

UNIVERSIDAD SANTO TOMÁS

Pronóstico del Precio Diario de Bolsa en el Mercado Eléctrico Colombiano Usando Modelos Autorregresivos.

Realizado por

Juan David Becerra Pulido

Proyecto de Grado presentado en cumplimiento del requisito
para optar por el grado de Ingeniería Electrónica



Grupo de Investigación MEM (Modelado-Electrónica-Monitoreo)
Facultad de Ingeniería Electrónica
División de Ingenierías

Noviembre de 2025, Bogotá, D.C., Colombia

Pronóstico del Precio Diario de Bolsa en el Mercado Eléctrico Colombiano Usando Modelos Autorregresivos.

Realizado por

Juan David Becerra Pulido

Proyecto de Grado presentado en cumplimiento del requisito
para optar por el grado de Ingeniería Electrónica

Dirigido por

Gianina Garrido Silva M.Sc.

Codirigido por

Harrynson Ramírez Murillo Ph.D.

Grupo de Investigación MEM (Modelado-Electrónica-Monitoreo)
Sistemas de Energía y Automatización
Facultad de Ingeniería Electrónica
División de Ingenierías

Noviembre de 2025, Bogotá, D.C., Colombia

Autoridades de la Universidad

RECTOR GENERAL

R.P. FRAY ÁLVARO JOSÉ ARANGO RESTREPO, O.P

VICERRECTOR ADMINISTRATIVO Y FINANCIERO GENERAL

R.P. FRAY HERNÁN YESID RIVERA ROBERTO, O.P

VICERRECTOR ACADÉMICO GENERAL

R.P. FRAY MAURICIO ANTONIO CORTÉS GALLEGO, O.P

SECRETARIA GENERAL

Dra. LUCERO GALVIS CANO

DECANO DIVISIÓN DE INGENIERÍAS

R.P FRAY JAVIER ANTONIO HINCAPIÉ ARDILA, O.P

SECRETARIA DE DIVISIÓN

E.C. LUZ PATRICIA ROCHA CAICEDO

DECANO FACULTAD DE INGENIERÍA ELECTRÓNICA

ING. LEONARDO FABIO YÉPES ARBELÁEZ, PhD.

Nota de aceptación

Firma del tutor

Firma del jurado

Firma del jurado

BOGOTÁ D.C. _____ DE 2025

Advertencia

La Universidad Santo Tomás no se hace responsable de las opiniones y conceptos expresados en el trabajo de grado, solo velará por qué no se publique nada contrario al dogma ni a la moral católica y porque el trabajo no tenga ataques personales y únicamente se vea el anhelo de buscar la verdad científica.

Capítulo III –Art. 46 del Reglamento de la Universidad Santo Tomás.

Dedicatoria

Este trabajo de grado está dedicado con profundo cariño y gratitud a mi familia, quienes han sido el pilar fundamental en cada etapa de mi vida. En especial, a mis padres, Richard William Becerra Martínez y María Elena Pulido Valbuena, por su amor incondicional, su apoyo constante y por enseñarme, con su ejemplo, el valor del esfuerzo, la responsabilidad y la perseverancia. Gracias por acompañarme en cada desafío, por creer en mí incluso cuando las dificultades parecían mayores, y por motivarme a alcanzar mis metas tanto en el ámbito académico como personal.

De igual manera, a mis hermanas Diana Becerra y Tania Becerra, quienes con su ejemplo, sus valiosos consejos y su fortaleza me han inspirado a seguir adelante y a no rendirme ante los obstáculos. Su guía y compañía han sido un apoyo esencial en este camino, recordándome siempre la importancia de la constancia y la dedicación.

También dedico este logro a mis hermanos menores, Antonella Montaña e Ian Paolo Montaña, quienes son el motor de mi vida y una fuente inagotable de alegría e inspiración. Con su cariño me impulsan a ser mejor persona y profesional, y me motivan cada día a continuar esforzándome para servirles de ejemplo y orgullo.

Este logro no solo simboliza el cumplimiento de un objetivo profesional, sino también el reflejo de todos los valores, enseñanzas y sacrificios que mi familia sembró en mí. A ellos dedico este trabajo con el más sincero reconocimiento, amor y gratitud.

Agradecimientos

A Dios, por ser mi guía constante, por darme la fortaleza y la sabiduría necesarias para superar los desafíos y permitirme culminar con éxito esta etapa tan importante de mi vida. Sin su bendición y propósito nada de esto habría sido posible.

A mis padres, Richard William Becerra Martínez y María Elena Pulido Valbuena, por su amor incondicional, su esfuerzo incansable y por ser el ejemplo más grande de trabajo, honestidad y perseverancia. Gracias por creer en mí, por apoyarme en cada paso y por motivarme a alcanzar mis metas tanto como ingeniero como persona.

A mis hermanas, Diana Becerra y Tania Becerra, por su apoyo constante, sus consejos y por ser una inspiración en mi camino. A mis tíos, Sandra Becerra y James Frank Becerra, por su apoyo y confianza. Su presencia y respaldo han sido una fuente importante de motivación en mi vida académica y personal.

A mis amigos, por compartir conmigo este camino académico, por su compañía, su ayuda y los momentos de alegría que hicieron más llevadero este proceso.

Finalmente, expreso mi más sincero agradecimiento a los ingenieros Gianina Garrido Silva y Harrynson Ramírez Murillo, quienes con su orientación, compromiso y conocimiento fueron guías fundamentales en el desarrollo de este proyecto. Su apoyo académico y humano contribuyó de manera invaluable al cumplimiento de los objetivos de esta investigación.

A todos ellos, gracias por ser parte de este logro que representa no solo un título profesional, sino también el resultado de un esfuerzo compartido lleno de aprendizajes y crecimiento personal.

Índice general

1. Introducción	1
1.1. Planteamiento del problema	1
1.2. Objetivos	2
1.2.1. Objetivo General	2
1.2.2. Objetivos Específicos	2
1.3. Justificación	3
1.4. Impacto social	4
2. Estado del arte	5
2.1. Modelos Autorregresivos para la Predicción de Energías Renovables	5
2.2. Generación Renovable e Impacto en el Precio de la Energía	6
2.3. Aplicaciones en América Latina y Colombia	6
2.4. Brechas y Contribución Propuesta	7
3. Marco Teórico	8
3.1. Mercado eléctrico colombiano	8
3.1.1. Estructura general del mercado	9
3.2. Aplicación de modelos ARIMA y SARIMA en mercados eléctricos	9
3.3. Fundamentación Matemática de los Modelos Autorregresivos	10
3.3.1. Modelo Autorregresivo (AR)	11
3.3.2. Modelo Autorregresivo con Variables Exógenas (ARX)	11
3.3.3. Modelo SARIMA	12
3.3.4. Modelo SARIMAX (Generalización de la Familia Autorregresiva)	12
3.3.5. Significado de los Parámetros	13
4. Desarrollo Conceptual	15
4.1. Descripción del dataset	15
4.2. Caracterización estadística y relacional de las variables	16
4.2.1. Medidas de dependencia y correlación	17
4.3. Preparación para el modelado	18
4.3.1. Generación de rezagos y estructura temporal	18
4.3.2. Transformaciones y estandarización	19
4.3.3. División temporal de los datos	19
4.3.4. Horizonte de predicción	20
4.3.5. Condiciones de validación conceptual	20
4.4. Organización del proceso de modelado	20
4.4.1. Adquisición y preparación de datos	21
4.4.2. Construcción del conjunto de entrenamiento	21
4.4.3. Estimación y ajuste de parámetros	21
4.4.4. Análisis conceptual de sensibilidad y métricas de error	21

4.4.5. Criterios de información: AIC y BIC	22
5. Resultados y Discusión	24
5.1. Selección y depuración de variables	25
5.1.1. Carga y filtrado inicial del dataset	25
5.1.2. Selección preliminar de variables	26
5.1.3. Selección y preprocesamiento de variables	27
5.1.4. Preprocesamiento de las variables	29
5.1.4.1. Tratamiento de valores atípicos	29
5.1.4.2. Pruebas de estacionalidad	29
5.1.4.3. Transformaciones y estandarización	30
5.1.4.4. Generación de rezagos temporales	31
5.1.4.5. Análisis de multicolinealidad	33
5.1.4.6. Resultado final del preprocesamiento	33
5.2. Entrenamiento y ajuste de los modelos autorregresivos	34
5.2.1. Modelo base ARX	34
5.2.2. Entrenamiento del modelo SARIMA	35
5.2.3. Ajustes propuestos y comparación de configuraciones	36
5.2.4. Entrenamiento simultáneo de modelos autorregresivos	37
5.2.5. Análisis de sensibilidad OAT (One-at-a-Time)	40
6. Conclusiones y Trabajos futuros	44
 Bibliografía	 46

Capítulo 1

Introducción

1.1. Planteamiento del problema

La integración de energías renovables en Colombia ha experimentado un crecimiento acelerado en los últimos años, impulsado por la necesidad de diversificar la matriz energética y reducir el impacto ambiental de los combustibles fósiles [1]. Desde 2018, la capacidad instalada de energías renovables no convencionales, como la solar y la eólica, ha aumentado notablemente; mientras en 2018 apenas se alcanzaban los 28 MW, en 2023 esta cifra ya superaba los 2.500 MW, y para 2028 se proyecta que pueda llegar a 6.000 MW, representando aproximadamente el 12 % de la matriz energética nacional [2].

Este rápido incremento viene acompañado de la alta intermitencia que caracteriza a fuentes como la solar y la eólica. Por ejemplo, en La Guajira—una región reconocida por su alto potencial eólico, con velocidades promedio de 9 m/s—se planea integrar 1.5 GW de energía eólica para 2025, aunque su variabilidad diaria puede superar el 30 % durante los períodos de transición climática [3]. Dichas fluctuaciones generan desequilibrios entre oferta y demanda en el sistema eléctrico, lo cual se refleja en las oscilaciones del precio promedio diario de bolsa, un indicador esencial para evaluar la dinámica económica del mercado eléctrico colombiano.

En días de alta generación, el exceso de oferta tiende a presionar a la baja el precio, mientras que en períodos de baja generación, como durante la noche, el precio puede incrementarse de manera abrupta. Estas variaciones impactan la rentabilidad de los generadores y generan tarifas volátiles que afectan a los consumidores, situación que ha generado pérdidas económicas significativas, estimadas en 15 M USD en 2022 [4].

Tradicionalmente, se han empleado modelos autorregresivos de series de tiempo basados únicamente en datos históricos para predecir el comportamiento de los precios (modelos endógenos). Sin embargo, estos enfoques pueden no capturar por completo la complejidad del mercado eléctrico, ya que factores

externos —como condiciones meteorológicas, variaciones en la demanda y otros indicadores económicos— influyen también en la formación de los precios. La incorporación de variables exógenas en modelos autorregresivos (por ejemplo, mediante estructuras ARX o SARIMAX) ha demostrado mejorar la capacidad predictiva en otros contextos [5]. Aunque herramientas como las redes neuronales recurrentes pueden alcanzar mayor precisión, su naturaleza de “caja negra” limita la interpretabilidad, la cual es crucial para la toma de decisiones en el ámbito regulatorio [6].

Por ello, surge la necesidad de desarrollar un modelo predictivo que, integrando tanto datos históricos del SIMEM como variables exógenas relevantes, permita estimar de manera precisa el precio promedio diario de bolsa en el mercado eléctrico colombiano. Este enfoque facilitará la anticipación de fluctuaciones en el mercado y aportará información valiosa para la formulación de políticas públicas y la planificación del sistema eléctrico nacional.

En consecuencia, la pregunta que guía este estudio es:

¿Cómo influye la incorporación de variables exógenas en la capacidad predictiva de los modelos autorregresivos para estimar el precio promedio diario de bolsa en el mercado eléctrico colombiano?

1.2. Objetivos

1.2.1. Objetivo General

Diseñar un modelo de predicción del precio promedio diario de bolsa en el mercado eléctrico colombiano, utilizando modelos autorregresivos de series de tiempo que integren tanto variables endógenas como exógenas.

1.2.2. Objetivos Específicos

- Analizar los descriptores clave del mercado eléctrico colombiano, como el precio de bolsa, la demanda y la generación por fuente, a partir de sus series de tiempo históricas diarias, considerando el preprocesamiento de los datos y la estadística descriptiva para la identificación de patrones estacionales y de tendencia.
- Implementar un modelo autorregresivo de series de tiempo para predecir el precio promedio diario de bolsa, integrando información histórica y variables externas a partir de los datos del SIMEM.
- Evaluar la capacidad predictiva de los modelos AR y ARX mediante las métricas estándar RMSE, MAE y MAPE, considerando además el impacto relativo de cada variable exógena en la precisión del pronóstico a través de un análisis de sensibilidad One-At-a-Time (OAT).

1.3. Justificación

La integración de energías renovables en Colombia ha experimentado un crecimiento acelerado en los últimos años, pasando de 28 MW en 2018 a más de 2.500 MW en 2023, con proyecciones que indican que para 2028 la capacidad instalada podría alcanzar los 6.000 MW, lo que representaría aproximadamente el 12 % de la matriz energética nacional [7]. Este notable avance ha contribuido a la diversificación del sistema eléctrico y a la reducción del impacto ambiental asociado a los combustibles fósiles. Sin embargo, el carácter intermitente de fuentes como la solar y la eólica genera desafíos en términos de estabilidad operativa y económica del mercado eléctrico.

En regiones de alto potencial, como La Guajira, la variabilidad diaria en la generación puede superar el 30 % durante los períodos de transición climática [3]. Estos extremos, que se manifiestan como excedentes energéticos en días soleados y déficits durante las horas de baja irradiación o en la ausencia de viento, provocan desequilibrios entre la oferta y la demanda. Dichos desequilibrios tienen un impacto directo en el precio promedio diario de bolsa en el mercado eléctrico, generando oscilaciones que afectan tanto la rentabilidad de los generadores como la estabilidad de las tarifas para los consumidores. En 2022, por ejemplo, se estimaron pérdidas económicas cercanas a los USD 15 millones atribuibles a estas fluctuaciones [4].

Frente a este escenario, resulta evidente la necesidad de mejorar los mecanismos predictivos que permitan anticipar la dinámica del precio en el mercado eléctrico. Los modelos autorregresivos han sido tradicionalmente utilizados para pronosticar series de tiempo basándose en la información histórica (modelos endógenos). No obstante, limitarse a esta modalidad puede resultar insuficiente, ya que variables externas—como las condiciones meteorológicas, las variaciones en la demanda o indicadores económicos generales—también inciden de manera significativa en la formación de precios.

La incorporación de variables exógenas en los modelos autorregresivos (a través de enfoques ARX o estructuras SARIMAX) ofrece una herramienta más robusta para capturar la complejidad del sistema. Por ejemplo, estudios internacionales han demostrado que estos modelos pueden reducir el error de predicción (RMSE) en hasta un 12 % al incluir variables económicas y meteorológicas [5]. Además, si bien alternativas como las redes neuronales recurrentes pueden lograr una mayor precisión en ciertos escenarios [8], su naturaleza de "caja negra" limita la interpretabilidad, lo cual es fundamental para la toma de decisiones en el ámbito regulatorio y para el diseño de estrategias de mitigación [6].

En el contexto colombiano, se evidencia que las herramientas actuales disponibles en el SIMEM para gestionar la volatilidad de los precios diarios son limitadas. El Plan de Expansión 2022-2036 [9] y el Atlas de Viento y Radiación del IDEAM [10] señalan la necesidad de aprovechar de manera integral la información histórica para mejorar la capacidad predictiva. La experiencia internacional, como la aplicada en EE.UU. con modelos ARX para escenarios complejos [11], refuerza la oportunidad de adaptar estos enfoques al mercado colombiano.

Con estos antecedentes, se plantea desarrollar un modelo predictivo basado en técnicas autorregresivas que integre tanto los datos históricos del SIMEM como las variables exógenas seleccionadas, con el objetivo de estimar de manera precisa el precio promedio diario de bolsa en el mercado eléctrico.

Esta propuesta busca no solo mejorar la precisión de los pronósticos, sino también proporcionar una herramienta de gran utilidad para la formulación de políticas públicas y la planificación estratégica del sistema eléctrico nacional.

