disponibles —precios del petróleo WTI y Brent e índice climático ONI— y para producir evaluaciones fuera de muestra con suficiente longitud histórica en cada ventana de entrenamiento.

Los nueve modelos candidatos cumplen roles metodológicos diferenciados en el experimento. Los tres benchmarks ingenuos —Naive 1 (paseo aleatorio puro), Naive S (pronóstico estacional mensual) y Naive 2 (paseo aleatorio desestacionalizado)— establecen la línea base estricta que cualquier modelo más elaborado debe superar para justificar su complejidad adicional. Naive 2 actúa específicamente como benchmark de referencia para el cálculo del OWA, en línea con la convención de las competencias M para series mensuales (Makridakis et al., 2020). La inclusión explícita de estos tres benchmarks es una prescripción metodológica de primer orden: sin una línea base exigente, una comparación entre modelos complejos no permite establecer si alguno genera valor predictivo real o simplemente replica el comportamiento de la serie más reciente.

ARIMA y ETS cumplen el rol de baseline estadístico univariado. No requieren covariables y producen pronósticos a partir exclusivamente de la historia pasada de y_t. Su presencia en el experimento permite aislar cuánto de la ganancia predictiva del VAR o del Random Forest se debe a la información adicional de las covariables y cuánto sería obtenible con la estructura temporal de la propia serie. La especificación de ambos modelos —ARIMA(p,d,q) con extensión estacional y la familia ETS sobre el espacio de modelos de espacio de estados— se selecciona por AIC sobre cada ventana de entrenamiento, lo que garantiza que la selección respeta el calendario informacional y

no introduce sesgos de mirada hacia adelante.

El VAR multivariado incorpora como covariables las series de WTI, Brent y ONI junto con y_t, y su función en el experimento es precisamente cuantificar el valor predictivo neto de la correlación dinámica entre el precio del café y esas variables macroeconómicas y climáticas. Si el VAR supera a los univariados, la evidencia sugiere que esas covariables contienen información predictiva incremental que los modelos de la historia propia de y_t no pueden capturar; si no los supera, el resultado indica que dicha correlación no es suficientemente estable fuera de muestra para generar ganancias netas. El orden de rezagos del VAR se selecciona también por AIC sobre la ventana de entrenamiento inicial de cada partición.

El Random Forest constituye el test no paramétrico del experimento. Su inclusión no parte de una presunción de superioridad, sino de una pregunta empírica: ¿puede un método que captura interacciones no lineales entre rezagos y covariables, sin imponer estructura funcional, mejorar el pronóstico fuera de muestra en esta serie? La implementación utiliza una configuración estándar —500 árboles, 12 rezagos de y_t y 4 rezagos de las covariables— sin búsqueda de hiperparámetros, lo que refleja el nivel de esfuerzo de calibración que sería razonable en una aplicación práctica sin datos adicionales para una validación interna más elaborada (Bojer, 2022).

Las dos estrategias de combinación de pronósticos —promedio simple y promedio ponderado por el inverso del RMSE acumulado— representan el nivel siguiente de complejidad: en lugar de seleccionar un único modelo, agregan las predicciones del conjunto completo de siete modelos base. El promedio simple evalúa si una agregación sin calibración captura señales complementarias entre modelos; el ponderado introduce una regla de actualización dinámica que asigna mayor peso a los modelos con menor error histórico acumulado al momento de cada pronóstico. Ambas combinaciones respetan estrictamente el calendario informacional: los pesos en cada origen τ se calculan con errores observados hasta $\tau - 1$, sin acceso a información futura (Timmermann, 2006).

3.2. Protocolo de evaluación fuera de muestra y métricas

El principio organizador del protocolo de evaluación es evitar cualquier forma de fuga de información que haga los resultados no reproducibles en un contexto operativo real. En el pronóstico de series temporales, la aleatorización del conjunto de prueba —procedimiento estándar en

aprendizaje automático transversal— viola la estructura temporal de los datos e introduce sesgos optimistas que sobreestiman el desempeño fuera de muestra (Hyndman & Athanasopoulos, 2021). El diseño adoptado en esta tesis respeta estrictamente el orden temporal: para cada origen τ, todos los modelos se estiman con las observaciones $\{y_1, \ldots, y_\tau\}$ y producen pronósticos $\hat{y}_{\tau+h|\tau}$ para horizontes $h \in \{1, 3\}$. El origen avanza secuencialmente de mes en mes dentro de la ventana de prueba, acumulando información —*rolling-origin expanding window*— de modo que la ventana de entrenamiento crece a medida que se avanza en el tiempo. Esta estructura reproduce fielmente la situación de un analista que actualiza su modelo cada mes con los datos disponibles, sin recurrir jamás a observaciones futuras.

La robustez de los resultados frente a la elección del período de prueba se asegura mediante tres particiones temporales independientes. La primera reserva los últimos 18 meses como observaciones de prueba ($n = 18$), la segunda los últimos 24 meses ($n = 24$) y la tercera los últimos 36 meses ($n = 36$). Cada partición define su propio conjunto de orígenes de pronóstico y su propia ventana de entrenamiento inicial. La adopción de múltiples particiones sigue la recomendación de Tashman (2000), quien argumenta que un único período de prueba puede ser no representativo del comportamiento de largo plazo de la serie, ya que los resultados de cualquier modelo son sensibles a las condiciones de volatilidad y estructura del subperíodo evaluado; usar múltiples ventanas "desensibiliza las medidas de error a eventos especiales en cualquier origen único". La consistencia del ordenamiento de los modelos a través de las tres particiones es, por tanto, el criterio de robustez más exigente disponible en este diseño, más informativo que cualquier resultado puntual.

Se evalúan dos horizontes de pronóstico: $h = 1$ (un mes adelante) y $h = 3$ (tres meses adelante). La elección de estos horizontes responde a su relevancia práctica para la gestión del riesgo de precio en el sector cafetero —donde las decisiones de cobertura y financiamiento se toman típicamente en el corto plazo— y a la posibilidad de comparar si la ventaja relativa de los modelos multivariados se mantiene o se erosiona al ampliar el horizonte. Los pronósticos de $h = 3$ se generan directamente —estrategia *direct forecasting*— para cada origen, lo que evita la propagación de errores propia de la estrategia recursiva pero impone una función de pérdida específica a cada horizonte.

La comparación de precisión predictiva se articula en torno a tres métricas complementarias cuya elección responde a criterios metodológicos explícitos, no a convención arbitraria. Sea N el

número de pronósticos evaluados, y_i el valor observado y $\hat{y}_i$ el pronóstico en la posición i. El sMAPE (*Symmetric Mean Absolute Percentage Error*) se define como:

$$\text{sMAPE} = \frac{200}{N} \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|}, \tag{1}$$

y expresa el error como porcentaje simétrico del valor observado, facilitando la comparación entre horizontes y períodos sin depender de la escala absoluta de la serie. El MASE (*Mean Absolute Scaled Error*), propuesto por Hyndman y Koehler (2006) como alternativa robusta al MAPE convencional y libre del riesgo de denominadores cercanos a cero, se define como:

$$\text{MASE} = \frac{\dfrac{1}{N} \displaystyle\sum_{i=1}^{N} |y_i - \hat{y}_i|}{\dfrac{1}{T-m} \displaystyle\sum_{t=m+1}^{T} |y_t - y_{t-m}|}, \tag{2}$$

donde $m = 12$ es la periodicidad estacional mensual y el denominador es el error absoluto medio del pronóstico estacional ingenuo calculado sobre la muestra de entrenamiento. Un MASE < 1 indica que el modelo supera al pronóstico estacional ingenuo en la escala del entrenamiento; un MASE > 1 indica lo contrario. La virtud del MASE en este contexto es que su denominador es fijo para cada partición, lo que lo hace comparable entre modelos y entre horizontes dentro del mismo experimento.

La métrica principal de comparación es el OWA (*Overall Weighted Average*), definida como el promedio aritmético de los cocientes sMAPE y MASE respecto al benchmark Naive 2:

$$\text{OWA} = \frac{1}{2} \left(\frac{\text{sMAPE}}{\text{sMAPE}_{\text{Naive2}}} + \frac{\text{MASE}}{\text{MASE}_{\text{Naive2}}} \right). \tag{3}$$

Esta métrica fue adoptada como criterio de clasificación oficial en la competencia M4 (Makridakis et al., 2020) por dos propiedades que la hacen especialmente adecuada para este estudio. Primero, al expresar el desempeño de cada modelo relativo al benchmark, el OWA es directamente interpretable: un valor inferior a uno indica que el modelo supera a Naive 2; un valor superior a uno indica inferioridad. Segundo, al combinar sMAPE y MASE, el OWA agrega señales de precisión desde dos perspectivas complementarias —relativa y escalada— reduciendo la dependencia respecto a la

elección de una única medida. Como métricas auxiliares, se reportan también el RMSE y el MAE en escala original, cuya función es permitir una lectura de la magnitud económica del error sin sustituir el criterio comparativo principal.

La significancia estadística de las diferencias de precisión se evalúa mediante el test de Diebold y Mariano (1995), que contrasta la hipótesis nula de igual capacidad predictiva entre dos modelos a partir de la serie de diferencias de pérdidas cuadráticas $d_t = L(e_{1t}) - L(e_{2t})$. Dado que el test original presenta distorsiones de tamaño pronunciadas en muestras pequeñas, se aplica la corrección de muestra finita propuesta por Harvey et al. (1997), que escala el estadístico DM por un factor de ajuste y lo contrasta contra una distribución $t(n-1)$ en lugar de la normal estándar. La varianza del diferencial de pérdidas se estima mediante un estimador HAC con $h-1$ rezagos de Newey-West, apropiado para errores de pronóstico multi-paso. El test se aplica para cada modelo frente al benchmark Naive 2, bajo la hipótesis alternativa unilateral de que el modelo evaluado incurre en menor pérdida esperada, y sus resultados se interpretan como evidencia complementaria al criterio OWA, no como el criterio principal de decisión. Esta jerarquía es metodológicamente correcta: con $n \leq 36$ observaciones de error disponibles por partición, el poder estadístico del test es insuficiente para detectar mejoras de magnitud moderada (Coroneo & Iacone, 2020), de modo que la ausencia de significancia formal no equivale a ausencia de diferencias reales entre modelos.

4. Datos

El análisis empírico descansa sobre un conjunto de datos maestro mensual construido específicamente para este proyecto, que integra en una sola base canónica la variable objetivo y las covariables potencialmente informativas bajo un calendario temporal común. El criterio de consolidación es funcional: cada fila representa un mes calendario, cada columna una variable, y el formato tabular ancho garantiza que todos los modelos reciban exactamente la misma estructura de información, independientemente de si son univariados o multivariados. La base abarca 315 observaciones mensuales con fechas de cierre de mes, desde enero de 2000 hasta marzo de 2026.

