**Logotipo

Descripción generada automáticamente**

**Predicción del precio ponderado de la energía en la bolsa colombiana mediante la relación de variables de precipitación “IDEAM” y del mercado energético “XM” con datos recolectados en un periodo de 10 años**.

Felipe Ruiz Zea

Yeferson Garcia Vasquez

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor

David Villanueva Valdés, M.Sc. Data Scientist

Universidad de Antioquia

Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2024

|  |  |
| --- | --- |
| **Cita** | (Ruiz Zea & Garcia Vasquez, 2024) |
| **Referencia**  **Estilo APA 7 (2020)** | Ruiz Zea, F., & Garcia Vasquez, Y. F. (2024). *Predicción del precio ponderado de la energía en la bolsa colombiana mediante la relación de variables de precipitación “IDEAM” y del mercado energético “XM” con datos recolectados en un periodo de 10 años*. Universidad de Antioquia, Medellín, Colombia. |

**** 

Especialización en Analítica y Ciencia de Datos, CohorteV.

Centro de Investigación Ambientales y de Ingeniería (CIA).

|  |  |
| --- | --- |
|  | Diagrama  Descripción generada automáticamente con confianza media |

Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

**Dedicatoria**

Texto de dedicatoria centrado.

**Agradecimientos**

Texto de agradecimientos centrado.

# Tabla de contenido

[Siglas, acrónimos y abreviaturas 9](#_Toc167737885)

[Resumen 10](#_Toc167737886)

[Abstract 11](#_Toc167737887)

[1. Descripción del problema 12](#_Toc167737888)

[1.1. Problema de negocio 12](#_Toc167737889)

[1.2. Aproximación desde la analítica de datos 12](#_Toc167737890)

[1.3. Origen de los datos 12](#_Toc167737891)

[1.4. Métricas de desempeño 13](#_Toc167737892)

[2. Objetivos 15](#_Toc167737893)

[2.1. Objetivo general 15](#_Toc167737894)

[2.2. Objetivos específicos 15](#_Toc167737895)

[3. Datos 16](#_Toc167737896)

[3.1. Datos originales 16](#_Toc167737897)

[3.2. Datsets 18](#_Toc167737898)

[3.2.1. DataSet Precipitaciones 18](#_Toc167737899)

[3.2.2. DataSets precio energia y volumen útil de embalse 19](#_Toc167737900)

[3.3. Analítica descriptiva 20](#_Toc167737901)

[3.3.1. precipitaciones 20](#_Toc167737902)

[3.3.2. Volumen 22](#_Toc167737903)

[3.3.3. Precio 24](#_Toc167737904)

[3.3.4. Correlaciones entre variables de precipitación y volumen almacenado 25](#_Toc167737905)

[3.3.5. Correlaciones entre volumen almacenado y precio 25](#_Toc167737906)

[Referencias 28](#_Toc167737907)

# Tablas

[Tabla 1. Descripción de columnas del dataset de precipitaciones. 15](#_Toc167742173)

[Tabla 2. Variables para Volumen útil de embalses. 16](#_Toc167742174)

[Tabla 3. Variables que describen el dataset, de precio de la energía en la bolsa nacional. 17](#_Toc167742175)

[Tabla 4. Estadística descriptiva del conjunto de datos de precipitaciones. 19](#_Toc167742176)

[Tabla 5. Estadística descriptiva de datos de energía almacenada embalses Cundinamarca. 21](#_Toc167742177)

[Tabla 6. Estadística descriptiva de los datos de precio de energía en la bolsa nacional. 22](#_Toc167742178)

# Tabla de Ecuaciones

[Ecuación 1. Mean Squared Error (MSE) 12](#_Toc167736865)

[Ecuación 2. Root Mean Squared Error (RMSE) 12](#_Toc167736866)

[Ecuación 3. Standard Deviation (STD). 13](#_Toc167736867)

[Ecuación 4. Adjusted Coefficient of Determination 13](#_Toc167736868)

# Tabla de Figuras

[Figura 1. ACF de precipitación. 19](#_Toc167742361)

[Figura 2 Grafica serie temporal datos de precipitación 20](#_Toc167742362)

[Figura 3. Energía almacenada de embalses Cundinamarca. 21](#_Toc167742363)

[Figura 4. ACF de volumen de embalses. 22](#_Toc167742364)

[Figura 5. ACF precios de energía. 23](#_Toc167742365)

[Figura 6. Búsqueda de mejor correlación Precipitación Vs Vol. embalses 23](#_Toc167742366)