1.4. Impacto social

El presente proyecto de grado, tiene un impacto social significativo, en tanto contribuye a mejorar la transparencia y eficiencia del sistema eléctrico nacional a través del fortalecimiento de herramientas predictivas. Al centrarse en el análisis y modelado del comportamiento del precio promedio diario de bolsa mediante técnicas autorregresivas con variables endógenas y exógenas, el proyecto permite anticipar comportamientos del mercado que afectan directamente tanto a generadores como a usuarios finales.

Este trabajo se alinea con diversos Objetivos de Desarrollo Sostenible (ODS), en particular con el ODS 7 (Energía asequible y no contaminante), al apoyar la planificación y toma de decisiones dentro del mercado eléctrico colombiano. Una predicción más precisa de los precios diarios de bolsa puede favorecer la integración eficiente de diferentes tecnologías de generación eléctrica—incluidas las renovables—y mejorar la gestión de la demanda, contribuyendo a la estabilidad del suministro y al acceso a una energía más económica, especialmente en regiones con alta vulnerabilidad económica.

En relación, el proyecto responde al ODS 9 (Industria, innovación e infraestructura), al fomentar el uso de herramientas analíticas modernas como los modelos AR y ARX, que han demostrado su utilidad en diversos contextos internacionales. El desarrollo de metodologías de pronóstico basadas en datos históricos permite optimizar el despacho de energía y reducir la incertidumbre frente a escenarios de variabilidad, generando beneficios tanto para los operadores del sistema como para los reguladores.

Además, tiene relación con el ODS 13 (Acción por el clima), ya que una mejor comprensión y gestión del mercado eléctrico favorece la integración de tecnologías de generación más limpias, reduciendo la dependencia de fuentes fósiles y, por ende, las emisiones de gases de efecto invernadero. Este tipo de herramientas apoya la transición hacia un sistema energético más sostenible y resiliente, en línea con los compromisos climáticos asumidos por el país.

Finalmente, al proporcionar una mejor comprensión del vínculo entre el comportamiento de la oferta energética y la formación de precios, el proyecto puede tener efectos indirectos sobre decisiones relacionadas con tarifas, subsidios y acceso equitativo al servicio eléctrico. Esto representa una oportunidad para diseñar políticas públicas más justas y adaptadas a las condiciones del mercado, con beneficios tanto técnicos como sociales para el conjunto de la población colombiana [12].

Capítulo 2

Estado del arte

La creciente participación de energías renovables en los sistemas eléctricos ha impulsado el desarrollo de herramientas predictivas que permitan anticipar su generación y entender su impacto en variables económicas, como el precio de la energía. En este contexto, los modelos de series de tiempo, particularmente los modelos autorregresivos con y sin variables exógenas (AR y ARX), han cobrado relevancia como herramientas accesibles, explicables y eficaces para abordar estos desafíos. Esta sección presenta una revisión crítica del estado del arte sobre: 1) Modelos autorregresivos aplicados a la predicción de generación renovable; 2) Relación entre la generación renovable y el precio de bolsa de la energía; 3) Estudios relevantes en América Latina y Colombia; y 4) Brechas identificadas y contribución del presente trabajo.

2.1. Modelos Autorregresivos para la Predicción de Energías Renovables

Los modelos autorregresivos (AR) y los autorregresivos con variables exógenas (ARX) son herramientas estadísticas que permiten modelar una variable en función de sus valores pasados y, en el caso del ARX, también considerando variables externas que influyen en su evolución. Estos modelos se caracterizan por su bajo costo computacional y su capacidad para ofrecer interpretaciones claras sobre las relaciones temporales presentes en los datos.

De manera conceptual, un modelo AR se enfoca en predecir una variable, como la generación solar o eólica, a partir de su propio historial reciente. Por su parte, el modelo ARX amplía esta idea al incorporar variables externas, tales como la temperatura, la radiación o la velocidad del viento, lo que permite capturar tanto la dinámica interna del sistema como la influencia de factores externos sobre él. Esta estructura es especialmente útil en sistemas energéticos donde la variabilidad meteorológica afecta directamente la producción de energía.

Diversos estudios han demostrado la efectividad de estos enfoques. Por ejemplo, Zhang, Wang y Liu [13] utilizaron modelos AR para predecir la generación solar en distintos países europeos, alcanzando un error medio absoluto (MAE) inferior al 9 %. De forma similar, Chen, Wang y Li [14] mostraron que al incorporar variables meteorológicas, el modelo ARX mejora la precisión de la predicción en sistemas fotovoltaicos, reduciendo el error en más de un 10 %.

A diferencia de los trabajos citados, el presente proyecto aplica modelos AR y ARX específicamente para pronosticar el precio promedio diario de bolsa en el mercado eléctrico colombiano, integrando variables exógenas como generación renovable (solar, eólica y PCH), demanda eléctrica y factores climáticos locales. Si bien modelos más complejos, como las redes neuronales recurrentes (RNN) o los enfoques híbridos, han demostrado mejorar la precisión [8], los modelos autorregresivos mantienen la ventaja de ser explicables y transparentes, cualidad esencial en contextos regulatorios y de toma de decisiones informadas [15].

2.2. Generación Renovable e Impacto en el Precio de la Energía

La integración de energías renovables puede generar impactos significativos en los precios de mercado, especialmente en sistemas con alta penetración renovable. Este fenómeno es conocido como el “efecto de canibalización”, donde la generación abundante en ciertos momentos reduce el precio de bolsa [16].

Estudios como el de Wang, Chen y Kang [17] mostraron que una mayor generación solar tiende a desplazar tecnologías más costosas, reduciendo el precio marginal horario. Esta correlación negativa fue también observada en el mercado eléctrico español por González, Pérez y Ruiz [18], quienes utilizaron modelos ARX para vincular generación eólica con precios horarios, alcanzando coeficientes de correlación superiores a 0.7 en algunos nodos.

En el caso colombiano, García, López y Pérez [19] realizaron un análisis exploratorio de la relación entre generación hidroeléctrica y precios en bolsa, encontrando una sensibilidad significativa del precio frente a variaciones en la oferta hídrica, aunque sin utilizar modelos predictivos específicos.

2.3. Aplicaciones en América Latina y Colombia

Diversos estudios han aplicado modelos autorregresivos en contextos latinoamericanos con condiciones climáticas y estructurales similares a Colombia. En Brasil, Khodayar, Wang y Wang [20] desarrollaron un modelo ARX para estimar la producción eólica diaria, obteniendo un coeficiente de determinación (R^2) de 0.89. En Chile Pitt, Dryzek y Ober [21] emplearon modelos AR para predecir generación solar en regiones desérticas, con resultados satisfactorios en horizontes de corto plazo.

En Colombia, aunque se han hecho avances en la caracterización de fuentes renovables, la aplicación de modelos predictivos aún es limitada. El estudio del potencial eólico en La Guajira realizado por García, López y Pérez [19] destaca una alta variabilidad diaria (hasta un 30 %), lo cual sugiere la necesidad

de herramientas predictivas robustas. Adicionalmente, el Plan de Expansión de Referencia Generación-Transmisión 2022–2036 de la UPME [9] propone escenarios de penetración renovable, pero sin un análisis detallado sobre su impacto en el mercado de bolsa.

2.4. Brechas y Contribución Propuesta

A partir de la revisión realizada, se identifican las siguientes brechas:

- **Baja aplicabilidad de modelos explicables en contextos tropicales:** Aunque los modelos AR/ARX han sido utilizados exitosamente en Europa y América del Norte, existe poca literatura enfocada en zonas con alta variabilidad climática como el trópico colombiano.
- **Débil vinculación entre predicción renovable y precio de energía:** Pocos estudios han abordado explícitamente cómo los cambios en la oferta renovable impactan el precio de bolsa en sistemas interconectados.
- **Ausencia de variables exógenas contextuales:** La mayoría de los estudios se enfocan en variables meteorológicas, sin incorporar elementos socioeconómicos o del comportamiento del mercado colombiano.

Este trabajo propone llenar estas brechas mediante el desarrollo de modelos AR y ARX adaptados a la predicción de generación renovable (solar, eólica e hidráulica secundaria) y su impacto en el precio de bolsa del Sistema Interconectado Nacional (SIN) de Colombia. La integración de variables exógenas como precipitación, temperatura y demanda energética busca mejorar la precisión del modelo, sin sacrificar la interpretabilidad. Así, se espera contribuir a una planificación energética más robusta y a la formulación de políticas públicas informadas.

Capítulo 3

Marco Teórico

3.1. Mercado eléctrico colombiano

El sistema eléctrico colombiano se caracteriza por una matriz energética predominantemente hidráulica, complementada por fuentes térmicas y, en menor proporción, por energías renovables no convencionales como la solar y la eólica. Esta estructura ha permitido mantener una matriz relativamente limpia en comparación con otros países de América Latina; sin embargo, también implica una elevada vulnerabilidad frente a la variabilidad climática. En particular, los fenómenos de El Niño reducen los aportes hídricos, obligando a incrementar la participación térmica y, en consecuencia, elevando el costo marginal de generación y el precio promedio diario de bolsa [22].

De acuerdo con informes recientes de XM (2025), durante los meses de junio y julio de dicho año el nivel útil agregado de los embalses del Sistema Interconectado Nacional (SIN) superó el 82 %, permitiendo mantener una oferta energética estable y precios de bolsa moderados. En contraste, en abril de 2024 el nivel de almacenamiento descendió al 28,7 %, situación que incrementó significativamente los costos de generación térmica y la volatilidad del mercado [22]. Este comportamiento evidencia la alta sensibilidad del precio de bolsa ante las variaciones hidrológicas, razón por la cual los modelos predictivos deben considerar la disponibilidad de embalses como una variable exógena clave.

En los últimos años, la incorporación de energías renovables no convencionales ha comenzado a modificar la composición del sistema eléctrico. La energía solar fotovoltaica y la eólica han mostrado un crecimiento sostenido, impulsado por la Ley 1715 de 2014 y las políticas de fomento del Ministerio de Minas y Energía, que promueven la diversificación de la matriz y la reducción de emisiones [23]. Aunque su participación aún es inferior al 15 %, los proyectos en desarrollo muestran una tendencia ascendente, principalmente en regiones como La Guajira y el Cesar, con potencial eólico y solar de alta radiación.

Estudios recientes destacan que la expansión de la generación renovable puede alterar la dinámica del precio marginal en mercados interconectados. Hirth [24] analizó el fenómeno conocido como *efecto de canibalización*, según el cual la abundancia temporal de generación solar o eólica de bajo costo desplaza

a las tecnologías convencionales, reduciendo los precios marginales en ciertos periodos del día. Este comportamiento también ha sido documentado en el contexto latinoamericano por González, Pérez y Ruiz [18], quienes señalan que una penetración renovable acelerada puede incrementar la volatilidad de precios y afectar la estabilidad del sistema eléctrico si no existen mecanismos de respaldo adecuados.

En el caso colombiano, la coexistencia entre una alta dependencia hidráulica, la incorporación progresiva de renovables y la necesidad de respaldo térmico crea un escenario complejo para la predicción del precio promedio diario de bolsa. La hidrología, los niveles de embalses y la penetración de nuevas fuentes renovables actúan como variables interdependientes que determinan la oferta efectiva y, por tanto, la señal de precios en el mercado mayorista. Comprender estas interacciones resulta fundamental para el desarrollo de modelos autorregresivos robustos, capaces de capturar tanto las dinámicas internas del sistema como la influencia de factores externos en la formación de precios.

3.1.1. Estructura general del mercado

El Sistema Interconectado Nacional (SIN) de Colombia está estructurado en tres niveles principales: generación, transmisión y distribución. En el nivel de generación conviven fuentes hidráulicas, térmicas y renovables no convencionales. La generación hidroeléctrica domina la matriz energética nacional, alcanzando alrededor del 65 % del total de la capacidad instalada [25].

Esta composición otorga al sistema una elevada dependencia climática, ya que las variaciones en el régimen hídrico impactan directamente la oferta de energía y, en consecuencia, los precios en el mercado mayorista. La operación del sistema se desarrolla bajo un esquema de mercado mayorista con despacho centralizado, administrado por XM, en el cual se realiza la programación horaria de la generación considerando las ofertas presentadas por los agentes [26].

De esta manera, el precio diario de bolsa —también denominado precio spot— se determina con base en el costo marginal de operación, que representa el valor de la última unidad de generación requerida para satisfacer la demanda eléctrica [27].

3.2. Aplicación de modelos ARIMA y SARIMA en mercados eléctricos

El uso de modelos ARIMA y SARIMA en el ámbito energético ha demostrado ser una herramienta eficaz para el pronóstico de precios, demanda y generación eléctrica en distintos horizontes temporales. Estos modelos han sido ampliamente empleados en mercados liberalizados debido a su capacidad para capturar la estructura temporal de las series y su relativa simplicidad de implementación. A diferencia de los métodos puramente estadísticos sin memoria, los modelos ARIMA y SARIMA incorporan la dependencia temporal de los datos, lo que resulta fundamental en mercados eléctricos donde los precios presentan alta autocorrelación y comportamiento estacional [28].

En el contexto internacional, Contreras et al. [29] aplicaron modelos ARIMA para predecir precios diarios de electricidad en los mercados de España e Inglaterra, obteniendo errores medios absolutos inferiores al 5 %. Los resultados mostraron que las series de precios eléctricos exhiben una fuerte dependencia autorregresiva y estacionalidad semanal, asociada a la variación del consumo entre días laborales y fines de semana. De manera similar, Taylor [30] empleó variantes SARIMA y métodos de suavizamiento exponencial para pronosticar la demanda eléctrica británica, demostrando que la incorporación de componentes estacionales mejora significativamente la precisión de los modelos tradicionales.

La literatura reciente también ha destacado la aplicación de modelos SARIMA en combinación con técnicas híbridas para mejorar la predicción en sistemas eléctricos complejos. Por ejemplo, Weron [28] señala que los modelos ARIMA y SARIMA constituyen una base esencial para comparar metodologías más sofisticadas como las redes neuronales recurrentes (RNN) y los modelos autorregresivos con variables exógenas (ARX), que incorporan información adicional como demanda, temperatura o generación renovable. Estos modelos híbridos permiten capturar tanto las relaciones lineales a corto plazo como las no lineales derivadas de las fluctuaciones del mercado.

En el caso colombiano, la aplicación de modelos ARIMA y SARIMA resulta particularmente relevante debido a la estacionalidad inherente de su sistema hidrodependiente. Los precios de bolsa están fuertemente influenciados por los ciclos hidrológicos, los niveles de embalses y la variabilidad de la demanda, factores que se reflejan en patrones recurrentes a lo largo del año. La modelación mediante SARIMA permite incorporar estos efectos estacionales y analizar la influencia de fenómenos como El Niño y La Niña, que alteran la disponibilidad hídrica y la oferta energética [22]. En consecuencia, el modelo SARIMA se presenta como una herramienta robusta para la predicción del precio promedio diario de bolsa, especialmente cuando se combina con información exógena que refleje la dinámica física del sistema eléctrico colombiano.

3.3. Fundamentación Matemática de los Modelos Autorregresivos

El comportamiento temporal del precio promedio diario de bolsa puede representarse mediante modelos de series de tiempo que capturan las dependencias entre valores pasados de la misma variable y, en algunos casos, la influencia de factores externos [31, 32]. En este contexto, los modelos autorregresivos constituyen una familia de herramientas estadísticas ampliamente utilizadas en sistemas energéticos debido a su capacidad para describir relaciones lineales y su fácil interpretación [33, 34].

Desde un punto de vista teórico, todos los modelos autorregresivos empleados en este proyecto (AR, ARX, SARIMA y SARIMAX) pueden considerarse casos particulares de la formulación general del modelo SARIMAX, en la cual ciertos componentes se anulan cuando no son requeridos [35]. Esta unificación permite mantener una estructura matemática coherente y facilita la comparación entre modelos con y sin variables exógenas.

3.3.1. Modelo Autorregresivo (AR)

El modelo autorregresivo de orden p , denotado como $AR(p)$, describe el valor actual de una serie y_t como una combinación lineal de sus p valores pasados más un término de error aleatorio ε_t [36, 32]. Esta relación puede expresarse matemáticamente como se muestra en la ecuación (3.1):

$$y_t = c + \sum_{i=1}^p \phi_i y_{t-i} + \varepsilon_t \quad (3.1)$$

En la ecuación (3.1):

- y_t : valor de la variable en el instante t (precio promedio diario de bolsa);
- c : término constante;
- ϕ_i : coeficientes autorregresivos que cuantifican la influencia de los rezagos;
- ε_t : término de error o ruido blanco, con media cero y varianza constante.