La variable objetivo es el precio externo mensual del café, aproximado por el contrato continuo KC1 de Bloomberg como proxy del benchmark internacional del café arábica. Esta definición exige una precisión conceptual importante: KC1 no es el precio spot ni el precio efectivo

de exportación colombiano, sino la referencia de futuros que organiza la formación de precios en el mercado externo. La justificación empírica para usarlo como proxy es sólida. La ICO reporta, para el grupo *Colombian Milds*, una regresión lineal entre precios físicos y futuros de Nueva York con $R^2 = 0{,}9212$ (International Coffee Organization, 2011), lo que implica que la mayor parte de la variación del precio spot del grupo es explicada por el benchmark de futuros arábica. En consecuencia, KC1 ofrece una aproximación operativamente defendible del precio externo relevante para Colombia cuando el objetivo es comparar modelos de pronóstico bajo una referencia internacional homogénea y replicable.

En una etapa preliminar se consideró la referencia COFECMNY Index difundida en Bloomberg por la Federación Nacional de Cafeteros, conceptualmente más cercana al precio de exportación colombiano. Sin embargo, su publicación regular se detiene en agosto de 2021, lo que impide sostener una muestra homogénea hasta marzo de 2026. Por esta razón, la corrida principal adopta KC1 como serie objetivo base. La versión consolidada en la base maestra no registra valores faltantes en ningún punto de la muestra analizada, condición necesaria para sostener un protocolo de evaluación con ventana expandible sin interrupciones en el calendario de entrenamiento. Como transformación complementaria, el dataset incluye el logaritmo natural del precio, disponible para especificaciones donde la estabilización de varianza sea conveniente, aunque los modelos de la corrida principal operan en nivel nominal. La Figura 1 resume la trayectoria de la serie objetivo en toda la muestra, incluyendo el ciclo de máximos históricos observado entre 2024 y 2025.

La elección de la frecuencia mensual no es una decisión puramente operativa. Frente a observaciones diarias o semanales, que capturan con mayor intensidad ruido de negociación, ajustes especulativos y movimientos transitorios de liquidez, la agregación mensual filtra parte de esa variabilidad de muy corto plazo y permite concentrar el ejercicio en relaciones más estables con fundamentos macroeconómicos y climáticos. Esta frecuencia también está determinada por la naturaleza de las covariables relevantes: indicadores como el Índice Oceánico Niño (ONI), central para aproximar choques climáticos asociados a la oferta cafetera, se publican en periodicidad mensual y no admiten una correspondencia diaria o semanal sin introducir supuestos de interpolación difíciles de defender. Finalmente, el horizonte mensual es coherente con el uso práctico de los pronósticos en el sector cafetero colombiano, donde las decisiones de cobertura, planeación presupuestal, financiamiento y seguimiento de política sectorial suelen evaluarse en ventanas mensuales

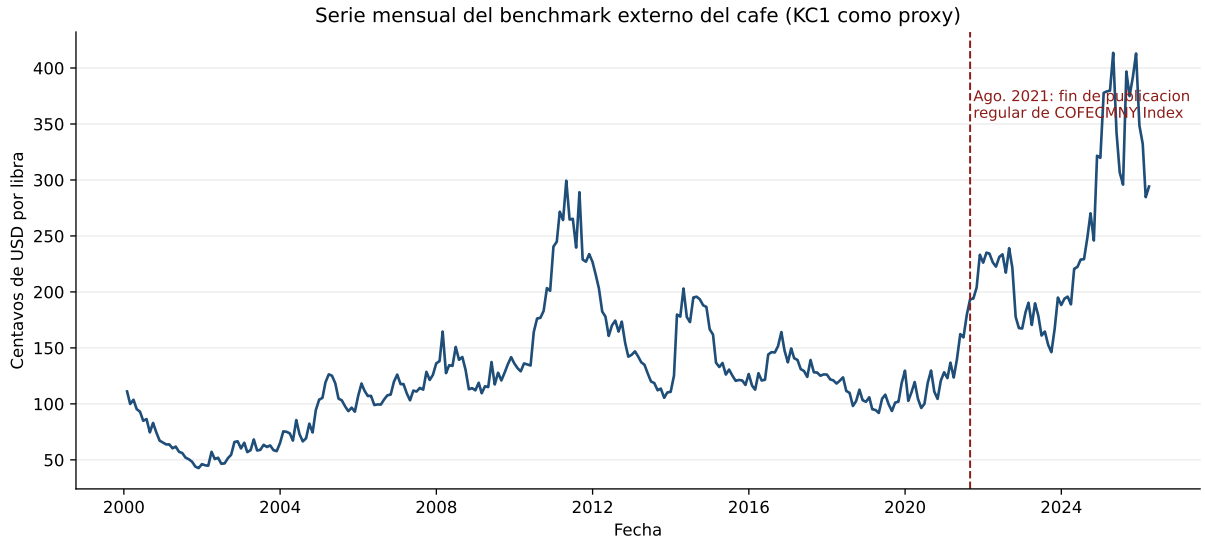

Figura 1: Serie objetivo mensual utilizada en la tesis: contrato continuo KC1 como proxy del precio externo del café. La línea vertical punteada marca agosto de 2021, fecha a partir de la cual deja de publicarse de forma regular la referencia COFECMNY Index utilizada en la etapa preliminar.

Nota. Elaboración propia con datos de Bloomberg (contrato continuo KC1, enero 2000 – marzo 2026).

o trimestrales, más que en fluctuaciones interdiarias propias del *trading*.

Para aislar el valor predictivo de la información macroeconómica y financiera, se incorporan tres covariables del entorno internacional del mercado de productos básicos: el precio del petróleo crudo West Texas Intermediate (WTI), el precio del petróleo Brent y el índice amplio del dólar estadounidense (DXY). Los precios del petróleo actúan como indicadores del ciclo económico global y de los costos de transporte y producción agrícola; el índice del dólar captura el efecto de la moneda de denominación sobre los precios internacionales de los productos básicos cotizados en dólares. La literatura reciente subraya además que los precios del WTI y el Brent absorben rápidamente choques globales de alta magnitud —incluyendo los vinculados a la pandemia de COVID-19 y al conflicto Rusia-Ucrania— y que la información contenida en sus revisiones de pronóstico es relevante para la evaluación del estado del ciclo económico global (Iregui et al., 2025), lo que refuerza la justificación de su inclusión como covariables informativas en el modelo VAR. WTI y Brent presentan cada uno un único valor faltante en toda la muestra, mientras que el DXY concentra 74 valores faltantes en las primeras observaciones del período, limitación que determina que esta variable no se utilice como covariable activa en los modelos de la corrida principal, aunque permanece en la base para análisis exploratorios. WTI y Brent, disponibles en

314 de las 315 observaciones, constituyen el núcleo del bloque macroeconómico efectivamente incorporado al VAR y al Random Forest.

La dimensión climática se representa mediante el Índice Oceánico Niño (ONI), indicador del fenómeno El Niño–Oscilación del Sur (ENSO) publicado mensualmente por la Administración Nacional Oceánica y Atmosférica de los Estados Unidos (NOAA). El ONI registra solo dos valores faltantes en toda la muestra y está disponible en 313 observaciones, cobertura suficiente para su incorporación sin imputaciones en el período de evaluación fuera de muestra. Su inclusión responde a la evidencia de que los ciclos ENSO afectan la producción cafetera en Brasil y Colombia a través de la precipitación y la temperatura, lo que genera correlaciones entre el estado climático contemporáneo y la dinámica futura de la oferta y del precio (Useche Mahecha, 2024).

El dataset incorpora adicionalmente cinco series de Google Trends relacionadas con el mercado del café: volumen de búsqueda asociado al precio del café, a los futuros de café, al café arábica y a eventos climáticos extremos —heladas y sequías— en Brasil. Estas series aproximan la atención digital como señal de anticipación informacional en el mercado. Su cobertura comienza en 2004, acumulando 48 valores faltantes en el tramo inicial de la muestra, lo que restringe su utilidad en la corrida principal, donde el período de entrenamiento se extiende desde el año 2000. Por esta razón, su rol en la tesis es exploratorio y no forman parte de los modelos cuyo desempeño se reporta en el capítulo de resultados. El criterio de exclusión no es conceptual sino operativo: su incorporación reduciría la longitud efectiva de la ventana de entrenamiento en un 15 % aproximadamente, alterando la comparabilidad del experimento entre modelos.

En cuanto a la gestión de los faltantes, la tesis adopta una postura de no imputación: ningún valor ausente fue reemplazado por estimaciones o interpolaciones. Los modelos utilizan únicamente la información efectivamente observada en cada fecha de corte, lo que garantiza que el experimento replica de manera fiel las condiciones de un analista operando en tiempo real. Esta elección tiene como consecuencia que las covariables activas —WTI, Brent y ONI— se seleccionan también por su cobertura casi completa a lo largo de toda la muestra, minimizando la pérdida de orígenes de pronóstico en el tramo de evaluación.

La base maestro incorpora directamente las etiquetas de partición temporal que codifican, para cada observación, si pertenece al tramo de entrenamiento o al de prueba. Las tres particiones que estructuran el experimento —18, 24 y 36 meses de observaciones de prueba— están definidas

como columnas binarias en el dataset, de modo que el pipeline experimental las lee directamente sin recalcularlas en tiempo de ejecución. En la partición de 18 meses, las observaciones de prueba van de octubre de 2024 a marzo de 2026; en la de 24 meses, de abril de 2024 a marzo de 2026; y en la de 36 meses, de abril de 2023 a marzo de 2026. Esta estructura garantiza que todos los modelos sean evaluados sobre exactamente los mismos tramos cronológicos, eliminando cualquier fuente de variación en el protocolo de comparación que pudiera atribuirse al diseño de la partición y no al modelo.

5. Implementación computacional

La implementación del experimento se estructuró como un pipeline modular en Python, cuyo principio de diseño central es la reproducibilidad total: dado el dataset maestro y el código, cualquier investigador puede regenerar exactamente los mismos pronósticos, métricas y tablas reportados en este documento ejecutando una única instrucción. Este principio no es un detalle técnico secundario, sino una condición metodológica de primer orden en la evaluación empírica de modelos de pronóstico, donde la transparencia del protocolo es tan relevante como los resultados mismos (Castle et al., 2021).