[Figura 7. Superposición escalada de series de tiempo precipitación Vs Volumen embalse. 24](#_Toc167742367)

[Figura 8. Búsqueda de mejor correlación Vol. embalses Vs Precio. 25](#_Toc167742368)

[Figura 9. Superposición escalada de series de tiempo Vol. Embalse Vs Precio. 25](#_Toc167742369)

# Siglas, acrónimos y abreviaturas

**mm.** Milímetros

**Kwh** Kilowatio Hora

**COP** Pesos colombianos

**UdeA** Universidad de Antioquia

**MAE** Error Cuadrático Medio

**RMSE** Raíz del Error Cuadrático Medio

**ACF** Función de autocorrelación

**STD** Deviación Estándar

# Resumen

En Colombia, el precio de la energía ha tendido al aumento irregular debido a muchos factores, entre ellos climáticos, especulativos y operacionales, lo cual ha llevado a un 37% de alza en su valor en 2023[1] debido a la alta dependencia de las centrales hidroeléctricas, cuya generación es sensible a las variaciones en las precipitaciones. La complejidad de anticipar estas fluctuaciones radica en la interconexión entre los sistemas meteorológicos, hidrológicos, energéticos y la competencia en el mercado. Para abordar este problema, se propone desarrollar un modelo de regresión que integre datos climatológicos del IDEAM y datos operacionales del sistema energético de XM, considerando variables como el volumen útil de los embalses, el aporte de ríos, la demanda energética y los precios históricos. Los objetivos del modelo incluyen predecir el impacto de las variaciones en las precipitaciones sobre la producción energética, anticipar las fluctuaciones en los precios de la energía y evaluar el desempeño del modelo con datos reales. La metodología empleará herramientas de ciencia de datos con Python y librerías como Scikit-learn para la limpieza, exploración y modelado de datos. Este modelo permitirá mejorar la planificación y gestión del sistema energético colombiano, aumentando su resiliencia ante las incertidumbres meteorológicas y optimizando la toma de decisiones estratégicas en la gestión de recursos energéticos a largo plazo.

**GitHub Repos:**

* [**Proyecto-Seminario**](https://github.com/Pipe1017/Proyecto-Seminario)

*Palabras clave*: Energía, Hidroeléctrica, Precipitaciones, Modelo de regresión, Predicción, Python, XM.

# Abstract

In Colombia, the price of energy has tended to increase irregularly due to many factors, including climatic, speculative, and operational factors. This has led to a 37% increase in its value in 2023 [1] due to the high dependence on hydroelectric power plants, whose generation is sensitive to variations in precipitation. The complexity of anticipating these fluctuations lies in the interconnection between meteorological, hydrological, and energy systems, as well as market competition. To address this problem, a regression model is proposed that integrates climatological data from IDEAM and operational data from the XM energy system, considering variables such as the usable volume of reservoirs, river inflows, energy demand, and historical prices. The objectives of the model include predicting the impact of precipitation variations on energy production, anticipating fluctuations in energy prices, and evaluating the model's performance with real data. The methodology will employ data science tools with Python and libraries such as Scikit-learn for data cleaning, exploration, and modeling. This model will improve the planning and management of the Colombian energy system, increasing its resilience to meteorological uncertainties and optimizing strategic decision-making in the long-term management of energy resources.

*Keywords***:** Energy, Hydroelectric, Precipitation, Regression model, Prediction, Python, XM, precio.

# Descripción del problema

La generación de energía hidroeléctrica es altamente sensible a las variaciones en las precipitaciones, esto hace que sea crítico para el panorama colombiano, donde el suministro eléctrico depende en un 70% por hidroeléctricas [2], lo que introduce una gran incertidumbre en el mercado energético del país. Este problema se base en anticipar las fluctuaciones en los precios de la energía, lo cual propone una tarea compleja debido a la interconexión entre los sistemas meteorológicos, hidrológicos y energéticos, así como de eventos fortuitos de interrupción causados por mal mantenimiento o sucesos imprevistos.

## Problema de negocio

En la actualidad, el precio de la energía en Colombia subió hasta un 37% en el 2023 [1] debido a la alta dependencia de las centrales hidroeléctricas. Anticipar las variaciones en estos indicadores es una tarea compleja por la interconexión entre estos sistemas, la difícil predicción de fenómenos meteorológicos y la alta competencia en el mercado de los diferentes agentes. La implementación de un modelo que integre los datos climatológicos de precipitación en Colombia (IDEAM) y los datos de operación del sistema Interconectado de Energía Mayorista de Colombia (XM) como el volumen útil de los embalses, aporte de ríos y demanda energética permitirá predecir y anticipar el impacto del alza de precios energía. Las métricas que se podrían incluir serían la evaluación del desempeño del modelo con datos reales de los precios del Sistema Interconectado Nacional (SIN).