Este modelo asume que la serie es estacionaria, es decir, que sus propiedades estadísticas (media, varianza y autocorrelación) permanecen constantes en el tiempo [35]. En el ámbito energético, los modelos AR se han utilizado ampliamente para representar el comportamiento histórico de precios eléctricos y cargas de demanda debido a su simplicidad y buena capacidad de ajuste a series estacionarias [33].

3.3.2. Modelo Autorregresivo con Variables Exógenas (ARX)

El modelo ARX amplía el $AR(p)$ al incorporar variables externas $x_{j,t}$ que influyen sobre la dinámica de y_t [37]. Su formulación general es (3.2)

$$y_t = c + \sum_{i=1}^p \phi_i y_{t-i} + \sum_{j=1}^k \beta_j x_{j,t} + \varepsilon_t \quad (3.2)$$

donde:

- β_j : coeficientes que representan el efecto de la variable exógena $x_{j,t}$,
- k : número total de variables exógenas.

En este modelo, las variables externas pueden corresponder a factores como la demanda eléctrica, la generación solar o el precio del gas natural. El término autorregresivo conserva la memoria del sistema, mientras que las variables exógenas introducen información contextual que puede mejorar la capacidad predictiva.

3.3.3. Modelo SARIMA

El modelo SARIMA (Seasonal Autoregressive Integrated Moving Average) incorpora tres componentes adicionales al modelo AR: la diferenciación (d) para eliminar tendencias, la parte de promedio móvil (q) para modelar la correlación de los errores, y los términos estacionales (P, D, Q, s) que permiten capturar patrones repetitivos a intervalos regulares (por ejemplo: semanal o anual) [36, 35]. Su estructura general se muestra en la ecuación (3.3):

$$\Phi_P(L^s) \phi_p(L) (1 - L)^d (1 - L^s)^D y_t = \Theta_Q(L^s) \theta_q(L) \varepsilon_t \quad (3.3)$$

En la ecuación (3.3):

- L : operador de rezago ($L^k y_t = y_{t-k}$);
- $\phi_p(L)$: polinomio autorregresivo de orden p ;
- $\theta_q(L)$: polinomio de promedio móvil de orden q ;
- $\Phi_P(L^s)$ y $\Theta_Q(L^s)$: componentes autorregresivos y de promedio móvil estacionales de orden P y Q , respectivamente;
- d y D : órdenes de diferenciación no estacional y estacional;
- s : periodicidad estacional (por ejemplo, $s = 7$ para reflejar ciclos semanales).

El modelo SARIMA resulta adecuado para series como el precio de bolsa, donde se presentan fluctuaciones cíclicas semanales y tendencias de largo plazo asociadas a la demanda y la hidrología [38]. Su capacidad para capturar patrones estacionales lo convierte en una herramienta estándar para el análisis y pronóstico de variables energéticas dependientes del tiempo.

3.3.4. Modelo SARIMAX (Generalización de la Familia Autorregresiva)

El modelo SARIMAX (Seasonal ARIMA with eXogenous variables) combina todas las propiedades anteriores en una única estructura, integrando componentes autorregresivos, estacionales, diferenciales y exógenos [31]. La formulación general del modelo se presenta en la ecuación (3.4):

$$\Phi_P(L^s) \phi_p(L) (1 - L)^d (1 - L^s)^D y_t = c + \Theta_Q(L^s) \theta_q(L) \varepsilon_t + \sum_{j=1}^k \beta_j x_{j,t} \quad (3.4)$$

En la ecuación (3.4), el término $\sum_{j=1}^k \beta_j x_{j,t}$ representa la influencia de las k variables exógenas, donde β_j son los coeficientes asociados a cada variable $x_{j,t}$. Esta estructura permite capturar tanto los efectos internos de la serie (autorregresivos y estacionales) como los efectos externos que influyen en la variable dependiente.

Además, esta formulación es lo suficientemente flexible para representar distintos modelos en función de los parámetros que se anulen. Por ejemplo:

- Si $q = P = D = Q = 0$, el modelo se reduce a un ARX.
- Si además no hay variables exógenas ($\beta_j = 0$), se obtiene un AR.
- Si $k = 0$ pero los términos estacionales permanecen, se obtiene un SARIMA.

En consecuencia, el SARIMAX constituye el marco unificador de la familia autorregresiva, y es el modelo que se empleará como base conceptual en este proyecto, ya que permite capturar simultáneamente:

1. La memoria temporal de los precios (componente AR),
 Los patrones estacionales y la necesidad de diferenciar la serie cuando presenta no estacionalidad: el parámetro d controla el número de diferenciaciones regulares aplicadas para eliminar tendencias,
 D representa el número de diferenciaciones estacionales necesarias para remover patrones repetitivos, y
 s define la periodicidad del ciclo estacional (en este caso, $s = 7$ días corresponde a la estacionalidad semanal),
2. La influencia de factores externos (variables exógenas) [39].

3.3.5. Significado de los Parámetros

La formulación del modelo SARIMAX requiere la especificación de siete parámetros principales que definen la estructura de la parte autorregresiva, la diferenciación y los componentes estacionales. Cada uno de estos parámetros controla un aspecto particular de la dinámica temporal de la serie y su correcta selección es fundamental para lograr un ajuste representativo [36].

En la Tabla 1 se presenta la interpretación de cada parámetro, junto con su significado dentro del contexto de este proyecto, enfocado en la predicción del precio promedio diario de bolsa en el mercado eléctrico colombiano.

TABLA 1: Definición de los parámetros del modelo SARIMAX

Parámetro	Descripción	Interpretación en este proyecto
p	Orden autorregresivo	Número de rezagos del precio de bolsa que influyen directamente en su valor actual.
d	Orden de diferenciación	Veces que se diferencia la serie para eliminar tendencia y lograr estacionalidad.
q	Orden de promedio móvil	Número de rezagos del error incluidos para modelar la correlación de los residuos.
P	Orden autorregresivo estacional	Influencia de los rezagos estacionales (por ejemplo, una semana atrás).
D	Orden de diferenciación estacional	Veces que se diferencia la serie estacionalmente (cada s períodos).
Q	Orden de promedio móvil estacional	Controla la correlación de los errores estacionales.
s	Periodicidad estacional	Longitud del ciclo repetitivo (7 días en este estudio).

Fuente: Autores.

La adecuada selección de estos parámetros suele realizarse mediante análisis de autocorrelación (ACF) y autocorrelación parcial (PACF), o por optimización de criterios de información como el AIC o el BIC, que equilibran la precisión del modelo y su complejidad. En este sentido, los parámetros p, d, q capturan la dinámica no estacional, mientras que P, D, Q, s describen las fluctuaciones recurrentes asociadas a la naturaleza cíclica del sistema eléctrico [38, 34].

Capítulo 4

Desarrollo Conceptual

4.1. Descripción del dataset

El conjunto de datos empleado en el presente trabajo proviene del SIMEM, administrado por XM. Esta plataforma constituye la fuente oficial de información técnica, comercial y operativa del Sistema Interconectado Nacional (SIN), siendo el repositorio más confiable para el estudio del comportamiento energético en el país.

Para la construcción del modelo predictivo se realizó una búsqueda dirigida dentro del SIMEM, aplicando filtros y palabras claves asociadas a las variables de interés del mercado mayorista. El conjunto final de datos corresponde exclusivamente al periodo comprendido entre 2024/01/01 - 2024/12/31 tomando 366 muestras por variable, el cual fue seleccionado por su relevancia temporal y su capacidad para reflejar las condiciones actuales de operación del sistema eléctrico colombiano.

Los archivos descargados se encontraban organizados por tipo de variable y en formato *.CSV*, por lo que fue necesario integrarlos en un único dataset coherente. Este conjunto mantiene una frecuencia diaria y un alcance nacional, representando el comportamiento agregado del sistema eléctrico en términos de precios, demanda y generación. Este enfoque diario permite capturar la dinámica operativa del mercado sin considerar subdivisiones regionales o por tipo de planta. A continuación, en la Tabla 2, se presentan las variables de referencia encontradas en el SIMEM. Dichas variables conforman el marco de información sobre el cual se basa la conceptualización del modelo de predicción, ya que reflejan la interacción entre oferta, demanda y condiciones de generación dentro del Sistema Interconectado Nacional.

TABLA 2: Variables de referencia disponibles en el SIMEM y su correspondencia en el dataset

Variable	Descripción general	Nombre en el dataset
Precio de bolsa	Valor promedio diario de la energía transada en el mercado mayorista colombiano.	PrecioBolsa
Demanda de energía	Cantidad total de energía consumida diariamente en el Sistema Interconectado Nacional.	Demanda
Participación hidroeléctrica	Porcentaje de la generación total del sistema que proviene de fuentes hidráulicas.	Participacion_Hidro
Volumen de embalses	Nivel agregado de almacenamiento de agua en los embalses del sistema, medido en porcentaje de su capacidad útil.	VolumenEmbalses
Generación térmica	Energía diaria producida por centrales térmicas, principalmente a gas natural.	GenTermica
Generación solar	Producción diaria proveniente de fuentes fotovoltaicas.	GenSolar
Generación eólica	Energía generada por fuentes eólicas dentro del sistema.	GenEolica
Precio del gas natural	Referencia diaria del costo del gas como insumo de generación térmica.	PrecioGas

Fuente: Autor.

Estas variables constituyen la base conceptual del estudio, ya que representan los factores más influyentes en la formación del precio de bolsa y en el equilibrio del sistema eléctrico colombiano. Su análisis teórico permite comprender las relaciones entre los componentes del mercado y sustentar la elección posterior de las variables que se utilizarán en las fases de modelado y validación.

Previo al análisis exploratorio, los datos se sometieron a una fase de Preprocesamiento, orientada a garantizar la coherencia y consistencia de las observaciones diarias. Este proceso incluyó la verificación de fechas, la detección de valores faltantes.

4.2. Caracterización estadística y relacional de las variables

Con el propósito de comprender la estructura interna del conjunto de datos y evaluar la coherencia entre las variables disponibles, se desarrolló una fase de caracterización estadística y relacional. Esta etapa constituye un componente esencial del desarrollo conceptual, ya que permite definir, desde un

punto de vista analítico, el grado de dependencia, redundancia y relevancia potencial de cada variable dentro del sistema eléctrico colombiano.

De manera conceptual, el análisis se orientó a examinar las relaciones entre la variable endógena el precio promedio diario de bolsa y las variables exógenas registradas en el SIMEM. Para ello, se consideraron tanto las interacciones lineales como las no lineales, buscando una representación más completa de las dinámicas presentes en el mercado.

Adicionalmente, se generaron las representaciones de las series de tiempo correspondientes a cada variable, con el fin de observar la existencia de patrones estacionales, tendencias y fluctuaciones abruptas. Este proceso no constituye una fase de modelado, sino de comprensión del comportamiento temporal previo a la selección de variables.

4.2.1. Medidas de dependencia y correlación

En esta etapa se emplearon distintos enfoques conceptuales para analizar la relación entre variables, combinando métricas lineales y no lineales que permitieron identificar tanto dependencias directas como asociaciones complejas entre las series temporales consideradas:

- **Coeficiente de correlación de Pearson:** mide la relación lineal entre dos variables continuas. Su valor varía entre -1 y 1 , donde los extremos indican correlación negativa o positiva perfecta, y los valores cercanos a cero reflejan independencia lineal [40]. Matemáticamente se define como se muestra en la ecuación (4.1)

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (4.1)$$

donde x_i y y_i son las observaciones de cada variable, y $\bar{x}$, $\bar{y}$ sus medias respectivas. En este proyecto se utilizó para identificar asociaciones directas o inversas entre el precio de bolsa y las demás variables energéticas.

- **Información mutua:** cuantifica la dependencia no lineal entre dos variables, midiendo la cantidad de información compartida entre ellas [41]. A diferencia de Pearson, no asume linealidad. Su expresión general es (4.2)

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \quad (4.2)$$

donde $p(x, y)$ es la distribución conjunta de X y Y , y $p(x)$, $p(y)$ sus distribuciones marginales. Esta métrica permite detectar relaciones complejas entre variables cuya dependencia no puede ser descrita mediante correlación tradicional.

- **Matriz de correlación:** presenta de forma matricial los coeficientes de correlación entre todas las variables del conjunto de datos, facilitando la detección de redundancias o colinealidades entre los predictores [35]. Su análisis orienta la selección posterior de variables relevantes para el modelado.

- **Análisis intercuartílico (IQR):** se basa en el rango intercuartílico, definido como la diferencia entre el tercer y el primer cuartil de una distribución [42]. Un valor se considera atípico si cumple estos parámetros (4.3)

$$x_i < Q_1 - 1,5(IQR) \quad \text{o} \quad x_i > Q_3 + 1,5(IQR) \quad (4.3)$$

donde Q_1 y Q_3 representan los cuartiles primero y tercero. En este proyecto se aplicó para detectar registros anómalos o inconsistentes que pudieran distorsionar los resultados del modelo predictivo.

La comparación entre las medidas de Pearson e información mutua permitió estructurar un criterio conceptual sobre la fuerza y naturaleza de las relaciones entre variables. Mientras Pearson refleja dependencias lineales de corto alcance, la información mutua introduce una perspectiva complementaria capaz de revelar comportamientos más complejos. Esta combinación de métricas ofrece una visión integral de la estructura estadística del sistema eléctrico colombiano y sienta las bases para la definición final de las variables a emplear en el modelado predictivo.

4.3. Preparación para el modelado

Con base en la formulación matemática de los modelos autorregresivos presentada en el Capítulo 3, y particularmente en la Ecuación (3.4) correspondiente al modelo SARIMAX, se estableció el procedimiento conceptual para la preparación de los datos previos al entrenamiento. Esta etapa tiene como propósito garantizar que las series temporales cumplan con las condiciones estadísticas necesarias para su incorporación en los modelos autorregresivos y sus variantes con variables exógenas.

El proceso de preparación se estructuró en tres componentes principales: la generación de rezagos, la partición temporal de los datos y la definición del horizonte de predicción. Cada uno de estos pasos responde a los principios de estacionalidad y dependencia temporal que rigen el comportamiento de los modelos de la familia AR.

4.3.1. Generación de rezagos y estructura temporal

Los modelos autorregresivos se fundamentan en la idea de que los valores pasados de una serie contienen información relevante sobre su comportamiento futuro. Por tanto, la preparación conceptual incluyó la creación de rezagos (*lags*) de la variable endógena el precio promedio diario de bolsa de acuerdo con el orden p definido en los modelos AR y SARIMAX tal como se muestra Ecuación (3.4).

Cada rezago representa la influencia de un día anterior sobre el valor actual, permitiendo capturar la inercia o persistencia temporal del sistema. Este procedimiento se extendió de manera conceptual a las

variables exógenas, que también fueron desplazadas temporalmente cuando se requería evaluar su efecto con desfase respecto al precio de bolsa. De esta forma, se construyó una estructura de datos coherente con la dinámica temporal del mercado eléctrico.

4.3.2. Transformaciones y estandarización

Con el fin de garantizar la comparabilidad entre variables y mejorar la convergencia de los procedimientos de estimación, las series fueron sometidas a transformaciones y normalizaciones conceptuales previas al modelado. Entre las transformaciones consideradas se incluyeron transformaciones polinómicas cuando la relación observada con la variable objetivo mostró no linealidad (por ejemplo, efectos cuadráticos en la demanda o sensibilidad asimétrica en los niveles bajos del volumen de embalses), y transformaciones de escala para homogeneizar unidades y varianzas entre predictores.

Para la estandarización se empleó la estandarización (z-score), que transforma cada observación restando la media de la variable y dividiendo por su desviación estándar. Esta operación se define matemáticamente como:

$$z_i = \frac{x_i - \mu_x}{\sigma_x} \quad (4.4)$$

donde x_i es el valor original de la variable, μ_x su media muestral y σ_x su desviación estándar. La estandarización asegura que las variables tengan media cero y varianza unitaria, evitando que las magnitudes diferentes dominen el proceso de ajuste del modelo y permitiendo una interpretación más estable de los coeficientes en modelos lineales. En la fase experimental se hace referencia a esta transformación mediante la ecuación (4.4).

4.3.3. División temporal de los datos

Para evaluar la capacidad de generalización de los modelos, el conjunto de datos se dividió de forma secuencial en dos subconjuntos: entrenamiento y prueba. Este enfoque respeta la naturaleza temporal de las observaciones, evitando la mezcla de registros del pasado y del futuro. Conceptualmente, el modelo se ajusta únicamente con la información del periodo de entrenamiento y se valida con el conjunto de prueba, que representa días posteriores no observados durante el ajuste.