El pipeline se organiza en módulos con responsabilidades bien delimitadas. Un módulo de configuración centraliza todas las constantes del experimento —rutas de archivos, nombre de la variable objetivo, periodicidad estacional $m = 12$, nombres de los splits y semilla aleatoria— de modo que cualquier cambio en los parámetros del diseño se propaga de forma consistente a todos los componentes sin modificaciones distribuidas. Un módulo de carga normaliza el índice temporal del dataset maestro, verifica la presencia de las columnas requeridas —`coffee_price`, las columnas de partición y las covariables activas— y rechaza la ejecución si el contrato mínimo de datos no se cumple. Esta verificación anticipada evita que errores de datos silenciosos se propaguen hasta las métricas finales. Los módulos de estimación —benchmarks ingenuos, modelos univariados, VAR y Random Forest— están implementados como funciones sin estado lateral: reciben como entrada únicamente la ventana de entrenamiento disponible hasta el origen τ y devuelven los pronósticos correspondientes, sin acceso a ninguna observación posterior. El módulo de evaluación calcula las métricas para cada combinación de split, horizonte y modelo, y un script de ejecución

coordina el flujo completo, produce los archivos de salida y genera las tablas y figuras del capítulo de resultados.

El núcleo del experimento es el bucle de rolling-origin. Para cada partición temporal y cada horizonte $h \in \{1,3\}$, el bucle itera sobre todas las posiciones del tramo de prueba, construye la ventana de entrenamiento expandible $\{y_1,\ldots,y_{\tau-h}\}$ en el origen τ, llama a cada uno de los nueve modelos con esa ventana y registra el pronóstico $\hat{y}_{\tau|\tau-h}$ junto con el valor observado y_τ. La restricción de que la ventana de entrenamiento llega solo hasta $\tau-h$ —y no hasta $\tau-1$— es la que garantiza que el pronóstico de h pasos adelante no contiene ninguna observación del período objetivo, eliminando cualquier forma de filtración de información futura. Esta distinción es especialmente importante para $h=3$: el modelo que pronostica la observación de marzo de 2025, por ejemplo, solo tiene acceso a datos hasta diciembre de 2024, no a enero ni febrero de 2025.

La estandarización de la selección de especificaciones responde a un principio de honestidad experimental: los hiperparámetros de cada modelo deben determinarse exclusivamente con información disponible en la fecha de corte, sin mirar el tramo de prueba. Para ARIMA, ETS y VAR, la especificación se selecciona por AIC sobre la muestra de entrenamiento inicial de cada partición —es decir, la ventana disponible antes del primer origen de prueba— y se mantiene fija a lo largo de todo el rolling-origin de esa partición. Esta decisión tiene una justificación metodológica clara: re-seleccionar la especificación en cada origen sería computacionalmente costoso sin aportar evidencia adicional sobre el problema de interés, y la estabilidad de la especificación a lo largo del tramo de prueba facilita la interpretación de los resultados. El ARIMA resultante en las tres particiones es SARIMA$(1,1,1)(0,1,1)_{12}$ —estructura que captura la dinámica autorregresiva de primer orden y la corrección de medias móviles tanto en la componente regular como en la estacional— y el ETS seleccionado es la especificación de suavizamiento exponencial simple, coherente con la dinámica cercana al paseo aleatorio que exhibe la serie. El VAR se estima en primeras diferencias con orden $p=2$, seleccionado por AIC, y los pronósticos en nivel se recuperan por acumulación de las diferencias pronosticadas. Una guarda de estabilidad verifica en cada origen que las raíces del polinomio característico del VAR se encuentren dentro del círculo unitario; si esta condición no se cumple, el pronóstico revierte a Naive 1. En la corrida final, esta guarda no fue activada en ningún origen de ninguna partición.

El Random Forest opera con una configuración de hiperparámetros fija para toda la corrida: 500 árboles, 12 rezagos de la variable objetivo y 4 rezagos de cada una de las tres covariables como características de entrada, para un total de 24 predictores por observación. La elección de no realizar búsqueda de hiperparámetros no es una limitación inadvertida sino una decisión de diseño deliberada: refleja el nivel de esfuerzo de calibración que sería razonable en una aplicación práctica donde no se dispone de un conjunto de validación interno independiente del tramo de prueba, y hace el experimento comparable con aplicaciones estándar de la literatura (Bojer, 2022). Para $h = 3$, el RF utiliza una estrategia recursiva: el pronóstico de $t+1$ se incorpora al historial de la variable objetivo, las covariables futuras se proyectan con su último valor observado, y el proceso se repite hasta obtener el pronóstico de $t+3$. Este enfoque acumula el error de pronóstico entre pasos, lo que es una limitación reconocida frente a la estrategia directa, pero mantiene coherencia con el esquema de características fijas utilizado para $h = 1$.

Las combinaciones de pronósticos se calculan dentro del mismo bucle de rolling-origin, lo que garantiza que los pesos del promedio ponderado se actualicen utilizando únicamente los errores acumulados hasta el origen previo, sin acceso a los resultados futuros del tramo de prueba. En el primer origen de cada partición, la combinación ponderada parte de pesos iguales entre los siete modelos base; a partir del segundo origen, los pesos son proporcionales al inverso del RMSE acumulado de cada modelo en los orígenes anteriores. Esta actualización recursiva implementa una forma sencilla de selección adaptativa que no requiere un conjunto de validación separado y es completamente trazable.

Al finalizar la ejecución, el pipeline escribe los resultados en archivos CSV estructurados: pronósticos por partición y horizonte, métricas por partición y el resumen agregado entre las tres particiones. Scripts de postprocesamiento independientes leen únicamente estos archivos de salida —sin acceso al código de estimación— para producir las tablas en formato LaTeX y las figuras del capítulo de resultados. Esta separación entre la generación de resultados y su presentación garantiza que las tablas y figuras del documento reflejen siempre la última corrida del experimento, sin posibilidad de discrepancias entre el texto y los datos subyacentes.

6. Resultados

6.1. Desempeño predictivo comparado: $h = 1$ frente a $h = 3$

Las especificaciones seleccionadas por AIC sobre la ventana de entrenamiento inicial de cada partición fueron idénticas en las tres: SARIMA$(1,1,1)(0,1,1)_{12}$ para `ARIMA`, suavizamiento exponencial simple para `ETS` y VAR(2) en primeras diferencias para `VAR`. Estas especificaciones se detallan en el Apéndice A. La estabilidad de las especificaciones a lo largo del tramo de prueba —y su convergencia independientemente de la longitud de la ventana inicial— es en sí misma un resultado: sugiere que la estructura dinámica dominante de la serie es robusta al período de estimación y no requiere re-selección en cada origen.

Para el horizonte de un mes (Tabla 1 y Figura 2), el `VAR` y `Naive1` se sitúan prácticamente empatados en el primer lugar con OWA promedio de 0,937 y 0,938 respectivamente, representando una mejora de aproximadamente 6,3 % sobre el benchmark `Naive2`. `ARIMA` y `ETS` los siguen con OWA de 0,956 y 0,957, de modo que los cuatro primeros modelos forman un bloque compacto con un rango interior de apenas 0,020 puntos. Lo más relevante de este resultado no es la magnitud puntual de las diferencias, sino su consistencia: los cuatro modelos superan a `Naive2` en las tres particiones evaluadas sin ninguna excepción, incluyendo la partición de 36 meses, donde el rango entre el primero y el cuarto se comprime aún más (0,012 puntos), indicando que con mayor cantidad de orígenes de prueba las diferencias entre métodos tienden a nivelarse pero el ordenamiento cualitativo se mantiene.

En el horizonte de tres meses, la jerarquía se reorganiza y los márgenes se estrechan (Tabla 2 y Figura 3). `Naive1` pasa a liderar con OWA 0,973, seguido de `VAR` (0,991) y `ETS` (0,992), mientras que `ARIMA` queda prácticamente en la frontera del benchmark con OWA 1,003. La contracción de márgenes respecto a $h = 1$ es sustancial: la mejor mejora relativa cae de 6,3 % a 2,7 %. La heterogeneidad entre la partición de 18 meses y las dos restantes merece atención: en la ventana más corta, que coincide con el período de máxima volatilidad histórica de la serie (octubre 2024 – marzo 2026), solo `Naive1` supera al benchmark en $h = 3$, mientras que en las particiones de 24 y 36 meses los cuatro modelos estadísticos lo superan. Esta sensibilidad a las condiciones del subperíodo evaluado es, precisamente, el argumento que motiva el uso de tres ventanas independientes siguiendo la recomendación de Tashman (2000): ninguna de las tres ventanas aisladas captura el

Cuadro 1: Comparación de modelos para $h = 1$, promedio entre las tres particiones temporales (18, 24 y 36 meses de prueba). Ordenado por OWA ascendente. Referencia: OWA = 1 corresponde a Naive 2.

#	Modelo	sMAPE (%)	MASE	OWA
1	VAR	9,1105	0,8097	0,9373
2	Naive 1	9,1183	0,8110	0,9376
3	ARIMA	9,2739	0,8286	0,9555
4	ETS	9,3036	0,8286	0,9570
5	Naive 2	9,7861	0,8587	1,0000
6	Comb. Ponderada	10,0904	0,8953	1,0365
7	Comb. Simple	11,6599	1,0240	1,1924
8	RF	17,0914	1,4716	1,7222
9	Naive S	37,0023	2,8699	3,5614

Nota. Elaboración propia. Valores calculados con el pipeline de evaluación fuera de muestra (*rolling-origin*).