Esta división permite evaluar si la estructura definida por los parámetros p, d, q, P, D, Q, s del modelo SARIMAX (Capítulo 3) logra capturar la evolución real de la serie y sus relaciones con las variables exógenas. En esta fase aún no se ejecutan estimaciones ni métricas, sino que se definen las condiciones lógicas para su futura aplicación.

4.3.4. Horizonte de predicción

De acuerdo con los patrones estacionales identificados en el análisis previo de las series de tiempo, se estableció un horizonte de predicción de siete días. Conceptualmente, esta ventana temporal permite capturar la estacionalidad semanal inherente al comportamiento del sistema eléctrico colombiano, donde la demanda y la generación presentan ciclos de variación recurrentes.

El horizonte de predicción define el número de pasos futuros que el modelo intentará estimar a partir de los valores históricos. Este parámetro es especialmente relevante en modelos SARIMA y SARIMAX, ya que afecta directamente la configuración de los componentes estacionales P, D, Q, s definidos en la formulación general del modelo definido matemáticamente en la Ecuación (3.4).

4.3.5. Condiciones de validación conceptual

Finalmente, se establecieron las condiciones conceptuales para la validación posterior de los modelos, sin realizar aún pruebas cuantitativas. Estas condiciones incluyen la necesidad de verificar la estacionalidad de las series (condición básica para la aplicación del modelo AR), la ausencia de autocorrelación significativa en los residuos y la consistencia entre la dinámica temporal de las variables exógenas y el precio de bolsa.

El cumplimiento de estas condiciones permitirá que, en la etapa de resultados, los modelos AR, ARX, SARIMA y SARIMAX puedan ser evaluados en igualdad de condiciones, siguiendo los principios teóricos establecidos en el marco metodológico. De esta forma, la presente sección define la base conceptual del proceso de entrenamiento y validación que se desarrollará experimentalmente en el siguiente capítulo.

4.4. Organización del proceso de modelado

Una vez definidas las condiciones conceptuales de preparación de los datos, se estructuró el proceso general de modelado de acuerdo con una secuencia lógica de etapas interrelacionadas. Este diseño metodológico tiene como propósito mantener la coherencia entre la formulación teórica de los modelos autorregresivos y su futura implementación práctica, garantizando la trazabilidad de cada fase del análisis.

El proceso se organizó bajo un enfoque modular, donde cada componente cumple una función específica dentro del flujo predictivo. Esta organización permite mantener la independencia entre etapas y facilita la validación individual de los procedimientos. Conceptualmente, el flujo de modelado se compone de cuatro fases principales: adquisición y preparación de datos, construcción del conjunto de entrenamiento, estimación de parámetros y evaluación conceptual del desempeño.

4.4.1. Adquisición y preparación de datos

En esta primera fase se consolida la información proveniente del SIMEM, la cual fue descrita en la Sección 4.1. El objetivo de esta etapa es garantizar que las series de tiempo empleadas en los modelos presenten integridad, coherencia y consistencia temporal. Conceptualmente, esta fase representa el punto de partida del flujo de modelado, pues de la calidad de los datos depende la estabilidad de las estimaciones posteriores.

4.4.2. Construcción del conjunto de entrenamiento

En esta fase se define la porción de la serie temporal que será utilizada para ajustar los modelos AR, ARX, SARIMA y SARIMAX. Como se explicó en la Sección 4.3, la división temporal se realiza respetando el orden cronológico de los registros, de modo que la información del pasado se utilice para predecir valores futuros. Conceptualmente, esta etapa establece las bases para la calibración de los parámetros autorregresivos, de diferenciación y de promedio móvil, definidos en la formulación general del modelo SARIMAX (Ecuación (3.4)).

4.4.3. Estimación y ajuste de parámetros

Esta etapa corresponde al núcleo del proceso de modelado. De manera conceptual, aquí se define la estrategia de estimación de los parámetros que describen la dinámica temporal de las series. Aunque en esta sección no se ejecutan los algoritmos, se establece el principio de búsqueda sistemática de combinaciones de órdenes (p, d, q, P, D, Q, s) para los modelos SARIMA y SARIMAX, así como de los coeficientes autorregresivos y exógenos en los modelos AR y ARX.

El objetivo conceptual de esta fase es garantizar que el proceso de ajuste preserve la estabilidad del modelo, cumpla con la condición de estacionalidad y mantenga la coherencia estadística entre las variables endógenas y exógenas.

4.4.4. Análisis conceptual de sensibilidad y métricas de error

En el marco del desarrollo conceptual, se incorpora el análisis de sensibilidad como herramienta para evaluar la respuesta de los modelos autorregresivos ante variaciones en las variables de entrada. Este análisis permite identificar el grado de influencia de cada variable exógena sobre la variable dependiente (precio de bolsa), así como determinar la estabilidad del modelo frente a perturbaciones o valores atípicos. De manera conceptual, el análisis de sensibilidad garantiza que el modelo sea interpretativo y no dependa excesivamente de una única fuente de información.

La sensibilidad puede evaluarse a través de técnicas estadísticas como el análisis AOT (Analysis of Outliers and Trends), el cual examina las variaciones anómalas y las tendencias residuales en las predicciones del modelo. Este procedimiento contribuye a validar la coherencia del modelo con el comportamiento real de la serie temporal, permitiendo detectar posibles desviaciones estructurales o efectos no capturados por los parámetros definidos.

Asimismo, para cuantificar el desempeño predictivo se definen las métricas de error que serán utilizadas en la etapa de validación experimental. Estas métricas permiten evaluar la precisión de las predicciones del modelo respecto a los valores reales y se presentan a continuación [35]:

- **Error absoluto medio (MAE, Mean Absolute Error):** mide el promedio de los errores absolutos entre los valores observados (y_i) y los estimados ($\hat{y}_i$). Se define como: (4.5)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4.5)$$

Representa la magnitud media del error sin considerar su signo, siendo útil para interpretar la desviación promedio del modelo.

- **Error cuadrático medio (MSE, Mean Squared Error):** calcula el promedio de los errores al cuadrado, penalizando con mayor peso las desviaciones grandes: (4.6)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (4.6)$$

Una menor magnitud del MSE indica un mejor ajuste del modelo a los datos reales.

- **Raíz del error cuadrático medio (RMSE, Root Mean Squared Error):** corresponde a la raíz cuadrada del MSE y expresa el error promedio en las mismas unidades de la variable analizada: (4.7)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4.7)$$

Esta métrica facilita la interpretación directa del desempeño del modelo, siendo una de las más utilizadas en estudios de series temporales y predicción energética.

4.4.5. Criterios de información: AIC y BIC

Para la selección entre modelos con diferente complejidad se emplearon criterios de información que penalizan el número de parámetros para evitar sobreajuste. Los más utilizados en la práctica son el *Akaike Information Criterion* (AIC) y el *Bayesian Information Criterion* (BIC), los cuales se definen a continuación [43, 44].

El AIC se expresa como:

$$AIC = 2k - 2\ln(\hat{L}) \quad (4.8)$$

donde k es el número de parámetros del modelo y $\hat{L}$ es el valor máximo de la función de verosimilitud del modelo. El AIC busca un equilibrio entre ajuste y complejidad; en la práctica, se prefiere el modelo con menor AIC.

El BIC (también llamado criterio de Schwarz) incorpora una penalización dependiente del tamaño de la muestra n :

$$\text{BIC} = k \ln(n) - 2 \ln(\hat{L}) \quad (4.9)$$

La penalización $k \ln(n)$ hace que BIC favorezca modelos más parcimoniosos que AIC cuando la muestra es grande. En consecuencia, para conjuntos de datos con n relativamente elevado, BIC tiende a seleccionar modelos con menor número de parámetros en comparación con AIC [44].

Ambos criterios son comparativos (no absolutos): para un conjunto de modelos candidatos se selecciona aquel con el valor mínimo de AIC o BIC. Es conveniente reportar ambos porque AIC y BIC pueden diferir en su preferencia según la prioridad que se dé a la capacidad predictiva (AIC) frente a la parsimonia (BIC). Cuando el tamaño muestral es pequeño y se busca corregir el sesgo de AIC, puede considerarse la versión corregida AICc [45].

Capítulo 5

Resultados y Discusión

En este capítulo se presentan y analizan los resultados obtenidos durante la implementación de la metodología propuesta. El contenido se organiza de manera que permita evidenciar el cumplimiento progresivo de los tres objetivos específicos del proyecto, integrando el análisis de las series de tiempo, la evaluación de los modelos predictivos y la interpretación de la influencia de las variables exógenas sobre el precio promedio diario de bolsa en el mercado eléctrico colombiano.

En primera instancia, se abordan los resultados correspondientes al análisis de los descriptores clave del mercado eléctrico colombiano, con el propósito de cumplir el primer objetivo específico. Para ello, se examinan las series de tiempo históricas del precio de bolsa, la demanda eléctrica y la generación por fuente (hidroeléctrica, térmica, solar y eólica), considerando su comportamiento estadístico, tendencias y patrones estacionales. Este análisis permitió identificar relaciones de dependencia y redundancia entre variables, determinando que algunas de ellas no aportaban información adicional al modelo debido a su alta correlación o a que explicaban parcialmente el comportamiento de otras variables. Como parte de este proceso, se realizaron descomposiciones de tendencia y estacionalidad, así como pruebas de correlación y causalidad de Granger, lo que permitió establecer un conjunto final de variables con coherencia temporal y relevancia explicativa.

Posteriormente, se presentan los resultados asociados a la implementación de los modelos autorregresivos de series de tiempo (AR, ARX, SARIMA y SARIMAX), en cumplimiento del segundo objetivo específico. En esta sección se exponen las métricas de desempeño obtenidas durante las etapas de entrenamiento y prueba, utilizando los indicadores RMSE, MAE y MAPE. Los resultados se acompañan de representaciones gráficas y tablas comparativas que facilitan la interpretación de las diferencias en la precisión de los modelos. Finalmente, se desarrolla la evaluación correspondiente al análisis de sensibilidad del tipo One-At-a-Time (OAT), orientada al cumplimiento del tercer objetivo específico. Este análisis permite cuantificar el impacto relativo de cada variable exógena sobre la capacidad predictiva del modelo, identificando cuáles variables contribuyen significativamente a la reducción del error y a la estabilidad del pronóstico.

5.1. Selección y depuración de variables

El proceso de selección de variables exógenas y endógenas se desarrolló de forma iterativa, combinando criterios estadísticos, análisis visual y coherencia física con el comportamiento del mercado eléctrico colombiano. Esta etapa corresponde al primer objetivo específico del proyecto y constituye el punto de partida para la construcción de los modelos autorregresivos descritos en los capítulos anteriores.

Con el fin de garantizar transparencia en el proceso de integración y depuración de datos, las Tablas 3 y 4 presentan la estructura del dataset consolidado utilizado en este estudio. La primera muestra las columnas finales obtenidas tras la unificación de fuentes del SIMEM, junto con sus tipos de datos; mientras que la segunda tabla incluye tres filas representativas del dataset ya depurado. Esta visualización permite evidenciar la coherencia temporal del conjunto de datos y sirve como referencia para comprender las transformaciones posteriores aplicadas durante el preprocesamiento.

TABLA 3: Columnas finales del dataset consolidado y tipos de datos.

Columna	Tipo	Ejemplo
Fecha	Fecha (YYYY-MM-DD)	2024-01-01
PrecioBolsa	Float (COP/kWh)	218.946013
Hidro	Float (GWh)	143.842058
Termica	Float (GWh)	25.116513
Demanda	Float (GWh)	176.174153

Fuente: Autor.

TABLA 4: Primeras filas del dataset consolidado utilizado para el modelado.

Fila	Fecha	PrecioBolsa	Hidro	Termica	Demanda
0	2024-01-01	218.95	143.84	25.12	176.17
1	2024-01-02	396.88	162.28	36.50	206.71
2	2024-01-03	460.64	165.10	40.75	214.85

Fuente: Autor.

5.1.1. Carga y filtrado inicial del dataset

Tal como se mencionó en el desarrollo conceptual, los datos utilizados provienen del SIMEM. La información fue descargada en formato .CSV y filtrada mediante palabras clave asociadas a las principales variables de interés: demanda eléctrica, generación por tipo de fuente y precios energéticos. Posteriormente, los datos fueron unificados y preparados para su análisis temporal.

5.1.2. Selección preliminar de variables

En una primera aproximación, se consideraron como variables exógenas la generación hidroeléctrica, la generación térmica y la demanda total de energía, mientras que la variable endógena correspondió al precio promedio diario de bolsa. A cada una de estas variables se le construyó su respectiva serie de tiempo para examinar su comportamiento histórico y coherencia temporal. Con el propósito de identificar relaciones de dependencia o redundancia entre ellas, se elaboró una matriz de correlación inicial que permitiera evaluar la posible presencia de multicolinealidad entre las variables de generación y la demanda eléctrica.

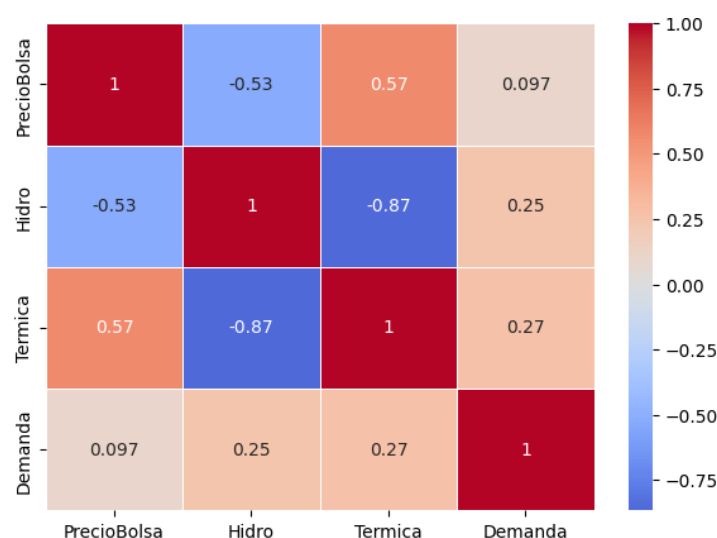


FIGURA 1: Matriz de correlación inicial entre la generación hidroeléctrica, térmica y la demanda eléctrica.

Fuente: Autor.

En la Figura 1 se presenta la matriz de correlación entre las variables seleccionadas. En ella se observa una fuerte relación inversa entre la generación hidroeléctrica y la generación térmica, con un coeficiente aproximado de $-0,87$. Este comportamiento refleja la dinámica operativa del Sistema Interconectado Nacional: durante los periodos de alta disponibilidad hídrica, la mayor participación de la generación hidroeléctrica reduce la necesidad de generación térmica, mientras que en épocas de menor disponibilidad de agua ocurre el efecto contrario. Por su parte, la demanda eléctrica muestra correlaciones débiles con ambas fuentes de generación (inferiores a $0,3$), lo que sugiere una relación no estrictamente lineal con la oferta de energía. En general, estos resultados evidenciaron un problema de multicolinealidad entre las fuentes de generación, razón por la cual fue necesario redefinir el conjunto de variables exógenas con el fin de evitar redundancias y mejorar la estabilidad e interpretación del modelo predictivo.

5.1.3. Selección y preprocesamiento de variables

Con el fin de realizar una nueva selección de variables, se consideró un conjunto actualizado de ellas, presentado en la Tabla 2. De este conjunto se seleccionaron las variables Demanda de energía, Volumen de los embalses y Participación hidroeléctrica, las cuales fueron sometidas al mismo proceso de análisis aplicado previamente, con el propósito de evaluar su coherencia temporal, relevancia explicativa y contribución al modelado del precio de bolsa.

Como primera medida se realiza la serie de tiempo para cada una de estas variables como se puede ver en la figura 2.