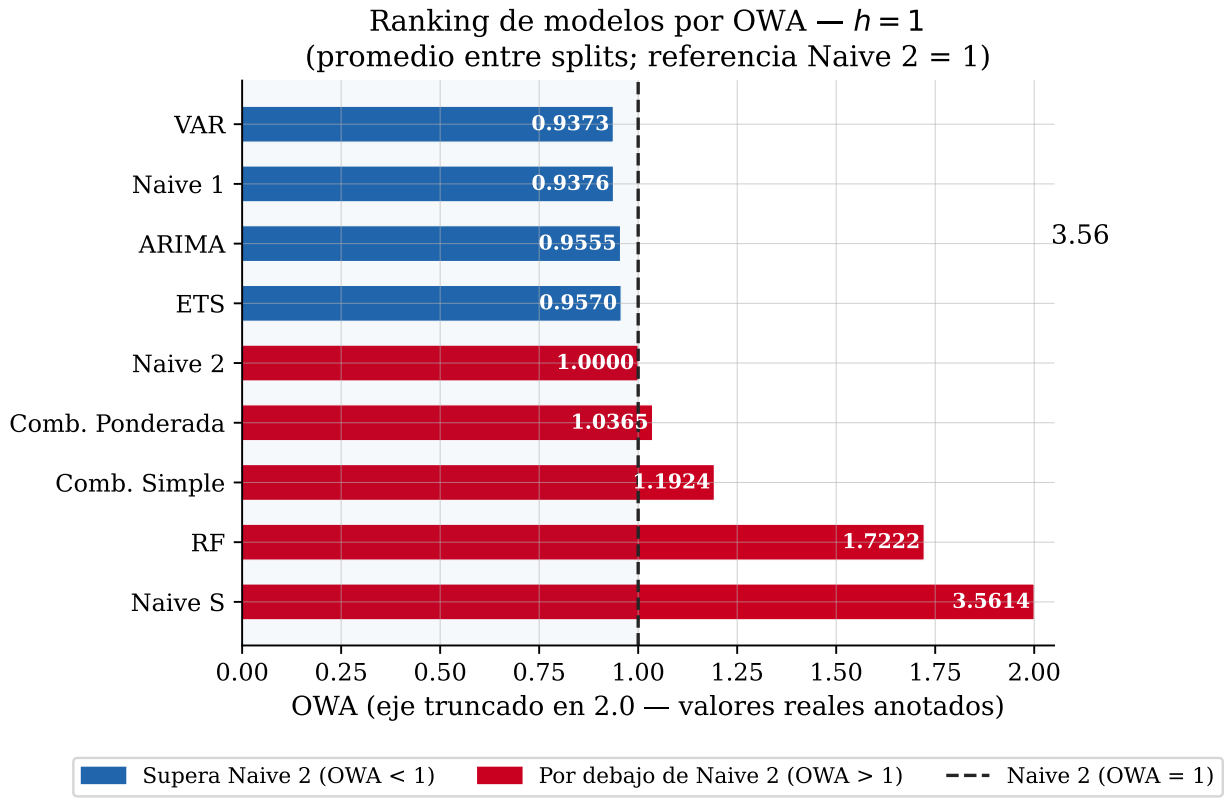

Figura 2: Ranking de modelos por OWA promedio, $h = 1$. La línea de referencia en OWA = 1 corresponde a `Naive2`. Los valores de `NaiveS` y `RF` se muestran completos en el eje; los demás modelos se ubican en el rango 0,9–1,2.

Nota. Elaboración propia.

comportamiento promedio con la misma fiabilidad que su combinación.

Cuadro 2: Comparación de modelos para $h = 3$, promedio entre las tres particiones temporales (18, 24 y 36 meses de prueba). Ordenado por OWA ascendente. Referencia: OWA = 1 corresponde a Naive 2.

#	Modelo	sMAPE (%)	MASE	OWA
1	Naive 1	17,1997	1,5256	0,9726
2	VAR	17,4824	1,5607	0,9913
3	ETS	17,5476	1,5546	0,9919
4	Naive 2	17,8824	1,5494	1,0000
5	ARIMA	17,6763	1,5794	1,0032
6	Comb. Ponderada	18,6139	1,6139	1,0407
7	Comb. Simple	19,2900	1,6510	1,0729
8	RF	23,8777	1,9651	1,2990
9	Naive S	37,0023	2,8699	1,9640

Nota. Elaboración propia. Valores calculados con el pipeline de evaluación fuera de muestra (*rolling-origin*).

La Figura 4 contrasta los OWA de todos los modelos en ambos horizontes y permite leer los movimientos de ordenamiento. El hallazgo más relevante es el desplazamiento del `VAR`: su OWA pasa de 0,937 en $h = 1$ a 0,991 en $h = 3$, lo que indica que la información de las covariables WTI, Brent y ONI tiene valor predictivo para el próximo mes —suficiente para situarlo al frente del campo— pero pierde utilidad marginal rápidamente al ampliar el horizonte, pasando del primer al segundo lugar. Este patrón es coherente con la evidencia de Kwas y Rubaszek (2021) sobre productos básicos en general: las correlaciones dinámicas con variables macroeconómicas son más explotables en el corto plazo, donde la señal aún no ha sido absorbida por la revisión de expectativas de los participantes del mercado. El ascenso relativo de `ETS` en $h = 3$ —de cuarto a tercer lugar— refleja que la suavización exponencial simple captura mejor que los modelos más estructurados la inercia de baja frecuencia que predomina cuando el horizonte se amplía. La Figura 5 confirma que estos movimientos no son artefactos de una sola ventana, sino que se reproducen en las tres particiones con variaciones de ordenamiento menores.

Las Figuras 6 y 7 ilustran los pronósticos de los cuatro mejores modelos en nivel para el período de prueba de la partición de 24 meses (abril 2024 – marzo 2026), que incluye el ciclo de máximos históricos registrado en el mercado del café. Su lectura subraya que la ventaja cuantitativa en OWA no implica que ningún modelo capture la dirección del ciclo de precios: los cuatro modelos tienden a subestimar los picos y a permanecer en un rango más conservador que la serie

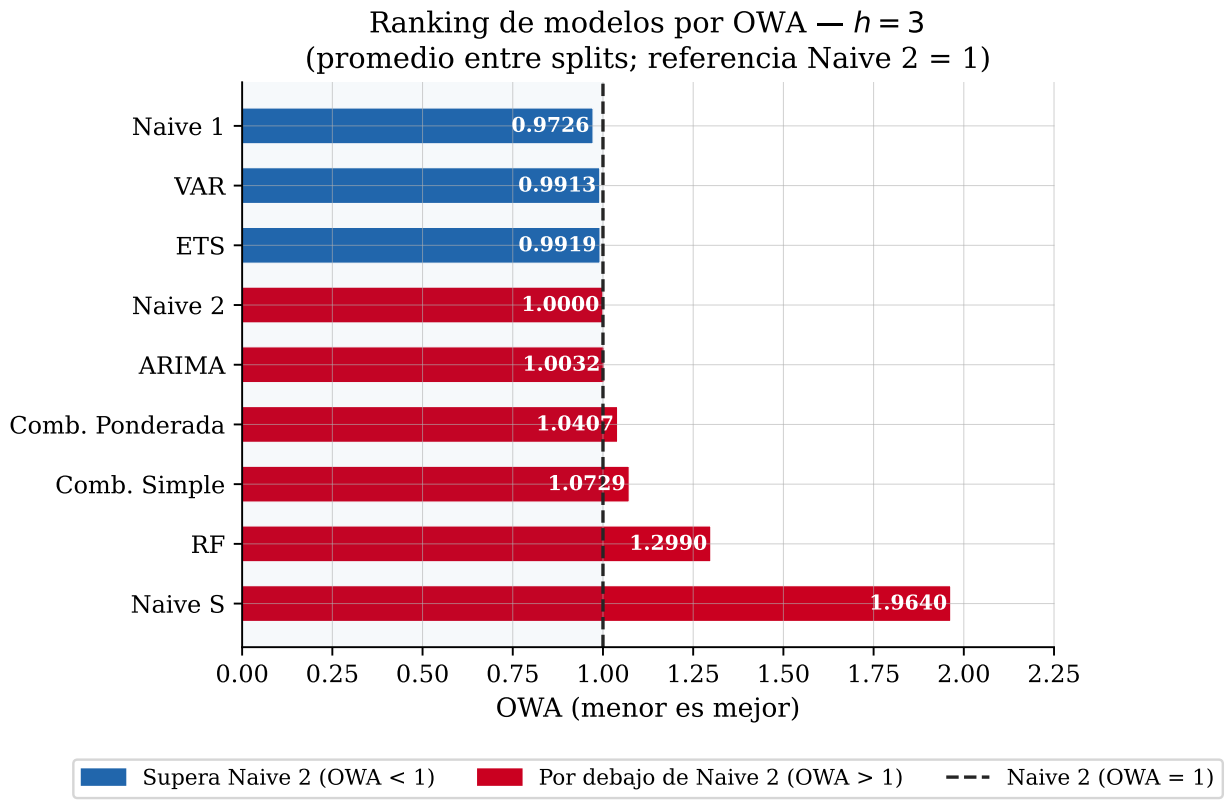

Figura 3: Ranking de modelos por OWA promedio, $h = 3$. La línea de referencia en OWA = 1 corresponde a `Naive2`.

Nota. Elaboración propia.

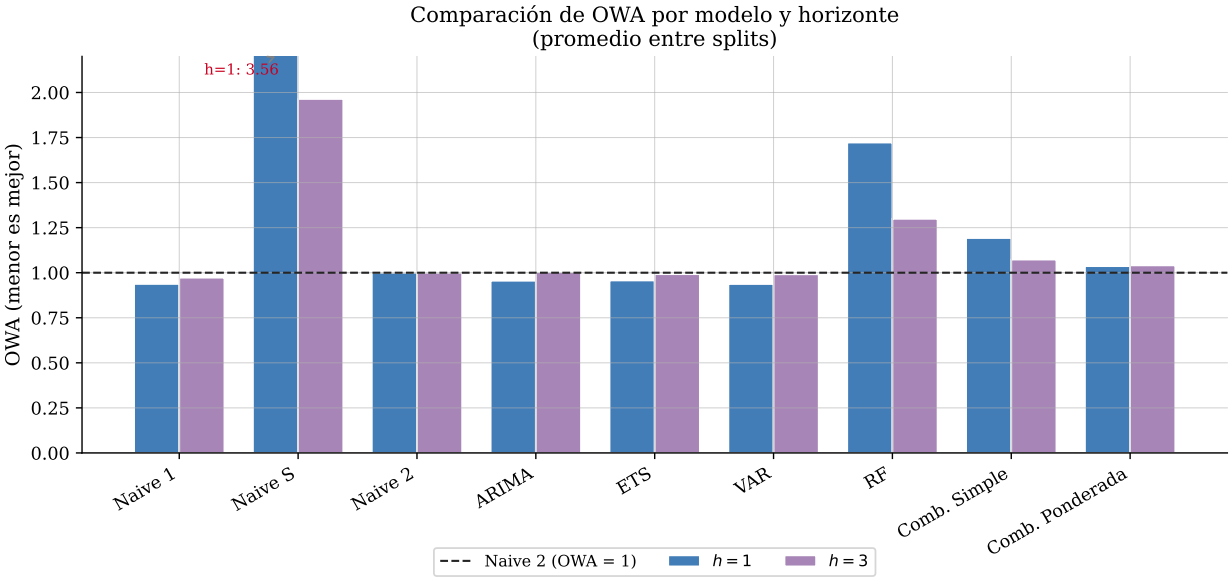

Figura 4: Comparación de OWA por modelo para $h = 1$ y $h = 3$ (promedio entre las tres particiones temporales). La línea horizontal indica OWA = 1 (`Naive2`).

Nota. Elaboración propia.