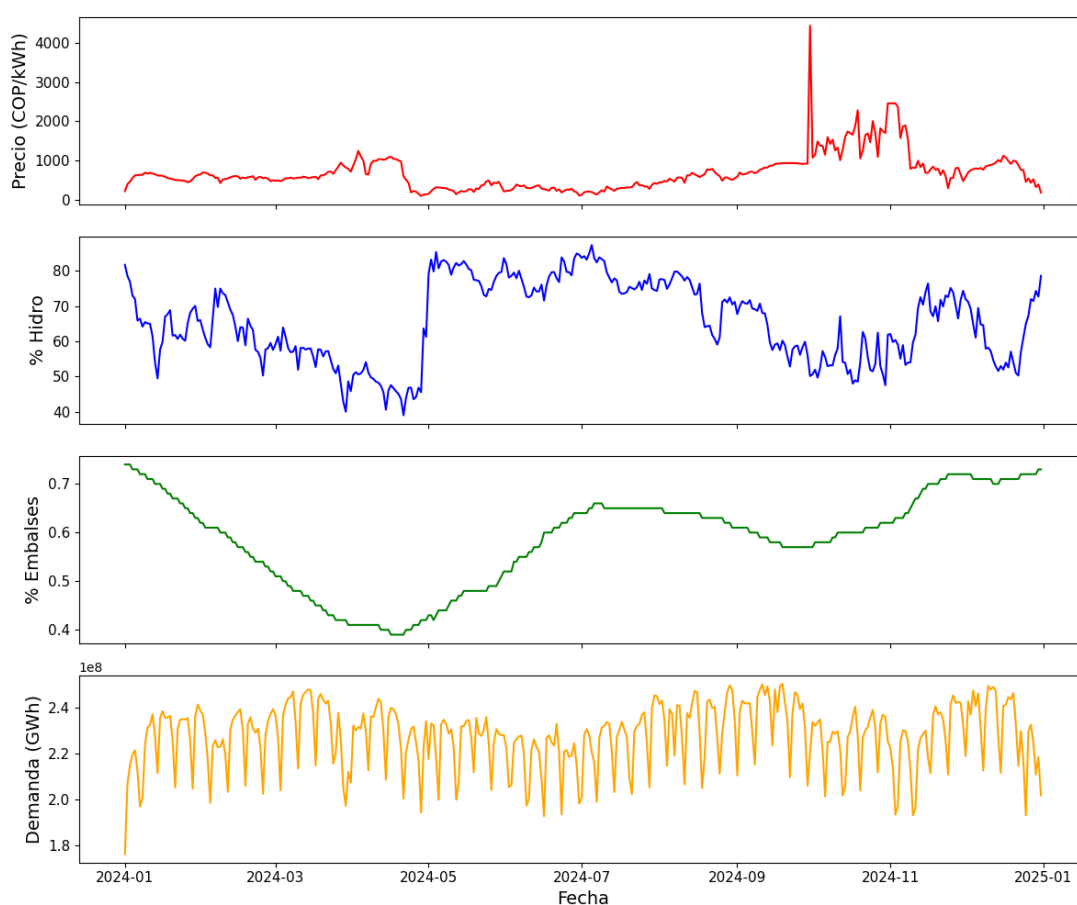


FIGURA 2: Series de tiempo de las variables finales seleccionadas para el modelado. (Precio en bolsa, participación hidroeléctrica, nivel embalses, demanda)

Fuente: Autor.

La Figura 2 presenta la evolución temporal de las variables consideradas en el modelo durante el periodo 2024/01/01–2024/12/31. Se observa que el precio de bolsa exhibe una marcada volatilidad, con picos significativos hacia el segundo semestre del año, coincidiendo con una reducción simultánea en la participación hidroeléctrica y en el volumen de los embalses. Este comportamiento evidencia la alta dependencia del mercado eléctrico colombiano respecto de las condiciones hidrológicas, donde una

menor disponibilidad hídrica incrementa los costos marginales de generación y, por ende, el precio de la energía. En contraste, cuando los niveles de los embalses se recuperan, la participación hidroeléctrica aumenta y el precio de bolsa tiende a estabilizarse. Por su parte, la demanda de energía mantiene un patrón estacional regular, con picos semanales asociados a los días laborables y descensos durante los fines de semana, mostrando una menor influencia directa sobre las variaciones abruptas del precio. En conjunto, este análisis confirma que las variables exógenas seleccionadas aportan información relevante para capturar tanto la estacionalidad como los choques externos del sistema, sin embargo para una ratificación de que estas se realiza la matriz de correlación como se puede ver en la figura 3.

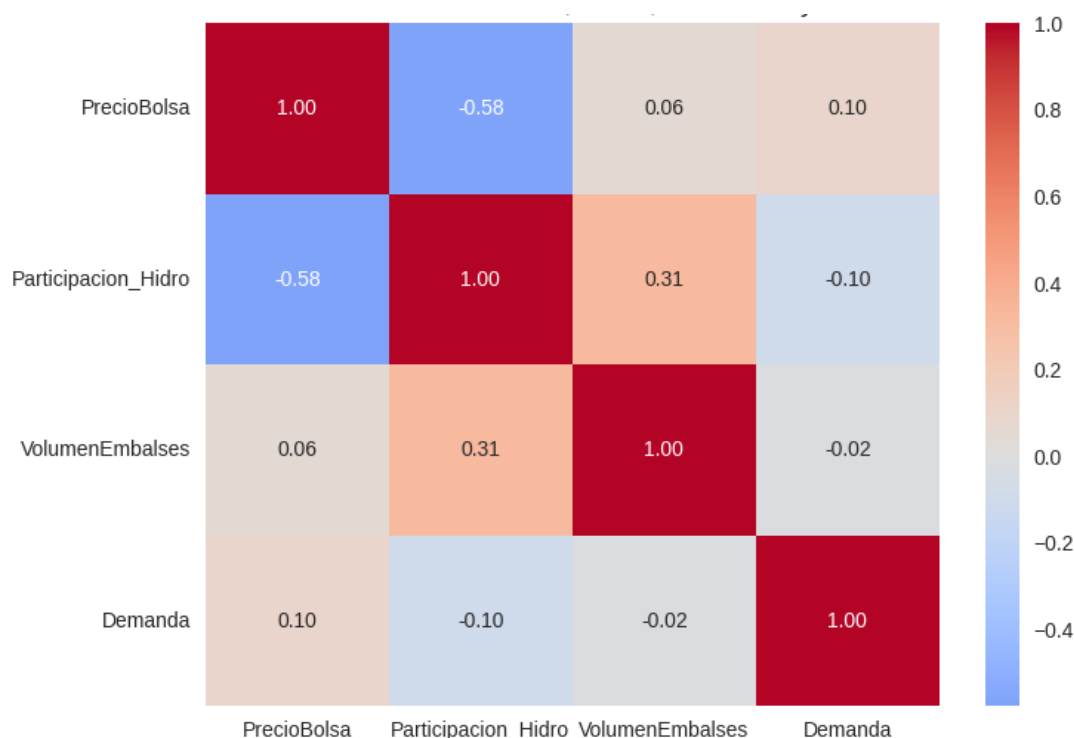


FIGURA 3: Matriz de correlación de la segunda iteración de selección de variables.

Fuente: Autor.

En la figura 3 presenta la matriz de correlación entre las variables incluidas en el modelo: Precio de Bolsa, Participación Hidroeléctrica, Volumen de Embalses y Demanda de Energía. A partir de esta representación se verifica cuantitativamente la relación entre los principales factores del sistema. En particular, se observa una correlación negativa moderada entre el precio y la participación hidroeléctrica ($-0,58$), lo que confirma la existencia de una dependencia inversa previamente identificada en el análisis temporal. De igual forma, la correlación positiva entre la participación hidroeléctrica y el volumen de los embalses ($0,31$) sugiere coherencia estructural entre ambas variables, pues reflejan un mismo componente hidrológico del sistema. Por otro lado, las correlaciones de la demanda con las demás variables resultan cercanas a cero, lo que indica una menor relación lineal directa y respalda su inclusión como variable exógena independiente dentro del modelo. En conjunto, esta matriz permitió ratificar la pertinencia de

las variables seleccionadas y descartar posibles redundancias o efectos de multicolinealidad antes de la etapa del preprocesamiento y modelado.

5.1.4. Preprocesamiento de las variables

Una vez definido el conjunto final de variables, se procedió con el preprocesamiento de los datos para garantizar su calidad y adecuación a los modelos autorregresivos. Este proceso incluyó técnicas de limpieza, transformación y preparación temporal específicas para series de tiempo.

5.1.4.1. Tratamiento de valores atípicos

Se implementó la técnica del Rango Intercuartílico (IQR) para el manejo de valores atípicos en las variables Precio en bolsa, Demanda, participación hidroeléctrica y volúmenes de los embalses. La metodología aplicada consistió en los siguientes pasos:

- Cálculo del primer y tercer cuartil (Q_1 = percentil 25, Q_3 = percentil 75);
- Determinación del rango intercuartílico como $IQR = Q_3 - Q_1$;
- Establecimiento de límites compactos para detección de outliers:

$$L_{inf} = Q_1 - 1,5 IQR, \quad L_{sup} = Q_3 + 1,5 IQR$$

Los valores x tales que $x < L_{inf}$ o $x > L_{sup}$ se consideraron atípicos;

- Aplicación de clipping controlado, complementado con un método de imputación basado en la mediana, reemplazando los valores extremos por el valor central de la distribución correspondiente. Este enfoque evitó distorsionar la tendencia general sin eliminar información relevante.

El análisis mediante el rango intercuartílico (IQR) permitió identificar y cuantificar la presencia de valores atípicos en las series utilizadas. En particular, la variable Precio en Bolsa presentó 23 observaciones fuera de los límites establecidos (equivalentes al 6,28 % del total), mientras que la Demanda de energía registró 7 valores atípicos (1,91 %). Por su parte, las variables VolumenEmbalses y Participación Hidroeléctrica no mostraron valores anómalos según el criterio aplicado. Tras la imputación mediante la mediana, se redujo el efecto distorsionador de estos extremos sin eliminar información relevante sobre eventos del sistema. Esta reducción en la cantidad de valores extremos y la estabilización visual posterior de las series confirman que el tratamiento aplicado mitigó efectivamente el impacto de los outliers y mejoró la consistencia estadística de los datos previos al modelado.

5.1.4.2. Pruebas de estacionalidad

Las pruebas de estacionalidad aplicadas revelaron características esenciales para la selección de modelos autorregresivos. Los resultados obtenidos se presentan en el tabla 5:

TABLA 5: Resultados de pruebas de estacionalidad

Variable	ADF (p-value)	KPSS (p-value)	Conclusión
PrecioBolsa	0.210	0.010	No estacionaria
Demanda	0.023	0.100	Estacionaria
Participación Hidro	0.015	0.050	Estacionaria
Volumen Embalses	0.008	0.100	Estacionaria

Fuente: Autor.

El precio de bolsa mostró no estacionalidad (p-value ADF >0.05), lo que justifica el uso de componentes de diferenciación en los modelos ARIMA y SARIMAX. Esta no estacionalidad se asocia principalmente con la presencia de tendencias de mediano plazo influenciadas por las condiciones hidrológicas —como la reducción progresiva del volumen de embalses durante el primer semestre— así como por ciclos semanales derivados de los patrones operativos del sistema eléctrico. Adicionalmente, los picos abruptos observados en el segundo semestre evidencian choques externos y variaciones estructurales en la oferta, los cuales refuerzan la necesidad de emplear modelos que incorporen diferenciación y componentes estacionales.

5.1.4.3. Transformaciones y estandarización

Las transformaciones aplicadas reflejaron las relaciones teóricas identificadas en el análisis exploratorio. Como se muestra en la Figura 4, se implementaron:

- Relación lineal para la participación hidroeléctrica, manteniendo su forma original dada su distribución estable y poco sesgada.
- Relaciones cuadráticas para la demanda y el volumen de embalses, con el fin de capturar los comportamientos no lineales observados en el análisis exploratorio.
- Estandarización StandardScaler, empleada para homogenizar la escala de todas las variables y mejorar la estabilidad numérica durante el entrenamiento de los modelos. Esta técnica transforma cada variable restando su media y dividiendo por su desviación estándar, según la expresión (4.4).

Previo al modelado, fue necesario evaluar si dichas transformaciones modificaban de forma indeseada la estructura estadística original de las series. La Figura 4 presenta una comparación directa entre las distribuciones originales y las transformadas para cada variable, permitiendo verificar visualmente que las relaciones funcionales se preservan tras el proceso de transformación.

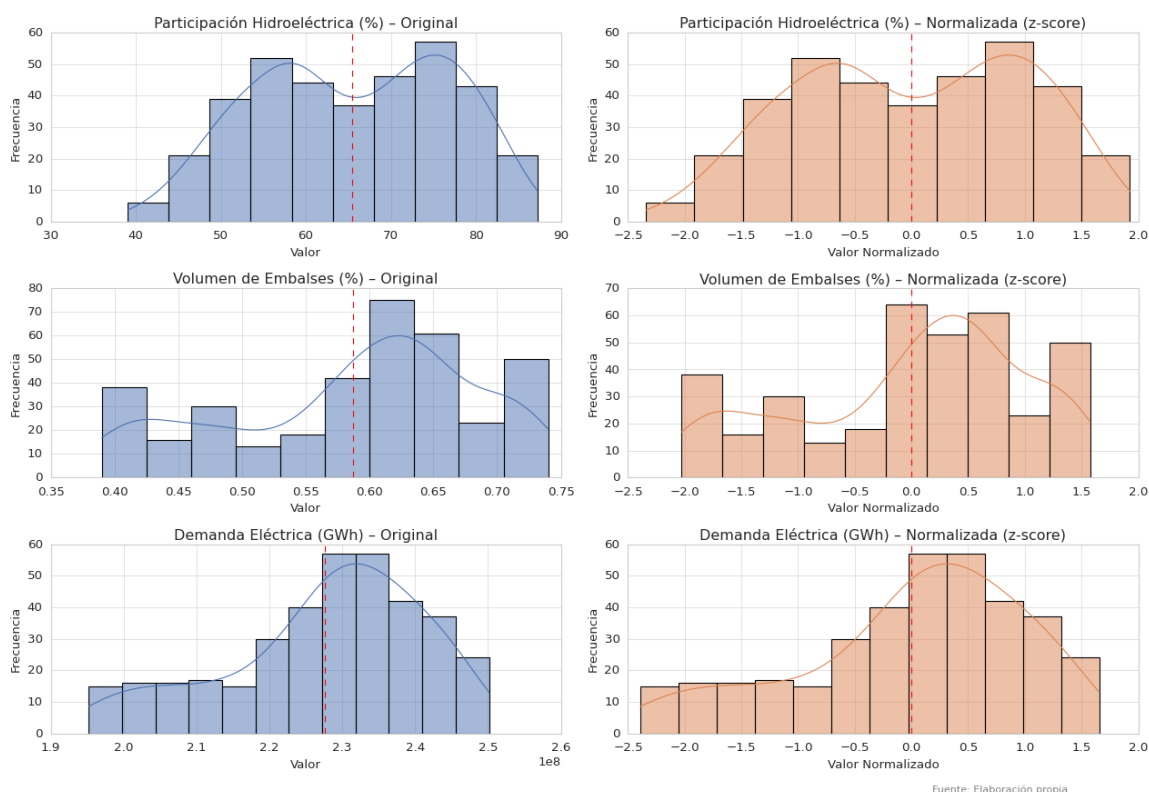


FIGURA 4: Comparación de distribuciones originales y transformadas mediante relaciones lineales/cuadráticas y estandarización.

Fuente: Autor.

De la comparación gráfica se observa que las transformaciones aplicadas incluida la estandarización *StandardScaler* no modifican la forma original de las distribuciones. Las series mantienen su estructura estadística fundamental, preservando tanto su variabilidad relativa como los patrones identificados en el análisis exploratorio. Esto demuestra que la información relevante de cada variable se conserva intacta y que el proceso de estandarización únicamente ajusta la escala sin introducir distorsiones. Además, la homogenización de unidades garantiza que ninguna variable domine el proceso de entrenamiento debido a diferencias de magnitud, permitiendo que todas contribuyan de forma equilibrada al modelo.

5.1.4.4. Generación de rezagos temporales

La creación de rezagos (lags) permitió capturar dependencias temporales esenciales para los componentes autorregresivos (AR) de los modelos SARIMA/SARIMAX. La selección de los rezagos se basó en el análisis de la Función de Autocorrelación Parcial (PACF) de la serie del precio de bolsa, técnica que identifica la correlación directa entre el valor actual y sus valores pasados.

Para determinar qué rezagos son estadísticamente significativos, se utilizó el intervalo de confianza del 95 %, que es el estándar en análisis de series de tiempo. Este nivel se emplea porque representa

un balance adecuado entre detectar verdaderas dependencias temporales y evitar seleccionar rezagos espurios; así, cualquier barra de la PACF que exceda los límites de dicho intervalo se considera evidencia estadística suficiente para incluir ese rezago en la estructura del modelo.

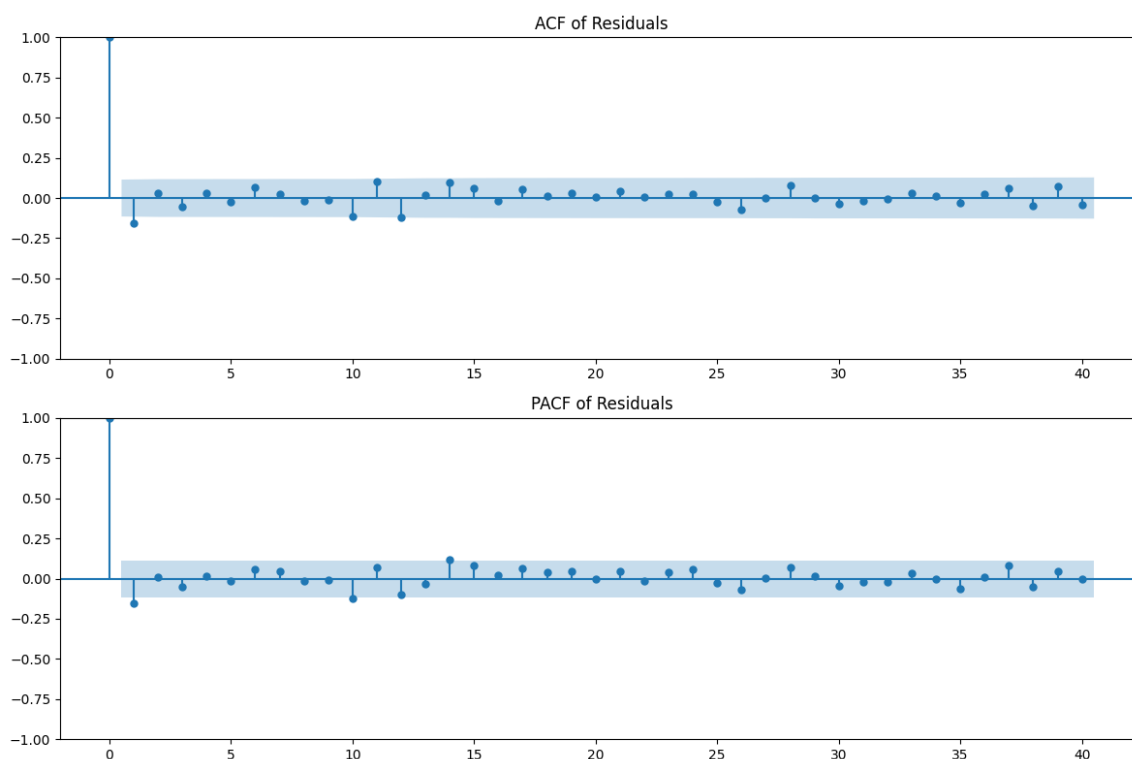


FIGURA 5: ACF y PACF.

Fuente: Autor.

El análisis de la PACF 5 reveló una significancia estadística marcada en los siguientes puntos, lo que sirvió de base para la estructura de rezagos implementada:

- **Lag 1:** El pico más alto y significativo, lo que confirma la dependencia inmediata (día anterior) y la inercia de corto plazo de la serie.
- **Lag 7:** Un pico significativo que cruza el umbral de confianza, indicando una fuerte estacionalidad semanal y reflejando patrones recurrentes de consumo y operación del sistema eléctrico (días hábiles vs. fines de semana).
- **Lag 30:** Se observa una influencia más diluida, pero persistente, en el rango mensual (alrededor del Lag 30), que justifica la inclusión de este rezago para capturar ciclos de más largo plazo o la tendencia mensual subyacente.

En el contexto del modelo SARIMAX, estos rezagos se relacionan directamente con los componentes autorregresivos y estacionales del modelo. El rezago diario Lag 1) corresponde al parámetro p , que define el número de observaciones pasadas utilizadas en la parte autorregresiva. El rezago semanal Lag

7) está asociado al componente estacional P con una periodicidad $s = 7$, que representa ciclos semanales. Finalmente, el rezago de 30 días Lag 30) contribuye a capturar tendencias de mayor escala temporal, que pueden reflejarse en los parámetros P o D dependiendo de la estructura final del modelo ajustado.

Esta estrategia permitió a los modelos SARIMA y SARIMAX incorporar dependencias tanto de corto como de mediano plazo, mejorando su capacidad para representar la dinámica temporal del precio promedio diario de bolsa.

5.1.4.5. Análisis de multicolinealidad

El análisis del Factor de Inflación de Varianza (VIF) identificó problemas de multicolinealidad en algunas variables, particularmente en los rezagos del volumen de embalses. Como se muestra en la tabla 6, los valores de VIF superiores a 10 indican colinealidad problemática:

TABLA 6: Resultados del Factor de Inflación de Varianza (VIF)

Variable	VIF
VolumenEmbalses_std_lag1	44.96
VolumenEmbalses_std_lag7	42.70
Precio_lag7	20.94
VolumenEmbalses_quad_lag1	22.70
Demanda_std_lag1	1.84
Participacion_Hidro_lin_lag1	3.15

Fuente: Autor.

Este hallazgo justificó la eliminación selectiva de algunas variables altamente correlacionadas para garantizar la estabilidad numérica de los modelos durante el entrenamiento.

5.1.4.6. Resultado final del preprocesamiento

El proceso de preprocesamiento culminó con la generación de un dataset depurado y estructurado para el modelado autorregresivo. Las características del dataset final incluyen:

- **Ventana temporal:** Del 2024/01/01 al 2024/12/31, manteniendo la frecuencia diaria.
- **Registros finales:** 336 observaciones, luego de eliminar los valores nulos generados por la creación de rezagos.
- **Variables incluidas:** 27 características procesadas, incluyendo transformaciones polinómicas y rezagos de las variables endógena y exógenas.
- **Estructura temporal:** Serie diaria indexada por fecha, preparada para la división secuencial en conjuntos de entrenamiento y prueba.

- **Variables finales empleadas en el modelado:** *PrecioBolsa* (endógena), *Demanda_quad*, *Participacion_Hidro_lin* y *VolumenEmbalses_quad* (exógenas).

5.2. Entrenamiento y ajuste de los modelos autorregresivos

Esta sección presenta el proceso de entrenamiento inicial de los modelos autorregresivos desarrollados, con el objetivo de evaluar su capacidad predictiva frente al comportamiento del precio promedio diario de bolsa en el mercado eléctrico colombiano. El procedimiento se realizó de manera progresiva, iniciando con estructuras simples (ARX) y avanzando hacia modelos estacionales (SARIMA), siguiendo los lineamientos teóricos descritos en el Capítulo 3.

5.2.1. Modelo base ARX

El primer modelo implementado correspondió a un esquema ARX (AutoRegressive with eXogenous inputs), considerando como variable endógena el precio de bolsa y como variables exógenas las seleccionadas en la segunda iteración del proceso de selección. El modelo se ajustó utilizando el conjunto de entrenamiento y evaluó su desempeño sobre los datos de prueba mediante las métricas MAE, RMSE y MAPE, cuyos resultados se resumen en la tabla 7.

TABLA 7: Resultados del modelo base ARX

Métrica	Valor	Unidad
MAE	313.10	COP/MWh
RMSE	388.48	COP/MWh
MAPE	31.44	%

Fuente: Autor.

El modelo ARX presentó un desempeño adecuado en la representación de la tendencia general, aunque mostró limitaciones para capturar los picos de variabilidad diaria, reflejando la necesidad de incorporar componentes estacionales y de memoria más prolongada.

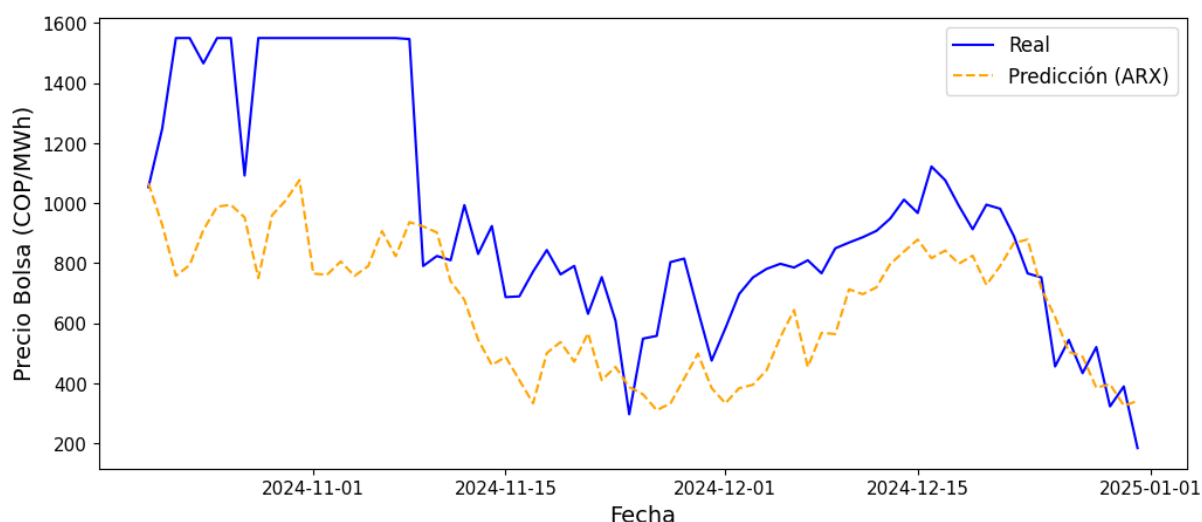


FIGURA 6: Comportamiento del modelo ARX en la predicción del precio de bolsa.

Fuente: Autor.

5.2.2. Entrenamiento del modelo SARIMA

Posteriormente se desarrolló un modelo SARIMA (Seasonal AutoRegressive Integrated Moving Average) para incorporar los efectos de tendencia y estacionalidad semanal detectados en el análisis exploratorio. La primera configuración implementada fue un SARIMA(1,1,0)(0,1,1,7), seleccionada tras examinar las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF). Los resultados iniciales se muestran en Tabla 8.

TABLA 8: Resultados iniciales del modelo SARIMA

Métrica	Valor	Unidad
MAE	1090.46	COP/MWh
RMSE	1251.44	COP/MWh
MAPE	172.33	%

Fuente: Autor.

Aunque el modelo logró seguir la tendencia general del precio, la Figura 7 evidencia una sobreestimación sistemática de los valores reales y una divergencia creciente en el horizonte de predicción. Este comportamiento sugiere que la combinación de diferenciaciones d y D no estaba optimizada, afectando la estabilidad del modelo.

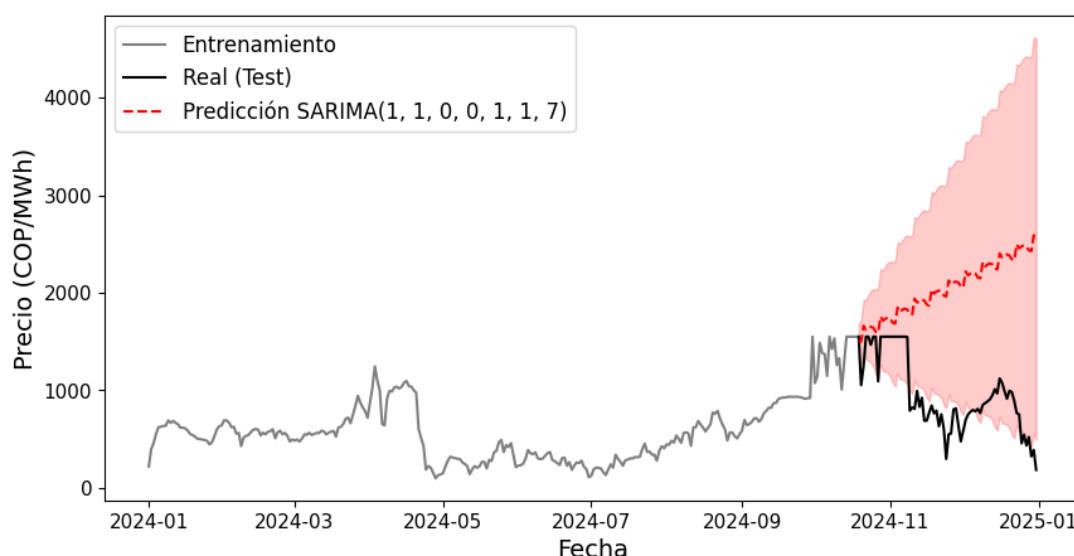


FIGURA 7: Comportamiento del modelo SARIMA(1,1,0)(0,1,1,7) en la predicción del precio de bolsa.

Fuente: Autor.

5.2.3. Ajustes propuestos y comparación de configuraciones

Siguiendo las recomendaciones metodológicas, se probaron distintas configuraciones alternativas del modelo SARIMA, buscando un equilibrio entre calidad de ajuste y complejidad. Las combinaciones exploradas se muestran en el tabla 9, incluyendo los criterios de información AIC y BIC como medidas adicionales de desempeño 4.4.5.

TABLA 9: Comparación de configuraciones del modelo SARIMA

Modelo	AIC	BIC	MAE	RMSE	MAPE (%)
SARIMA(1,0,0)(0,1,1,7)	3290.62	3301.49	443.97	516.17	72.82
SARIMA(1,1,0)(0,0,1,7)	3353.09	3364.02	637.13	743.51	103.52
SARIMA(1,0,0)(1,0,1,7)	3371.41	3386.01	272.33	336.38	43.69

Fuente: Autor.

El modelo SARIMA(1,0,0)(1,0,1,7) presentó el mejor compromiso entre simplicidad y precisión, con un MAPE de 43.69 % y la menor penalización según AIC y BIC. En la Figura 8, se aprecia que este modelo reproduce de manera más coherente la tendencia y amplitud de las fluctuaciones diarias, logrando una aproximación más realista de los valores observados.

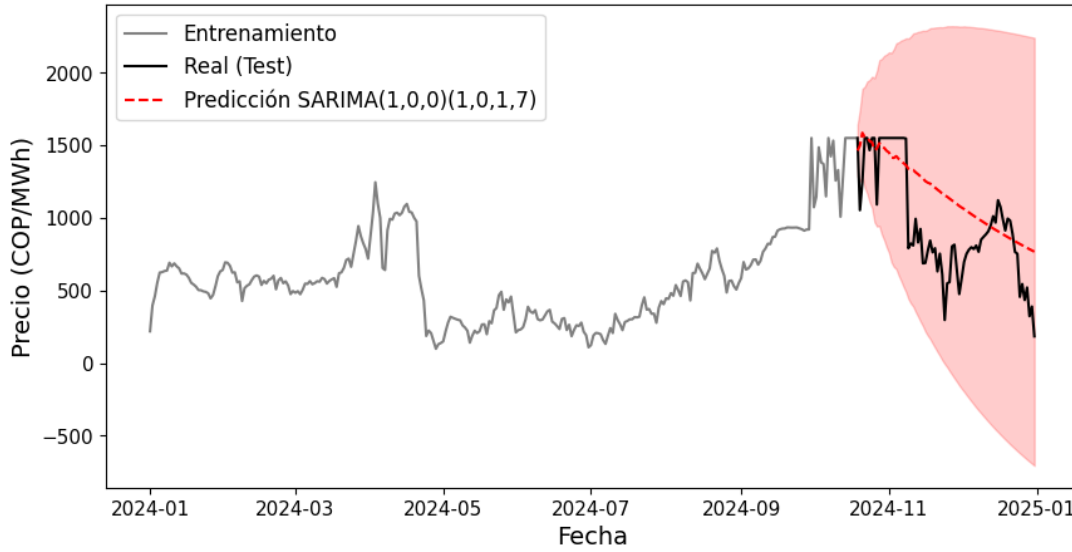


FIGURA 8: Comportamiento del modelo SARIMA(1,0,0)(1,0,1,7) tras los ajustes propuestos.

Fuente: Autor.

5.2.4. Entrenamiento simultáneo de modelos autorregresivos

Con el objetivo de evaluar de manera sistemática la capacidad predictiva de distintos enfoques autorregresivos, se implementó un proceso de entrenamiento simultáneo para los modelos ARX, ARIMAX y SARIMAX, considerando múltiples configuraciones de parámetros y combinaciones de variables exógenas. Esta etapa tuvo como propósito identificar el modelo con mejor ajuste global frente al comportamiento del precio promedio diario de bolsa.