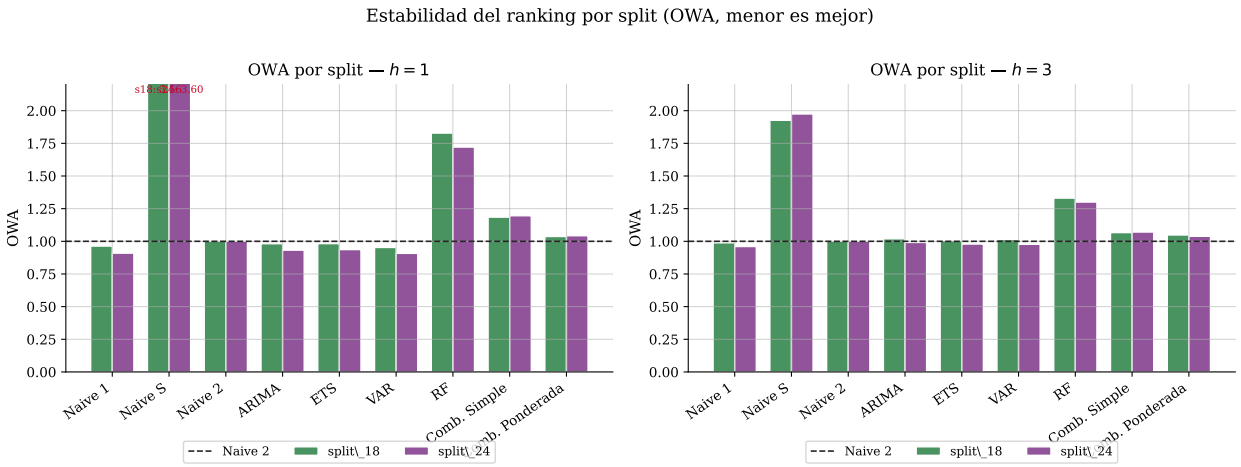

Figura 5: OWA por modelo y partición temporal para $h = 1$ (panel superior) y $h = 3$ (panel inferior). La línea horizontal indica OWA = 1.

Nota. Elaboración propia.

observada, comportamiento esperable dado que ninguno incorpora información sobre los factores estructurales que determinaron la ruptura de nivel documentada en 2024–2025.

6.2. Significancia estadística y el poder de la muestra

La comparación basada en OWA establece un ordenamiento robusto entre modelos, pero no responde a la pregunta de si las diferencias observadas son estadísticamente distinguibles del ruido muestral. Para abordarla, se aplica el test de Diebold y Mariano (1995), que contrasta la hipótesis nula de igual capacidad predictiva entre dos modelos:

$$H_0 : \mathbb{E}[d_t] = 0 \qquad \text{frente a} \qquad H_1 : \mathbb{E}[d_t] < 0, \tag{4}$$

donde $d_t = L(e_{1t}) - L(e_{2t})$ es la diferencia de pérdidas cuadráticas entre el modelo evaluado y `Naive2`. La hipótesis alternativa unilateral $H_1 : \mathbb{E}[d_t] < 0$ corresponde a superioridad predictiva del modelo candidato. El estadístico incorpora la corrección de muestra finita de Harvey et al. (1997), que escala el DM original y lo contrasta contra una $t(n-1)$ en lugar de la normal estándar; la varianza del diferencial se estima con un estimador HAC de $h-1$ rezagos Newey-West, apropiado para errores multi-paso.

La Tabla 3 reporta los resultados para $h = 1$ sobre la partición de 36 meses, que es la muestra

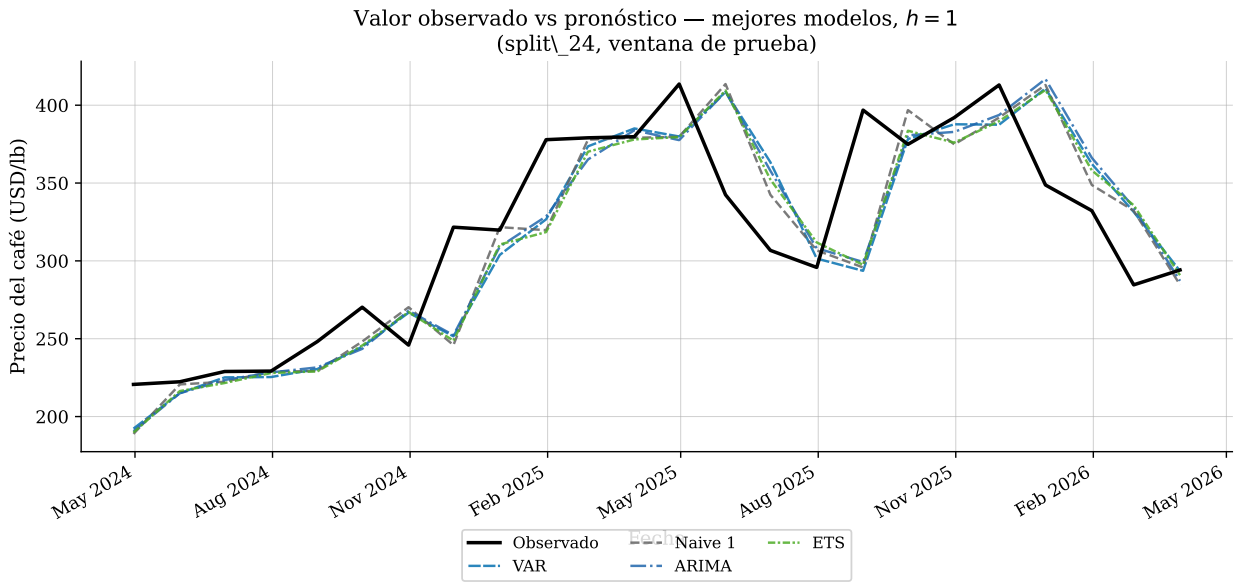

Figura 6: Valores observados y pronósticos ($h = 1$) de los cuatro mejores modelos según OWA, para el período de prueba de la partición de 24 meses (abril 2024 – marzo 2026).

Nota. Elaboración propia con datos de Bloomberg (contrato continuo KC1).

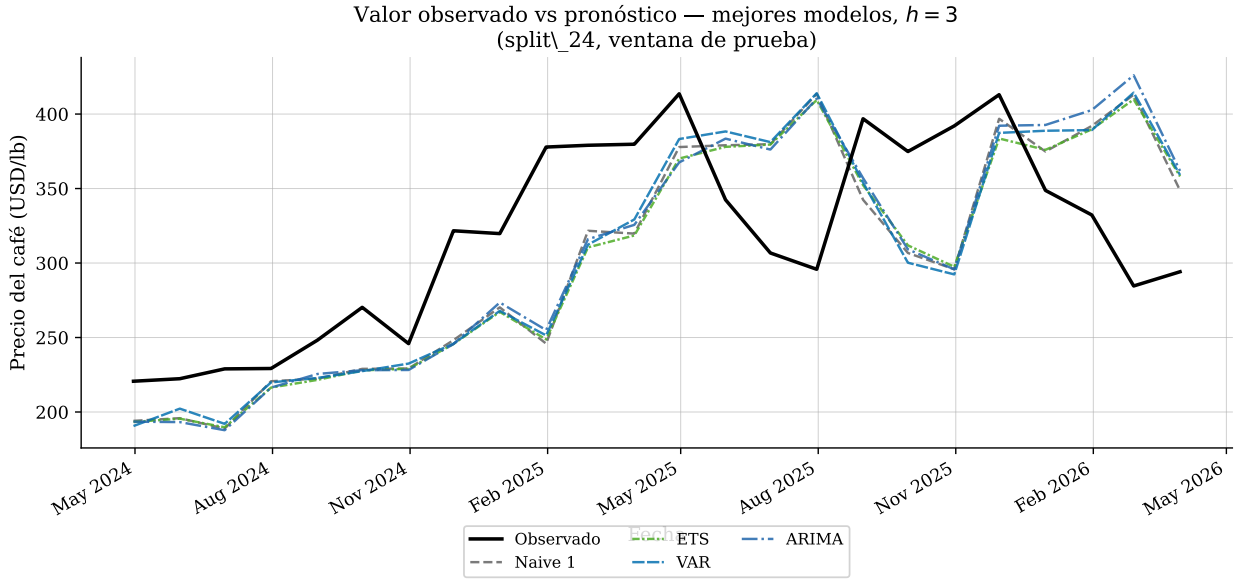

Figura 7: Valores observados y pronósticos ($h = 3$) de los cuatro mejores modelos según OWA, para el período de prueba de la partición de 24 meses (abril 2024 – marzo 2026).

Nota. Elaboración propia con datos de Bloomberg (contrato continuo KC1).

independiente más amplia disponible en el diseño ($n = 36$). Ninguna diferencia alcanza el nivel de significancia del 5 %. Sin embargo, el resultado estadísticamente más relevante no es el p-valor en sí, sino la dirección de las diferencias de pérdida: los cuatro modelos con mejor OWA (`ARIMA`, `ETS`, `Naive1` y `VAR`) presentan $\bar{d} < 0$ sin excepción, lo que confirma que sus pérdidas cuadráticas son sistemáticamente menores que las del benchmark en todos los orígenes disponibles, no como consecuencia de un período particular sino como patrón agregado.

Cuadro 3: Test de Diebold-Mariano vs. `Naive2`, $h = 1$, partición de 36 meses ($n = 36$). H_1: modelo con menor pérdida cuadrática que el benchmark (test unilateral). $\bar{d}$: diferencia media de pérdidas (negativo indica menor pérdida del modelo).

Modelo	*DM*	p-valor	$\bar{d}$	Sig.
ARIMA	−0,4149	0,340	−109,32	—
ETS	−0,4385	0,332	− 94,48	—
Naive 1	−0,4022	0,345	− 85,94	—
VAR	−0,3647	0,359	− 91,27	—
Comb. Ponderada	0,1906	0,575	26,92	—
Comb. Simple	2,6516	0,994	442,84	—
RF	2,7826	0,996	2735,36	—
Naive S	5,1778	1,000	10792,09	—

Nota. Elaboración propia. Corrección de muestra finita HLN (Harvey et al., 1997) aplicada sobre la partición de 36 meses. Ninguna diferencia alcanza significancia al 10 %.