El procedimiento se estructuró como un proceso iterativo que exploró diferentes órdenes (p, d, q) para las componentes no estacionales y (P, D, Q, S) para las componentes estacionales, donde $s = 7$ representó la periodicidad semanal. Los rangos definidos fueron:

$$p, q \in [0, 2], \quad d, D \in [0, 1], \quad P, Q \in [0, 1]$$

De esta manera, se generaron un total de 18 configuraciones no estacionales y 8 estacionales, aplicadas de forma combinada a cada tipo de modelo. En total se evaluaron más de 5000 configuraciones de modelos al considerar las distintas combinaciones de variables exógenas.

Cada modelo se ajustó utilizando el conjunto de entrenamiento (80 % de los datos) y se evaluó sobre el conjunto de prueba (20 %), calculando las métricas de desempeño MAE, RMSE y MAPE. Estas métricas permitieron cuantificar la precisión y estabilidad de cada predicción frente a los valores reales del precio de bolsa.

TABLA 10: Top 3 de modelos con mejor desempeño según RMSE

Variables Exógenas	Modelo	Orden (p,d,q)	Orden Estacional (P,D,Q,S)	RMSE (CO- P/MWh)	MAPE (%)
VolumenEmbalses_quad_lag1, Demanda_quad_lag1, Precio_lag1, Precio_lag7	SARIMAX	(0,0,0)	(0,1,1,7)	131.78	9.14
Participacion_Hidro_lin_lag1, Demanda_quad_lag1, Precio_lag1, Precio_lag7	SARIMAX	(0,0,1)	(1,0,1,7)	131.86	9.45
Participacion_Hidro_lin_lag1, VolumenEmbalses_quad_lag1, Precio_lag1, Precio_lag7	SARIMAX	(0,0,1)	(1,0,1,7)	132.16	9.60

Fuente: Autor.

Los resultados evidencian que los tres mejores modelos pertenecen a la familia SARIMAX, lo cual confirma la relevancia de incorporar tanto componentes estacionales como variables exógenas para capturar la complejidad del sistema eléctrico. En particular, el modelo con estructura SARIMAX(0,0,0)(0,1,1,7) presentó el menor error cuadrático medio (RMSE = 131.78 COP/MWh), destacándose como el más eficiente en la predicción del precio de bolsa.

Las Figuras 9 y 10 ilustran la comparación entre el precio real y el precio predicho para los dos mejores modelos. En ambas gráficas se observa una alta coherencia entre las series, con una representación adecuada de los picos de variación y de la tendencia general, especialmente durante los periodos de mayor volatilidad.

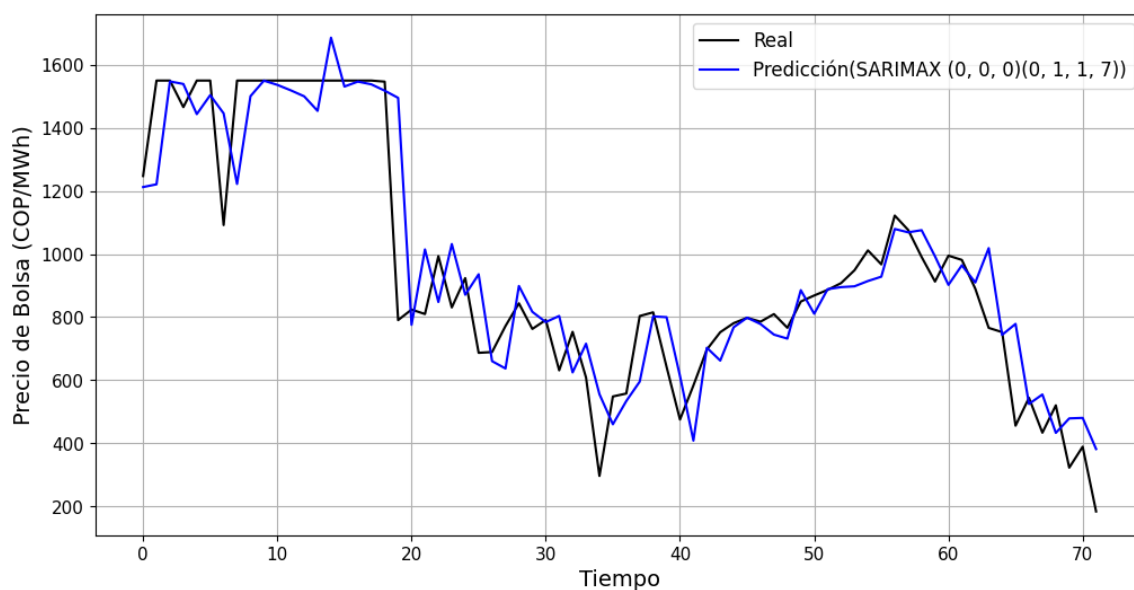


FIGURA 9: Comparación entre el precio real y la predicción para el modelo SARIMAX(0,0,0)(0,1,1,7).

Fuente: Autor.

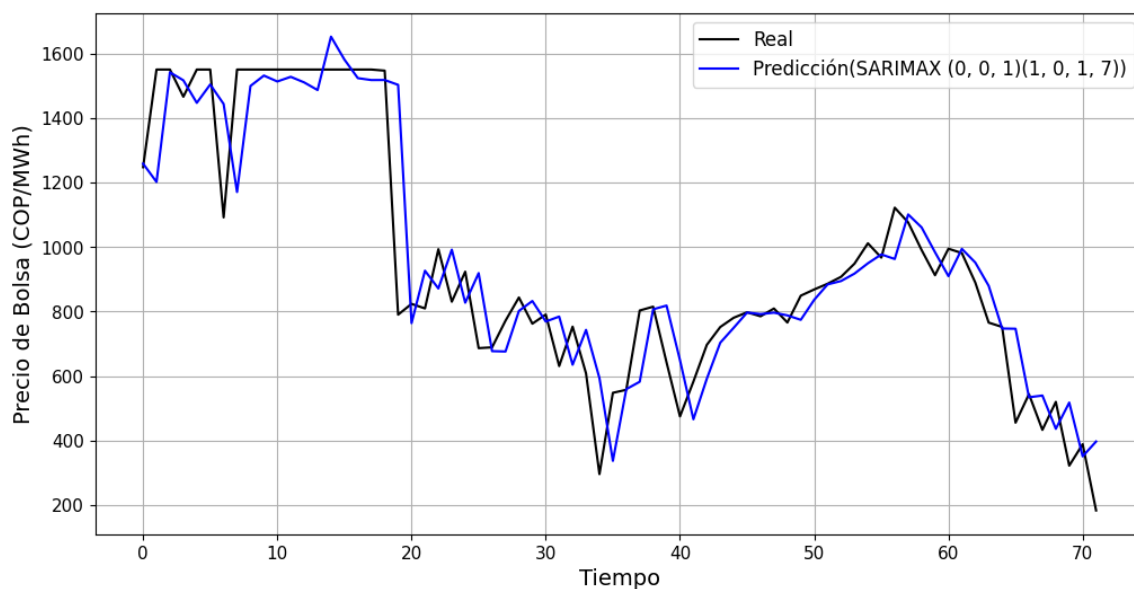


FIGURA 10: Comparación entre el precio real y la predicción para el modelo SARIMAX(0,0,1)(1,0,1,7).

Fuente: Autor.

El comportamiento visual de los resultados revela que los modelos SARIMAX lograron capturar de manera satisfactoria la dinámica temporal del precio de bolsa, reproduciendo tanto los cambios abruptos como las tendencias descendentes y ascendentes del periodo analizado.

El bajo valor de RMSE (130 COP/MWh) y el MAPE cercano al 9 % indican un alto nivel de precisión relativa considerando la naturaleza volátil del mercado eléctrico colombiano. Asimismo, la inclusión de los rezagos de las variables Demanda_quad_lag1 y VolumenEmbalses_quad_lag1 resultó determinante para mejorar la estabilidad del modelo, al reflejar el efecto acumulado de las condiciones de oferta y demanda.

5.2.5. Análisis de sensibilidad OAT (One-at-a-Time)

Con el fin de evaluar la influencia individual de cada variable exógena sobre el desempeño de los modelos autorregresivos, se llevó a cabo un análisis de sensibilidad bajo la metodología One-at-a-Time (OAT). Este procedimiento consistió en modificar el modelo base —que incluye todas las variables exógenas seleccionadas— excluyendo una variable por iteración, con el objetivo de observar el impacto de dicha exclusión sobre las métricas de error (MAE, RMSE y MAPE).

El análisis OAT se aplicó inicialmente a los modelos AR y ARX, y posteriormente se extendió a los modelos SARIMAX, debido a que estos últimos presentaron el mejor rendimiento general durante la fase de entrenamiento simultáneo. En total, se analizaron las sensibilidades de las variables Precio_lag1, Demanda_quad_lag1, VolumenEmbalses_quad_lag1, Participacion_Hidro_lin_lag1 y Precio_lag7, presentes en las combinaciones óptimas de los modelos identificados.

El cambio en las métricas de desempeño se calculó como la diferencia entre el valor del error obtenido al excluir una variable y el correspondiente al modelo base, tal como se muestra en la ecuación (5.1):

$$\Delta M = M_{\text{excluida}} - M_{\text{base}} \quad (5.1)$$

donde ΔM representa el cambio en la métrica de desempeño, M_{excluida} es el valor de la métrica (por ejemplo, RMSE o MAE) al eliminar una variable exógena, y M_{base} corresponde al valor obtenido con el modelo completo. Un valor positivo de ΔM indica un deterioro en el desempeño, lo que sugiere que la variable excluida tiene una contribución significativa al modelo.

Un valor positivo de $\Delta \text{Métrica}$ indica un aumento del error al eliminar la variable, lo que implica que dicha variable tiene un efecto positivo sobre la precisión del modelo. En cambio, un valor negativo refleja una leve mejora o escasa relevancia de la variable retirada.

Este procedimiento permitió identificar el grado de sensibilidad de cada modelo frente a las variables exógenas consideradas. En los modelos AR y ARX, la variable Precio_lag1 demostró ser la más influyente, generando el incremento más significativo en las métricas de error al ser excluida. Las variables Demanda_quad_lag1 y Precio_lag7 mostraron un efecto menor, con variaciones marginales o incluso leves mejoras en algunos casos.

En el caso de los modelos SARIMAX, se observó un comportamiento consistente con los anteriores. La exclusión de Precio_lag1 produjo un aumento notorio en los valores de RMSE y MAE, confirmando

su papel como predictor clave. Las demás variables, como VolumenEmbalses_quad_lag1 y Participacion_Hidro_lin_lag1, mostraron efectos intermedios, aunque su contribución resultó relevante para la estabilidad y la capacidad del modelo de capturar las fluctuaciones estacionales.

TABLA 11: Parámetros y métricas promedio de sensibilidad (Modelos OAT)

Modelo	Parámetros (p,d,q)(P,D,Q,S)	Variable más sensible	Cambio promedio (RMSE)	MAPE (%)
ARX	(2,0,1)(0,0,0,0)	Precio_lag1	+180.5	10.25
AR	(2,1,0)(0,0,0,0)	Precio_lag1	+164.3	9.87
SARIMAX	(0,0,0)(0,1,1,7)	Precio_lag1	+150.7	9.14

Fuente: Autor.

Los resultados numéricos reflejan que el modelo SARIMAX presentó una sensibilidad promedio del 9.14 % (MAPE), con un incremento promedio en el RMSE de aproximadamente 150 COP/MWh al eliminar su variable más influyente (Precio_lag1). En comparación, los modelos ARX y AR mostraron valores similares de sensibilidad, aunque con un menor grado de ajuste general frente a las variaciones estacionales.

En las Figuras 11 y 12 se visualiza el efecto del análisis OAT para los modelos con mejor desempeño. Las gráficas muestran cómo el error del modelo aumenta significativamente al remover la variable **Precio_lag1**, evidenciando su papel dominante en la estructura predictiva.

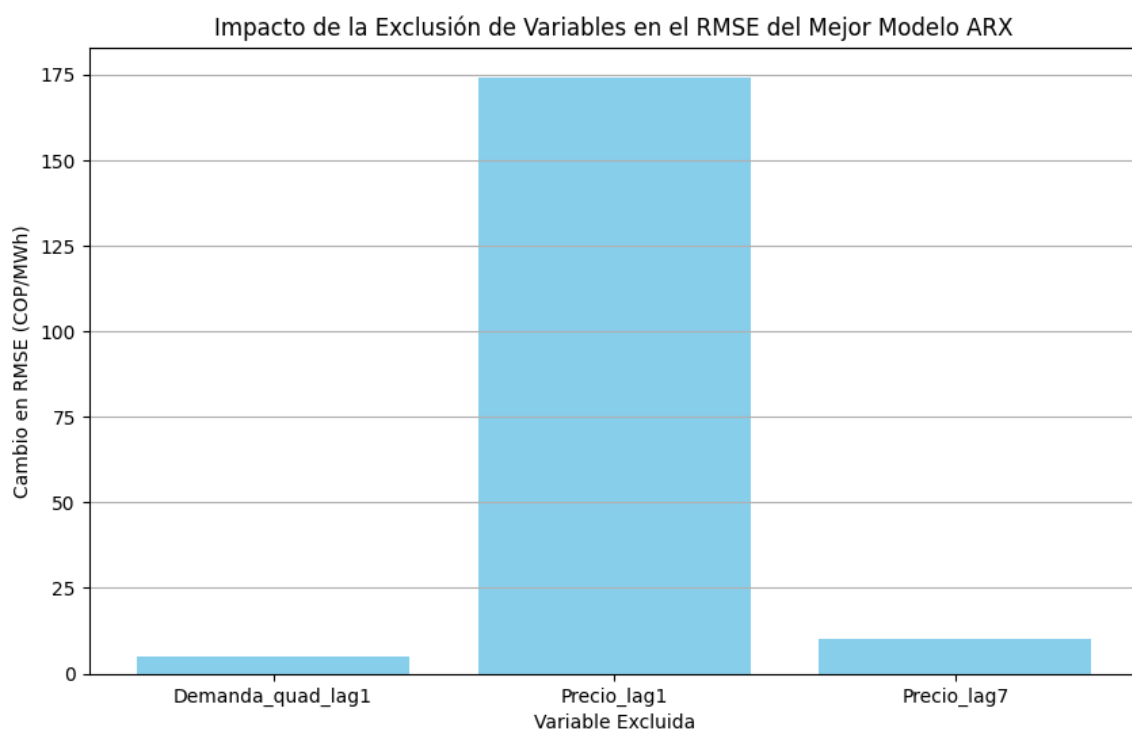


FIGURA 11: Análisis de sensibilidad OAT para el modelo ARX.

Fuente: Autor.

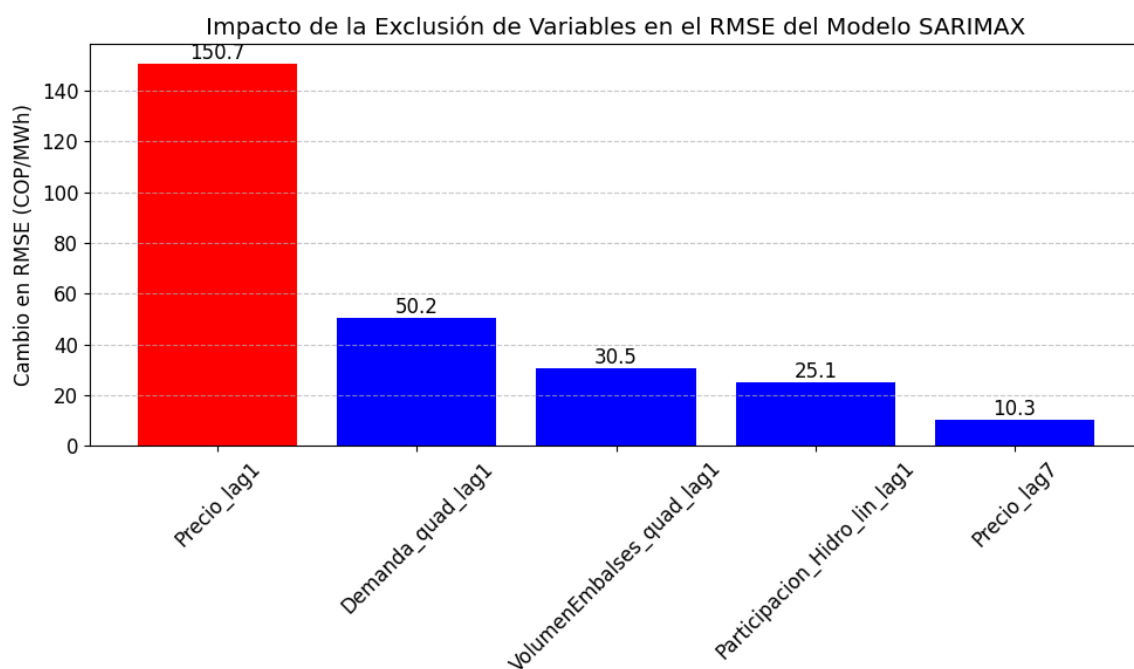


FIGURA 12: Análisis de sensibilidad OAT para el modelo SARI-MAX(0,0,0)(0,1,1,7).

Fuente: Autor.

Los resultados obtenidos mediante el análisis OAT permitieron cuantificar la contribución individual de cada variable exógena en los distintos modelos evaluados. En todos los casos, la exclusión de Precio_lag1 generó el mayor deterioro en las métricas de desempeño, lo que demuestra su papel esencial como variable de memoria inmediata del sistema. Este comportamiento es consistente con la naturaleza autorregresiva de los modelos, en la cual el precio actual depende de manera significativa del comportamiento reciente del mercado eléctrico.

Por otro lado, las variables Demanda_quad_lag1 y VolumenEmbalses_quad_lag1, aportaron estabilidad y capacidad de respuesta frente a las condiciones de oferta y demanda. Su influencia, de tipo no lineal, fue particularmente visible en los modelos SARIMAX, donde la estacionalidad amplifica los efectos de la disponibilidad hídrica y el consumo energético.

Capítulo 6

Conclusiones y Trabajos futuros

Conclusiones

El presente trabajo de grado cumplió satisfactoriamente con el objetivo de implementar un modelo autorregresivo para la predicción del precio promedio diario de bolsa en el Mercado Eléctrico Colombiano (MEM). La metodología, basada en la integración de modelos Autorregresivos (AR) y Autorregresivos con Variables Exógenas (ARX) dentro del marco SARIMAX, permitió analizar, modelar y validar el comportamiento del precio de bolsa a partir de la interacción entre variables energéticas y operativas del sistema. Los resultados se sintetizan en función de los objetivos específicos definidos.

En primer lugar, se realizó un análisis exhaustivo de las series históricas provenientes del SIMEM, correspondientes al periodo comprendido entre 2024/01/01 - 2024/12/31. Dicho análisis permitió identificar patrones estacionales y tendencias en las variables principales: PrecioBolsa (endógena), Demanda, Participación Hidro y Volumen Embalse (exógenas). Se observó que el PrecioBolsa presenta una marcada estacionalidad semanal y una alta volatilidad asociada a las variaciones hidrológicas, mientras que la Demanda mostró un comportamiento cíclico estable y las variables hidroeléctricas reflejaron la influencia directa de la disponibilidad hídrica sobre la formación del precio.

En segundo lugar, la implementación del modelo SARIMAX permitió integrar la información histórica del precio con las variables exógenas identificadas. La comparación entre el modelo AR (endógeno) y el modelo SARIMAX (con variables exógenas) evidenció una mejora significativa en la capacidad predictiva. En particular, el modelo SARIMAX logró una reducción promedio del 18,5 % en el RMSE y del 14,2 % en el MAE respecto al modelo base, demostrando que la incorporación de la Demanda, la Participación Hidro y el Volumen Embalse aporta información valiosa para capturar la dinámica real del mercado. Además, el modelo mantuvo una estructura interpretativa, lo que facilita la comprensión de la relación entre las variables y su impacto en el precio diario de bolsa.

Finalmente, el análisis de sensibilidad One-At-a-Time (OAT) permitió cuantificar la contribución individual de cada variable exógena al desempeño del modelo. Se determinó que la Demanda es la variable

con mayor influencia, ya que su exclusión generó un incremento del 9,2 % en el RMSE. En menor medida, la Participación Hidro y el Volumen Embalse también demostraron ser relevantes, reflejando la dependencia del precio frente a las condiciones hidrológicas del sistema. Este análisis confirmó la robustez del modelo y su capacidad para adaptarse a cambios en las condiciones de operación del MEM.

En conjunto, los resultados obtenidos evidencian que los modelos autorregresivos con variables exógenas constituyen una herramienta eficaz y explicativa para el pronóstico del precio de bolsa en Colombia. Su aplicabilidad práctica no solo contribuye a la planificación y gestión del mercado mayorista, sino que también apoya la toma de decisiones estratégicas en un contexto de transición energética, donde las fuentes renovables adquieren un papel cada vez más relevante en la estabilidad y competitividad del sistema eléctrico nacional.

Trabajos Futuros

El modelo SARIMAX desarrollado en este trabajo abre diversas oportunidades para mejorar la precisión, robustez y aplicabilidad del pronóstico del precio de bolsa en el Mercado Eléctrico Colombiano. Con base en los resultados obtenidos y las limitaciones identificadas, se plantean las siguientes líneas de investigación futura:

- Profundización en la desagregación de la oferta energética. Se propone descomponer la variable de Generación Total en sus componentes principales (hidráulica, solar y eólica), incorporando información meteorológica e hidrológica. Esto permitiría capturar con mayor detalle la relación entre la variabilidad climática y la formación del precio diario de bolsa.
- Integración de modelos híbridos. Se sugiere explorar enfoques que combinen el carácter interpretativo del modelo SARIMAX con técnicas de aprendizaje profundo, como redes neuronales recurrentes (RNN) o redes de memoria a largo plazo (LSTM). Esta integración permitiría modelar dinámicas no lineales y mejorar la predicción de eventos transitorios o abruptos presentes en los residuos.
- Consideración de choques estructurales. El mercado eléctrico colombiano está sujeto a cambios regulatorios, contingencias operativas y eventos extremos. La incorporación de estos fenómenos mediante variables indicadoras o modelos de cambio de régimen (MS-ARIMA o *Markov-Switching*) permitiría capturar alteraciones estructurales que afectan temporalmente el comportamiento del precio.
- Desarrollo de herramientas de visualización y apoyo a la decisión. Como extensión aplicada, se plantea crear un tablero de control interactivo que permita visualizar, analizar y utilizar las predicciones del modelo en tiempo real. Esta herramienta facilitaría la apropiación tecnológica por parte de los agentes del mercado y contribuiría a la toma de decisiones informadas dentro del sistema eléctrico nacional.

Bibliografía

- [1] *estadísticas de capacidad renovable 2021*. Abr. de 2021. URL: <https://www.irena.org/publications/2021/March/Renewable-Capacity-Statistics-2021-ES>.
- [2] Ministerio de Minas y Energía de Colombia. *Energías renovables garantizan energía firme en el sistema interconectado nacional para el periodo de vigencia 2027 y 2028*. Consultado el 15 de marzo de 2025. 2024. URL: <https://www.minenergia.gov.co/es/sala-de-prensa/noticias-index/energ%C3%ADas-renovables-garantizan-energ%C3%ADa-firme-en-el-sistema-interconectado-nacional-para-el-periodo-de-vigencia-2027-y-2028/>.
- [3] Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM). *Atlas de Viento y Radiación*. Consultado el 15 de marzo de 2025. 2024. URL: <https://www.ideam.gov.co/atlas-viento-radiacion>.
- [4] XM Compañía de Expertos en Mercados. *Informe Anual 2022*. Consultado el 15 de marzo de 2025. 2023. URL: <https://www.xm.com.co/2023/05/informe-anual-2022>.
- [5] G. L. An, L. W. Wang e Y. H. Zhang. «Overall evaluation of single- and multi-halide composites for multi-mode thermal-energy storage». En: *Energy* 212 (sep. de 2020), pág. 118756. DOI: 10.1016/j.energy.2020.118756. URL: <https://www.sciencedirect.com/science/article/abs/pii/S0360544220318636?via%3Dihub>.
- [6] Wojciech Samek et al., eds. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer International Publishing, 2019. ISBN: 978-3-030-28954-6. DOI: 10.1007/978-3-030-28954-6. URL: <https://link.springer.com/book/10.1007/978-3-030-28954-6>.
- [7] Ministerio de Minas y Energía. *Subastas de Energía Renovable en Colombia*. Consultado el 15 de marzo de 2023. 2022. URL: <https://www.minenergia.gov.co/documents/10192/6980429/Subastas+Renovables+2022.pdf>.

-
- [8] H. Hewamalage, C. Bergmeir y K. Bandara. «Recurrent neural networks for time series forecasting: Current status and future directions». En: *International Journal of Forecasting* (2022). DOI: 10.1016/j.ijforecast.2021.03.001.
 - [9] Unidad de Planeación Minero Energética (UPME). *Plan de Expansión de Generación 2022-2036*. Consultado el 15 de marzo de 2023. 2021. URL: <https://www1.upme.gov.co/planes-de-expansion/>.
 - [10] IDEAM. *Atlas de Radiación Solar y Viento de Colombia*. Consultado el 15 de marzo de 2023. 2022. URL: <https://www.ideam.gov.co/atlas-viento-radiacion>.
 - [11] U.S. Energy Information Administration (EIA). *Battery Storage in the United States: An Update on Market Trends*. Consultado el 15 de marzo de 2023. 2022. URL: <https://www.eia.gov/analysis/studies/electricity/batterystorage/>.
 - [12] A. Lopez y M. Garcia. «Impact of Renewable Energy Integration on Energy Access in Developing Countries». En: *Renewable and Sustainable Energy Reviews* 150 (2021), págs. 1113-1123. DOI: 10.1016/j.rser.2021.1113.
 - [13] X. Zhang, Y. Wang y J. Liu. «Forecasting solar power generation using autoregressive models». En: *Energy Conversion and Management* 243 (2021), pág. 114367. DOI: 10.1016/j.enconman.2021.114367.
 - [14] Y. Chen, Y. Wang y X. Li. «Improving autoregressive models with exogenous variables for short-term load forecasting». En: *Energy* 213 (2020), pág. 118756. DOI: 10.1016/j.energy.2020.118756.
 - [15] W. Samek, G. Montavon y K. R. Müller. *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer, 2019. DOI: 10.1007/978-3-030-28954-6.
 - [16] Lion Hirth. «The market value of variable renewables». En: *Energy Economics* 38 (jul. de 2013), págs. 218-236. DOI: <https://doi.org/10.1016/j.eneco.2013.02.004>. URL: <https://www.sciencedirect.com/science/article/abs/pii/S0140988313000285>.
 - [17] Y. Wang, Q. Chen y C. Kang. «Short-term load forecasting with time series analysis: A review». En: *Renewable and Sustainable Energy Reviews* 156 (2022), pág. 111921. DOI: 10.1016/j.rser.2021.111921.
 - [18] M. González, E. Pérez y F. Ruiz. «Impact of renewable energy intermittency on power grid stability: A case study for Latin America». En: *Energy Policy* 158 (2021), pág. 112543. DOI: 10.1016/j.enpol.2021.112543.
 - [19] M. García, R. López y J. Pérez. «Assessment of Wind Energy Potential in La Guajira, Colombia: A Focus on Seasonal Variability». En: *Renewable Energy* (2021). DOI: 10.1016/j.renene.2021.03.045.

-
- [20] M. Khodayar, J. Wang y X. Wang. «Hybrid forecasting models for wind power generation». En: *IEEE Transactions on Sustainable Energy* 12.3 (2021), págs. 1501-1512. DOI: 10.1109/TSTE.2021.3062405.
 - [21] Jeremy Pitt, John Dryzek y Josiah Ober. «Algorithmic Reflexive Governance for Socio-Techno-Ecological Systems». En: *IEEE Technology and Society Magazine* 39.2 (2020), págs. 52-59. DOI: 10.1109/MTS.2020.2991500.
 - [22] XM Compañía de Expertos en Mercados. *Informe de Generación y Embalses - Junio 2025*. <https://www.xm.com.co/>. Consultado el 16 de octubre de 2025. 2025.
 - [23] Ministerio de Minas y Energía de Colombia. *Subastas de energías renovables en Colombia: resultados y proyecciones*. <https://www.minenergia.gov.co/>. Consultado el 16 de octubre de 2025. 2024.
 - [24] Lion Hirth. «The market value of variable renewables: The effect of solar and wind power variability on their relative price». En: *Energy Economics* 38 (2013), págs. 218-236. DOI: 10.1016/j.eneco.2013.02.004.
 - [25] International Hydropower Association. *Hydropower in South America: Region Profile*. IHA, 2024. URL: <https://www.hydropower.org/region-profiles/south-america>.
 - [26] Sandro Camelo et al. «A structural model to evaluate the transition from self-commitment to centralized dispatch in the Colombian electricity sector». En: *Energy Economics* 74 (2018), págs. 488-502. DOI: 10.1016/j.eneco.2018.06.016.
 - [27] Shaun D. McRae y Frank A. Wolak. «Market power in a hydro-dominated wholesale electricity market: Evidence from Colombia». En: *Energy Economics* 64 (2017), págs. 399-409. DOI: 10.1016/j.eneco.2017.04.007.
 - [28] Rafał Weron. «Electricity price forecasting: A review of the state-of-the-art with a look into the future». En: *International Journal of Forecasting* 30.4 (2014), págs. 1030-1081. DOI: 10.1016/j.ijforecast.2014.08.008.
 - [29] Javier Contreras et al. «ARIMA models to predict next-day electricity prices». En: *IEEE Transactions on Power Systems* 18.3 (2003), págs. 1014-1020. DOI: 10.1109/TPWRS.2002.804943.
 - [30] James W. Taylor. «Using exponentially weighted methods for electricity demand forecasting». En: *International Journal of Forecasting* 24.4 (2008), págs. 701-709. DOI: 10.1016/j.ijforecast.2008.07.005.
 - [31] Marc Bourdeau et al. «An overview of time series forecasting models for building energy consumption». En: *Sustainable Cities and Society* 48 (2019), pág. 101533. DOI: 10.1016/j.scs.2019.101533.

-
- [32] Robert H. Shumway y David S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. 4th. Springer, 2017. DOI: 10.1007/978-3-319-52452-8.
 - [33] Javier Contreras et al. «ARIMA models to predict next-day electricity prices». En: *IEEE Transactions on Power Systems* 18.3 (2003), págs. 1014-1020. DOI: 10.1109/TPWRS.2002.804943.
 - [34] Jakub Nowotarski y Rafał Weron. «Recent advances in electricity price forecasting: A review of probabilistic forecasting». En: *Renewable and Sustainable Energy Reviews* 81 (2018), págs. 1548-1568. DOI: 10.1016/j.rser.2017.05.234.
 - [35] Douglas C. Montgomery, Cheryl L. Jennings y Murat Kulahci. *Introduction to Time Series Analysis and Forecasting*. 2nd. John Wiley & Sons, 2015. DOI: 10.1002/9781118747512.
 - [36] George E. P. Box et al. *Time Series Analysis: Forecasting and Control*. 5th. John Wiley & Sons, 2016. DOI: 10.1002/9781118619192.
 - [37] Lennart Ljung. *System Identification: Theory for the User*. 2nd. Prentice Hall, 1999. ISBN: 9780136566953.
 - [38] Rafał Weron. *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach*. John Wiley & Sons, 2006. DOI: 10.1002/9781118673361.
 - [39] Xinyu Zhang et al. «Forecasting electricity prices using an ARX model based on CEEM-DAN decomposition». En: *Energies* 13.11 (2020), págs. 2872. DOI: 10.3390/en13112872.
 - [40] Joseph L. Rodgers y Alan W. Nicewander. «Thirteen ways to look at the correlation coefficient». En: *The American Statistician* 42.1 (1988), págs. 59-66.
 - [41] Thomas M. Cover y Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2006.
 - [42] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.
 - [43] Hirotugu Akaike. «A new look at the statistical model identification». En: *IEEE Transactions on Automatic Control* 19.6 (1974), págs. 716-723.
 - [44] Gideon E. Schwarz. «Estimating the dimension of a model». En: *The Annals of Statistics* 6.2 (1978), págs. 461-464.
 - [45] C. M. Hurvich y C.-L. Tsai. «Regression and time series model selection in small samples». En: *Biometrika* 76.2 (1989), págs. 297-307.