La ausencia de rechazo formal es coherente con las propiedades del test en muestras de esta magnitud y no debe interpretarse como evidencia de igualdad predictiva. La distinción es importante. Coroneo y Iacone (2020) demuestran formalmente que la inferencia asintótica estándar del test DM "entrega pruebas de precisión predictiva con tamaño incorrecto cuando solo un número reducido de observaciones fuera de muestra está disponible", un problema que persiste incluso con la corrección HLN cuando n es pequeño. El propio Diebold (2015), en una revisión retrospectiva de su test, advierte explícitamente sobre el riesgo de sobreinterpretar comparaciones en contextos de evaluación con muestras reducidas: la inferencia frecuentista requiere que n sea suficientemente grande para que la convergencia asintótica sea operativa, condición que $n = 36$ no satisface para detectar mejoras de la magnitud observada. Con una mejora relativa de aproximadamente 6 % en OWA y la variabilidad propia de una serie de productos básicos, se requeriría $n \geq 100$ para alcanzar un poder estadístico razonable frente a esta alternativa. Esta limitación es estructural al diseño, no una deficiencia del modelo.

Más allá del poder estadístico, la literatura sobre pronóstico de precios de productos básicos documenta sistemáticamente que las diferencias entre modelos alternativos y el paseo aleatorio son habitualmente modestas en magnitud aun cuando son económicamente relevantes (Kwas & Rubaszek, 2021). Como argumenta Timmermann (2006), las ventajas predictivas persistentes son difíciles de sostener en mercados donde los participantes las explotan activamente, lo que implica que las mejoras esperables sobre el random walk son, por construcción, de magnitud contenida. En este contexto, la evidencia de robustez más sólida disponible en este diseño no es el p-valor del test DM, sino la consistencia del ordenamiento OWA a través de las tres particiones independientes: los cuatro modelos estadísticos superiores superan a `Naive2` en las seis combinaciones de horizonte y ventana donde el resultado es claro, una regularidad difícilmente atribuible al azar muestral de una sola evaluación.

6.3. Los límites de la complejidad: aprendizaje automático y combinaciones

El resultado más alejado de las expectativas del experimento es el del Random Forest. Con un OWA promedio de 1,722 en $h = 1$ y 1,299 en $h = 3$, el `RF` no supera a `Naive2` en ninguna partición ni horizonte evaluado, situándose entre los peores modelos del universo comparado. Este resultado no es inexplicable: la serie del precio del café presenta propiedades que desfavorecen precisamente los mecanismos de los que se benefician los métodos de ensamble basados en árboles. La autocorrelación condicional se agota rápidamente más allá del primer rezago, la variabilidad idiosincrásica es alta y la señal predictiva de los rezagos de WTI, Brent y ONI es débil para el horizonte mensual en el período analizado. En este entorno, un modelo con 500 árboles y 24 características de entrada tiene escasa estructura estadística que explotar, y la riqueza del espacio de hipótesis se convierte en una fuente de sobreajuste en lugar de una ventaja. La estrategia recursiva utilizada para $h = 3$ agrava el problema: la acumulación del error de pronóstico entre pasos propaga los errores del primer rezago a los horizontes subsiguientes, sin que las covariables puedan compensar.

El contraste con los entornos donde los métodos de machine learning sí dominan es metodológicamente informativo. En la M5, los métodos de gradient boosting lideraron de forma consistente (Makridakis et al., 2022) en un experimento diseñado con miles de series jerárquicas de

demanda minorista, covariables calendario ricas y validación cruzada intensiva: exactamente las condiciones que esta tesis no puede replicar con una sola serie mensual de un producto básico. La configuración implementada en este estudio es deliberadamente estándar —sin búsqueda de hiperparámetros, sin ingeniería de características específica al dominio, sin variables de alta frecuencia— en línea con el nivel de esfuerzo de calibración que sería razonable en una aplicación práctica de pronóstico del sector cafetero (Bojer, 2022). Bajo esas condiciones, el resultado negativo del `RF` es un hallazgo, no una anomalía: delimita las condiciones bajo las cuales la complejidad adicional no genera valor predictivo neto, condición necesaria para saber cuándo podría aportarlo.

Las dos estrategias de combinación de pronósticos tampoco superan al mejor modelo individual. La Combinación Ponderada (OWA 1,037 en $h = 1$) mejora notablemente sobre la Combinación Simple (OWA 1,192), lo que confirma que el mecanismo de ponderación por inverso del RMSE acumulado funciona como se espera —asigna menor peso a los peores modelos— pero no logra neutralizar el efecto de los componentes más débiles del pool. La fuente del problema es estructural: la inclusión de `NaiveS` (OWA 3,561) y `RF` (OWA 1,722) en el promedio eleva el promedio ponderado por encima del umbral del benchmark incluso cuando los cuatro modelos estadísticos de primer nivel producen señales consistentemente buenas. Timmermann (2006) argumenta que las combinaciones son más efectivas cuando los modelos componentes son razonablemente homogéneos en calidad: un pool heterogéneo como el de este experimento —que incluye por diseño modelos de muy distinta precisión— invierte este principio. Una estrategia de selección ex ante del subconjunto de modelos a combinar, que excluyera a priori los modelos de bajo desempeño, probablemente produciría un resultado distinto; su evaluación queda como línea directa de trabajo futuro.

La lectura conjunta de estos resultados lleva a una observación de orden más general. La evidencia empírica de esta tesis —nueve modelos, tres particiones independientes, dos horizontes, protocolo de evaluación sin fuga de información— no respalda la hipótesis de que la complejidad metodológica es, por sí sola, un predictor del desempeño fuera de muestra en series de precios de productos básicos. Los modelos que superan sistemáticamente al benchmark son precisamente los más parsimoniosos: el paseo aleatorio simple y los modelos estadísticos clásicos de una o pocas ecuaciones. Esta regularidad es coherente con el cuerpo de evidencia acumulado en las competencias de pronóstico (Makridakis et al., 2020): la complejidad añade valor cuando hay

suficiente señal en los datos para explotar, y esa condición no está garantizada en series financieras de alta variabilidad idiosincrásica. La consistencia del ordenamiento OWA a través de las tres ventanas de evaluación —y no el resultado puntual de ninguna de ellas— es la evidencia más sólida que este estudio puede ofrecer.

7. Conclusiones

Esta tesis se propuso evaluar la capacidad predictiva de nueve modelos para el pronóstico mensual del precio externo del café colombiano bajo un protocolo de validación fuera de muestra riguroso: ventana expandible de entrenamiento, tres particiones temporales independientes de 18, 24 y 36 meses, y dos horizontes de pronóstico. La pregunta central era cuál de los enfoques evaluados —benchmarks ingenuos, modelos estadísticos univariados, un modelo multivariado y aprendizaje automático— ofrecía la mayor precisión relativa en condiciones comparables de información. La respuesta que emerge del experimento es, a la vez, clara en su jerarquía y matizada en su interpretación. Para $h = 1$, el `VAR` y `Naive1` empatan prácticamente en el primer lugar (OWA promedio de 0,937 y 0,938 respectivamente), mejorando al benchmark en aproximadamente 6 % de forma consistente en las tres particiones sin ninguna excepción; la correlación dinámica con los mercados de energía y con el índice climático ONI contiene, pues, información predictiva de corto plazo que un modelo lineal multivariado puede capturar de forma regular. Para $h = 3$, esta ventaja se evapora: el paseo aleatorio simple (`Naive1`, OWA 0,973) pasa a liderar, los modelos estadísticos se agrupan en torno a la frontera del benchmark y la mejora máxima sobre `Naive2` cae al 2,7 %. Este patrón es coherente con la evidencia de Kwas y Rubaszek (2021) para productos básicos en general: la información de covariables macroeconómicas tiene valor predictivo en horizontes cortos pero lo pierde rápidamente cuando el horizonte se amplía, a medida que la incertidumbre acumulada reduce su utilidad marginal y la dinámica de la serie vuelve a dominar.

A la luz de estos resultados, la evidencia de superioridad predictiva debe leerse con precisión metodológica. El test de Diebold y Mariano (1995) con corrección de muestra finita (Harvey et al., 1997) no rechaza la hipótesis nula de igual capacidad predictiva al 5 % para ningún modelo en ninguna partición. Sin embargo, como demuestran Coroneo y Iacone (2020), la inferencia asintótica estándar del test DM entrega pruebas con tamaño incorrecto cuando el número de obser-

vaciones fuera de muestra es reducido; con $n \leq 36$, el poder estadístico es insuficiente para detectar mejoras de la magnitud observada sin incurrir en una tasa de falsos negativos elevada, y se requeriría un panel de $n \geq 100$ orígenes de prueba para alcanzar poder razonable frente a alternativas de esta magnitud. La ausencia de rechazo formal no equivale, por tanto, a ausencia de diferencias reales. Lo que sí resulta inequívoco es la dirección de las pérdidas diferenciales: los cuatro modelos estadísticos de mejor desempeño presentan $\bar{d} < 0$ para $h = 1$ en todas las particiones evaluadas, confirmando que sus pérdidas cuadráticas son sistemáticamente menores que las del benchmark como patrón agregado y no como consecuencia de un período particular. La consistencia del ordenamiento OWA a través de tres ventanas temporales independientes —y no el p-valor de ninguna prueba aislada— constituye el criterio de robustez más exigente disponible en este diseño y el fundamento empírico central de las conclusiones.

La lección metodológica más nítida del experimento concierne a la relación entre complejidad y desempeño predictivo. El Random Forest, con OWA de 1,722 en $h = 1$ y 1,299 en $h = 3$, no supera al benchmark en ninguno de los escenarios evaluados; las dos estrategias de combinación de pronósticos tampoco lo hacen. Este resultado no es una anomalía sino una consecuencia directa de las propiedades de la serie: baja autocorrelación condicional más allá del primer rezago, alta variabilidad idiosincrásica y señal predictiva débil en los rezagos de las covariables para el horizonte mensual. En este entorno, la riqueza del espacio de hipótesis de un ensamble de 500 árboles no tiene estructura estadística que explotar y degenera en sobreajuste. Las combinaciones de pronósticos adolecen, por su parte, de un problema de composición: la inclusión de modelos de muy distinto nivel de precisión en el pool —en particular `NaiveS` y `RF`— eleva el promedio por encima del umbral del benchmark incluso cuando los modelos estadísticos de primer nivel producen señales consistentemente buenas, lo que ilustra la observación de Timmermann (2006) de que las combinaciones son más efectivas cuando los componentes son razonablemente homogéneos en calidad. La complejidad añade valor cuando hay suficiente señal en los datos para explotarla; documentar con precisión cuándo esa condición no se cumple es, en sí mismo, una contribución.

Estos hallazgos deben leerse considerando las limitaciones que delimitan su alcance interpretativo. La evaluación se realizó sobre una sola serie en un período determinado (enero 2000 – marzo 2026), lo que impide generalizaciones directas a otras series de café, a otros productos básicos o a subperíodos con estructuras de mercado distintas. Los modelos de aprendizaje automático

se implementaron con una configuración estándar y sin búsqueda sistemática de hiperparámetros, reflejo del nivel de esfuerzo de calibración razonable en una aplicación práctica, pero una limitación reconocida frente a protocolos de optimización más intensivos. Las combinaciones se evaluaron sobre el pool completo de modelos, sin selección ex ante del subconjunto a combinar. Solo se produjeron pronósticos puntuales en dos horizontes, dejando fuera la evaluación de calibración probabilística y los horizontes de mediano plazo. Cada una de estas limitaciones es, a la vez, una línea de investigación directamente motivada por los resultados.

Frente a estas limitaciones, las direcciones de trabajo futuro más prometedoras son tres. La primera es la selección del pool de modelos combinados: una estrategia de poda basada en el desempeño histórico acumulado —o en "trimmed means" que excluyan los extremos del pool— tendría una motivación teórica sólida en Timmermann (2006) y podría revertir el resultado negativo de las combinaciones observado aquí. La segunda es la extensión hacia modelos globales de cross-learning entrenados sobre un conjunto amplio de series de productos básicos relacionadas: la evidencia de Semenoglou et al. (2021) muestra que estos enfoques pueden superar tanto a modelos locales como a benchmarks clásicos cuando la información compartida entre series es suficiente y diversa, y el precio del café comparte covariables estructurales —meteorología, ciclos de demanda, dinámica energética— con otros mercados de materias primas que podrían constituir un panel natural de aprendizaje. La tercera es la extensión al pronóstico probabilístico y a horizontes más largos ($h = 6$, $h = 12$): la evaluación de la calibración de los modelos —más allá del error puntual— es especialmente relevante para la gestión del riesgo de precio en el sector cafetero, donde las decisiones de cobertura requieren estimaciones de incertidumbre y no solo valores esperados. Esta tesis produce evidencia empírica acotada, replicable y útil como punto de partida; la invitación que deja abierta es a continuar el experimento con datos más ricos, pools más selectivos y horizontes más amplios.

Referencias

Aduteye, E. K., Sete, T. T., & Chi, Y. N. (2023). Time Series Analysis of Global Prices of Coffee: Insights into a Complex Market. *International Journal of Business and Economics*, *8*(2), 138-151.

Bates, J. M., & Granger, C. W. J. (1969). The Combination of Forecasts. *Operational Research Quarterly*, *20*(4), 451-468. https://doi.org/10.1057/jors.1969.103

Bojer, C. S. (2022). Understanding machine learning-based forecasting methods: A decomposition framework and research opportunities. *International Journal of Forecasting*, *38*(4), 1555-1561. https://doi.org/10.1016/j.ijforecast.2021.11.003

Box, G. E. P., & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control* (revised). Holden-Day.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Cardozo Rueda, K. S. (2022). Aplicación de redes neuronales artificiales para el pronóstico de precios de café. *Revista Colombiana de Tecnologías de Avanzada*, *1*(39), 113-117.

Castle, J. L., Doornik, J. A., & Hendry, D. F. (2021). Forecasting Principles from Experience with Forecasting Competitions. *Forecasting*, *3*(1), 138-165. https://doi.org/10.3390/forecast3010010

Cepeda C., F. (1981). A propósito de los modelos sobre café en Colombia. *Revista Colombiana de Matemáticas*, *3*.

Chen, Y. (2024). Research on Machine Learning-based Prediction of Coffee Futures Prices. *Highlights in Science, Engineering and Technology*, *92*, 199-209.

Chen, Z., Goh, H. S., Sin, K. L., Lim, K., Chung, N. K. H., & Liew, X. Y. (2021). Automated agriculture commodity price prediction system with machine learning techniques. *Advances in Science, Technology and Engineering Systems Journal*, *6*(2).

Coroneo, L., & Iacone, F. (2020). Comparing predictive accuracy in small samples using fixed-smoothing asymptotics. *Journal of Applied Econometrics*, *35*(4), 391-409. https://doi.org/10.1002/jae.2756

Deina, C., Prates, M. H. d. A., Alves, C. H. R., Martins, M. S. R., Trojan, F., & Stevan, S. L. J. (2022). A methodology for coffee price forecasting based on extreme learning machines. *Information Processing in Agriculture*, *9*(4), 556-565. https://doi.org/10.1016/j.inpa.2021.07.003

Díaz-Pinzón, J. E. (2025). Análisis económico-social del precio interno base del café pergamino seco en Colombia: modelamiento, predicción y proyección 2025–2028. *Revista Facultad de Ciencias Económicas*, *33*(1), 97-112. https://doi.org/10.18359/rfce.7760

Diebold, F. X. (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests. *Journal of Business & Economic Statistics*, *33*(1), 1-9. https://doi.org/10.1080/07350015.2014.983236

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253-263. https://doi.org/10.1080/07350015.1995.10524599

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, *29*(1), 108-121. https://doi.org/10.1016/j.ijforecast.2012.06.004

González Casimiro, M. P. (2009). *Técnicas de predicción económica*. Universidad del País Vasco.

Hamouda, F., Arfaoui, N., & Naeem, M. A. (2025). Forecasting Energy Commodity Prices Amidst Worldwide Energy Transitions Using Artificial Intelligence Models. *The Energy Journal*, *46*(5), 215-244. https://doi.org/10.1177/01956574251340012

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, *13*(2), 281-291. https://doi.org/10.1016/S0169-2070(96)00719-4

Humérez Quiroz, J. (2012). Combinación de pronósticos: una aplicación a la inflación de Bolivia. *Revista de Análisis*, *16*, 59-93.

Hwase, T. K., & Fofanah, A. J. (2021). Machine Learning Model Approaches for Price Prediction in Coffee Market. *International Journal of Scientific Research in Science and Technology*, *8*(6), 10-48.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3.ª ed.). OTexts.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001

International Coffee Organization. (2011). *Relationship between Coffee Prices in Physical and Futures Markets* (Study N.º ICC 107-4). International Coffee Organization. London, United Kingdom.

Iregui, A. M., Núñez, H. M., & Otero, J. (2025). Testing the Efficiency of Oil Price Forecast Revisions in Times of COVID-19 and the Russia–Ukraine Conflict. *Journal of Commodity Markets*, *40*, 100513. https://doi.org/10.1016/j.jcomm.2025.100513

Kipkoech, J., Bett, H., & Kirui, O. (2023). Forecasting agricultural commodity prices using ARIMA model: A case of rice and wheat prices in Kenya. *International Journal of Management Sciences*.

Kwas, M., & Rubaszek, M. (2021). Forecasting Commodity Prices: Looking for a Benchmark. *Forecasting*, *3*(2), 447-459. https://doi.org/10.3390/forecast3020027

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer. https://doi.org/10.1007/978-3-540-27752-1

Ly, R., Traore, F., & Dia, K. (2021). Forecasting Commodity Prices Using Long Short-Term Memory Neural Networks. https://doi.org/10.48550/arXiv.2101.03087

Madrigal Espinoza, S. D. (2011). *Pronóstico de series temporales con estacionalidad* [Tesis doctoral]. Universidad Autónoma de Nuevo León.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54-74. https://doi.org/10.1016/j.ijforecast.2019.04.014

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, *38*(4), 1346-1364. https://doi.org/10.1016/j.ijforecast.2021.11.013

Medeiros, M. C., Vasconcelos, G. F. R., Veiga, Á., & Zilberman, E. (2021). Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods. *Journal of Business & Economic Statistics*, *39*(1), 98-119. https://doi.org/10.1080/07350015.2019.1637745

Milas, C., Otero, J., & Panagiotidis, T. (2004). Forecasting the Spot Prices of Various Coffee Types Using Linear and Non-Linear Error Correction Models. *International Journal of Finance and Economics*, *9*, 277-288. https://doi.org/10.1002/ijfe.245

Naveena, K., Singh, S., Rathod, S., & Singh, A. (2017). Hybrid Time Series Modelling for Forecasting the Price of Washed Coffee (Arabica Plantation Coffee) in India. *International Journal of Agriculture Sciences*, *9*(10), 4004-4007.

Pinto-Rodriguez, V.-H., Cobos-Lozada, C.-A., & Nieto-Muñoz, A.-M. (2025). Systematic Review of Methods for Specialty Coffee Price Estimation. *Revista Facultad de Ingeniería*, *34*(71), e18089. https://doi.org/10.19053/uptc.01211129.v34.n71.2025.18089

Semenoglou, A.-A., Spiliotis, E., Makridakis, S., & Assimakopoulos, V. (2021). Investigating the accuracy of cross-learning time series forecasting methods. *International Journal of Forecasting*, *37*(3), 1072-1084. https://doi.org/10.1016/j.ijforecast.2020.11.009

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*, *48*(1), 1-48. https://doi.org/10.2307/1912017

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, *16*(4), 437-450. https://doi.org/10.1016/S0169-2070(00)00065-0

Timmermann, A. (2006). Forecast Combinations. En G. Elliott, C. W. J. Granger & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (pp. 135-196, Vol. 1). Elsevier. https://doi.org/10.1016/S1574-0706(05)01004-9

Useche Mahecha, E. E. (2024). *Análisis estadístico de la fluctuación de precios pagados por una compañía agroindustrial en la compra de lotes de café de acuerdo con la zonificación geográfica entre 2021–2023* [Trabajo de grado]. Universidad Nacional Abierta y a Distancia (UNAD).

Wallis, K. F. (2014). Revisiting Francis Galton's Forecasting Competition. *Statistical Science*, *29*(3), 420-424. https://doi.org/10.1214/14-STS468

Wang, X., Hyndman, R. J., Li, F., & Kang, Y. (2023). Forecast combinations: An over 50-year review. *International Journal of Forecasting*, *39*(4), 1518-1547. https://doi.org/10.1016/j.ijforecast.2022.11.005

A. Especificaciones seleccionadas por modelo

Esta sección documenta las especificaciones de los modelos seleccionadas por AIC sobre la muestra inicial de entrenamiento de cada partición. Las especificaciones resultaron idénticas en las tres particiones temporales.

Cuadro 4: Especificaciones de los modelos para la corrida final. Para ARIMA, ETS y VAR la especificación se seleccionó por AIC en la muestra inicial del *split* y permaneció fija para todos los orígenes del rolling-origin de ese *split*.

Modelo	**Especificación**	**Criterio de selección**
ARIMA	SARIMA$(1,1,1)(0,1,1)_{12}$	AIC, muestra inicial del *split*
ETS	ETS(A,N,N): error aditivo, sin tendencia, sin estacionalidad	AIC, muestra inicial del *split*
VAR	VAR(2) en primeras diferencias, sistema de 4 variables	AIC, máx. 12 rezagos
RF	500 árboles, `max_features`$=\sqrt{p}$, `min_samples_leaf`$=5$	Fijo (sin selección)

Nota. Elaboración propia.

Las cuatro series del sistema VAR son: Δ`coffee_price`, Δ`wti`, Δ`brent` y Δ`oni`. Los pronósticos en diferencias se reconvierten a nivel mediante acumulación desde el último valor observado en la muestra de entrenamiento.

Para el RF, las 24 características incluyen 12 rezagos de `coffee_price` ($y_{t-1},\ldots,y_{t-12}$) más 4 rezagos de cada covariable. Para $h=3$, el pronóstico es recursivo con *forward-fill* de covariables.

B. Tablas de resultados por partición y horizonte

Las tablas de resultados completas (sMAPE, MASE, OWA, RMSE, MAE) por cada combinación de partición y horizonte se generan automáticamente por el pipeline y se incluyen a continuación. El bloque contiene seis tablas en formato *booktabs*: cuatro con el ordenamiento canónico por partición y horizonte, y dos tablas de comparación agregada (una por horizonte) ordenadas por OWA ascendente. Todas las cifras proceden directamente de los archivos de salida del pipeline y son idénticas a las del archivo de comparación final.

Cuadro 5: Resultados por modelo — split_18, $h = 1$

Modelo	sMAPE (%)	MASE	OWA	RMSE	MAE
Naive 1	10.0501	0.9479	0.9634	44.4138	34.0444
Naive S	39.7977	3.2442	3.5554	126.3457	116.5250
Naive 2	10.4699	0.9802	1.0000	45.3087	35.2069
ARIMA	10.2134	0.9679	0.9814	43.9773	34.7630
ETS	10.2402	0.9666	0.9821	44.2284	34.7186
VAR	9.9247	0.9387	0.9528	44.2847	33.7163
RF	19.5917	1.7502	1.8284	77.0459	62.8612
Comb. Simple	12.3856	1.1633	1.1849	51.2887	41.7825
Comb. Ponderada	10.7859	1.0226	1.0367	45.9326	36.7292

Cuadro 6: Resultados por modelo — split_18, $h = 3$

Modelo	sMAPE (%)	MASE	OWA	RMSE	MAE
Naive 1	19.0717	1.7880	0.9883	72.6865	64.2194
Naive S	39.7977	3.2442	1.9264	126.3457	116.5250
Naive 2	19.4933	1.7911	1.0000	72.2829	64.3331
ARIMA	19.5817	1.8518	1.0192	74.1865	66.5125
ETS	19.3901	1.8178	1.0048	72.6055	65.2902
VAR	19.5315	1.8378	1.0140	73.8080	66.0105
RF	26.8308	2.3013	1.3306	96.3402	82.6579
Comb. Simple	20.9193	1.8983	1.0665	76.3043	68.1805
Comb. Ponderada	20.4607	1.8755	1.0484	74.1905	67.3623

Cuadro 7: Resultados por modelo — split_24, $h = 1$

Modelo	sMAPE (%)	MASE	OWA	RMSE	MAE
Naive 1	9.0292	0.8226	0.9092	39.4818	28.9250
Naive S	38.6683	2.9911	3.5959	116.1001	105.1687
Naive 2	10.0592	0.8934	1.0000	40.8566	31.4145
ARIMA	9.2478	0.8447	0.9324	39.1528	29.7011
ETS	9.3124	0.8464	0.9365	39.3859	29.7585
VAR	9.0129	0.8215	0.9077	39.3408	28.8843
RF	17.5012	1.5212	1.7212	68.0994	53.4873
Comb. Simple	12.0293	1.0678	1.1955	46.4448	37.5459
Comb. Ponderada	10.4453	0.9352	1.0426	41.1119	32.8831

Cuadro 8: Resultados por modelo — split_24, $h = 3$

Modelo	sMAPE (%)	MASE	OWA	RMSE	MAE
Naive 1	17.4718	1.5700	0.9600	64.7317	55.2021
Naive S	38.6683	2.9911	1.9747	116.1001	105.1687
Naive 2	18.4668	1.6121	1.0000	65.0318	56.6835
ARIMA	17.9843	1.6263	0.9913	66.0821	57.1819
ETS	17.8279	1.6002	0.9790	64.7285	56.2644
VAR	17.7220	1.6031	0.9770	65.6033	56.3660
RF	24.6293	2.0425	1.3003	85.8613	71.8166
Comb. Simple	19.8867	1.7165	1.0708	68.8429	60.3527
Comb. Ponderada	19.1574	1.6749	1.0382	66.6053	58.8899

Cuadro 9: Resultados por modelo — split_36, $h = 1$

Modelo	sMAPE (%)	MASE	OWA	RMSE	MAE
Naive 1	8.2756	0.6626	0.9403	33.2360	23.1847
Naive S	32.5409	2.3744	3.5329	98.6481	83.0861
Naive 2	8.8290	0.7025	1.0000	34.3081	24.5805
ARIMA	8.3604	0.6733	0.9527	33.0228	23.5594
ETS	8.3584	0.6728	0.9523	33.1570	23.5438
VAR	8.3940	0.6688	0.9514	33.2598	23.4010
RF	14.1814	1.1435	1.6171	56.4200	40.0143
Comb. Simple	10.5648	0.8408	1.1967	39.0189	29.4204
Comb. Ponderada	9.0399	0.7280	1.0301	34.5393	25.4731

Cuadro 10: Resultados por modelo — split_36, $h = 3$

Modelo	sMAPE (%)	MASE	OWA	RMSE	MAE
Naive 1	15.0555	1.2189	0.9694	54.4358	42.6514
Naive S	32.5409	2.3744	1.9908	98.6481	83.0861
Naive 2	15.6872	1.2450	1.0000	54.6197	43.5649
ARIMA	15.4630	1.2601	0.9989	55.4968	44.0930
ETS	15.4249	1.2459	0.9920	54.4457	43.5954
VAR	15.1937	1.2412	0.9827	55.0002	43.4319
RF	20.1730	1.5513	1.2660	71.5248	54.2843
Comb. Simple	17.0641	1.3383	1.0814	57.8714	46.8300
Comb. Ponderada	16.2236	1.2912	1.0357	55.8977	45.1830

Cuadro 11: Comparación de modelos, $h = 1$ (promedio entre splits). Ordenado por OWA ascendente. Referencia OWA = 1: Naive 2.

#	Modelo	sMAPE (%)	MASE	OWA	RMSE	MAE
1	VAR	9.1105	0.8097	0.9373	38.9618	28.6672
2	Naive 1	9.1183	0.8110	0.9376	39.0439	28.7181
3	ARIMA	9.2739	0.8286	0.9555	38.7176	29.3412
4	ETS	9.3036	0.8286	0.9570	38.9237	29.3403
5	Naive 2	9.7861	0.8587	1.0000	40.1578	30.4006
6	Comb. Ponderada	10.0904	0.8953	1.0365	40.5279	31.6951
7	Comb. Simple	11.6599	1.0240	1.1924	45.5841	36.2496
8	RF	17.0914	1.4716	1.7222	67.1884	52.1209
9	Naive S	37.0023	2.8699	3.5614	113.6979	101.5933

Cuadro 12: Comparación de modelos, $h = 3$ (promedio entre splits). Ordenado por OWA ascendente. Referencia OWA = 1: Naive 2.

#	Modelo	sMAPE (%)	MASE	OWA	RMSE	MAE
1	Naive 1	17.1997	1.5256	0.9726	63.9513	54.0243
2	VAR	17.4824	1.5607	0.9913	64.8039	55.2695
3	ETS	17.5476	1.5546	0.9919	63.9265	55.0500
4	Naive 2	17.8824	1.5494	1.0000	63.9782	54.8605
5	ARIMA	17.6763	1.5794	1.0032	65.2551	55.9291
6	Comb. Ponderada	18.6139	1.6139	1.0407	65.5645	57.1451
7	Comb. Simple	19.2900	1.6510	1.0729	67.6729	58.4544
8	RF	23.8777	1.9651	1.2990	84.5754	69.5863
9	Naive S	37.0023	2.8699	1.9640	113.6979	101.5933

Cambios de ranking entre horizontes

Cuadro 13: Cambios de ranking por OWA entre $h = 1$ y $h = 3$ (promedio entre *splits*). La columna "Dirección" indica si el modelo mejoró (↑), empeoró (↓) o mantuvo su posición (=) al pasar de $h = 1$ a $h = 3$.

Modelo	**Rank** $h = 1$	**Rank** $h = 3$	**Dirección**
VAR	1	2	↓
Naive 1	2	1	↑
ARIMA	3	5	↓
ETS	4	3	↑
Naive 2	5	4	↑
Comb. Ponderada	6	6	=
Comb. Simple	7	7	=
RF	8	8	=
Naive S	9	9	=

Nota. Elaboración propia.

C. Detalles del pipeline computacional

Estructura de archivos

- `config.py` — parámetros globales (rutas, variable objetivo, $m = 12$, benchmark).
- `load_master_dataset.py` — carga y validación del dataset maestro.
- `benchmarks.py` — funciones `naive1`, `naiveS`, `naive2`.
- `models.py` — selección y estimación de ARIMA y ETS.
- `var_model.py` — orden por AIC, estimación en diferencias, reconstrucción en nivel.
- `rf_model.py` — construcción de características, entrenamiento, pronóstico recursivo.
- `evaluation.py` — métricas por *split* y resumen agregado.
- `metrics.py` — implementaciones de sMAPE, MASE, OWA, RMSE y MAE.
- `run_experiment.py` — coordinación de la corrida completa.

- `build_tables.py` — consolidación de CSV y generación de tablas LaTeX.
- `build_figures.py` — generación de figuras PNG y PDF.

Dependencias principales

Python 3.11; `pandas`, `numpy`, `statsmodels` (ARIMA, ETS, VAR), `scikit-learn` (RF), `matplotlib` (figuras).

D. Nota de reproducibilidad

Para reproducir los resultados desde el código fuente:

1. Verificar que `data/master_monthly_dataset.csv` esté presente con las columnas requeridas documentadas en el capítulo de datos.
2. Desde `code/`, ejecutar `python run_experiment.py` para generar pronósticos y métricas en `outputs/`.
3. Ejecutar `python build_tables.py` para generar las tablas consolidadas.
4. Ejecutar `python build_figures.py` para generar las figuras.
5. Compilar `latex/Tesis.tex` con `latexmk -pdf` (requiere `biber` para la bibliografía).

Los resultados de ARIMA, ETS y VAR son determinísticos. Para el RF se recomienda fijar `random_state` en `rf_model.py` para reproducibilidad exacta entre corridas.