## Aproximación desde la analítica de datos

Para abordar este problema, se investigan casos exitosos en la literatura para identificar posibles modelos que se adapten a las características de nuestro problema, como el Vector Autoregression (VAR)[3], XgBoost[4] o modelos de Facebook como Prophet[5]. Estos modelos serán evaluados en función del error cuadrático medio (MSE), R cuadrado ajustado entre los datos de precipitación, el volumen útil disponible, la demanda y otros; además, se evaluarán las mejores correlaciones entre las diferentes variables disponibles.

## Origen de los datos

Los datos utilizados en este proyecto provienen de fuentes de acceso público. La información de precipitaciones se extraerá utilizando la página oficial de "Datos Abiertos Colombia". Por otro lado, los datos del sistema energético interconectado se obtendrán utilizando la base de datos oficial de XM, la cual es la entidad reguladora en el país para las tarifas eléctricas, a través de su API o su biblioteca disponible en Python. Ambos conjuntos de datos abarcan un período de 4 años, desde 2019 hasta 2022.

|  |  |  |
| --- | --- | --- |
| **DataSet** | **Fuente** | **Tipo de Acceso** |
| Precipitaciones | [Precipitación](https://www.datos.gov.co/en/Ambiente-y-Desarrollo-Sostenible/Precipitaci-n/s54a-sgyg/about_data) ([www.datos.gov.co/](http://www.datos.gov.co/)) | Público (Consumo API o descarga directa) |
| Energía/Precio, Nivel de embalses, vertimientos, consumos… | [XM](https://sinergox.xm.com.co/),  [Pydataxm](https://pypi.org/project/pydataxm/) | Público (Consumo API o descarga directa) |

Tabla 1. Fuentes de datasets.

## Métricas de desempeño

Para la evaluación del modelo analítico se pueden usar métricas estadísticas como el RMSE, MSE, STD, y combinaciones de estas para el precio estimado y el precio real. Sobre las métricas del negocio, es posible observar los diferentes porcentajes en el alza de los precios debido a periodos de sequía. Por ejemplo, simular escenarios con el modelo entrenado disminuyendo la carga de la generación hidroeléctrica compensada con otras fuentes que no dependan de recursos hídricos como térmicas y estabilizar el comportamiento del precio. A continuación, se muestran las formulaciones matemáticas correspondientes[6].

Ecuación 1. Mean Squared Error (MSE)

Ecuación 2. Root Mean Squared Error (RMSE)

Ecuación 3. Standard Deviation (STD).

Ecuación 4. Adjusted Coefficient of Determination

Donde:

* *n*: Número total de observaciones.
* ​: Valor observado en la *i*-ésima observación.
* : Valor predicho en la *i*-ésima observación.
* : Valor del 𝑖-ésimo dato.
* : Media (promedio) de todos los valores de x.
* : Coeficiente de Determinación no ajustado.
* 𝑘: Número de variables predictoras en el modelo.

# Objetivos

## Objetivo general

Desarrollar un modelo de regresión completo destinado a prever el impacto de las fluctuaciones de precipitación en la producción de energía, centrándose específicamente en la disponibilidad y la dinámica de precios. Esto implica la recopilación, limpieza, exploración, modelación y visualización de conjuntos de datos públicos relacionados con la precipitación, y la generación de energía hidroeléctrica como volumen útil del embalse, vertimientos, consumos y precio.

## Objetivos específicos

* Limpiar y preprocesar los datos recopilados para asegurar su calidad y consistencia como eliminar o imputar valores faltantes, remover datos anormales, y normalizar o estandarizar las variables.
* Explorar los datos mediante análisis descriptivo y visualización para establecer relaciones preliminares entre las precipitaciones y la producción de energía hidroeléctrica con el cálculo de correlaciones entre las variables principales e identificar posibles variables independientes adicionales que influyan en la producción energética (por ejemplo, nivel de embalses, consumo, salidas, ubicación de sensores).
* Definir y construir el modelo de regresión, y evaluar el rendimiento del modelo de regresión, mediante la división de los datos en conjuntos de entrenamiento y prueba para posteriormente, utilizar métricas de evaluación (MAE, RMSE, R²) para medir la precisión del modelo.
* Interpretar los resultados del modelo de regresión con base en el análisis de los coeficientes del modelo para entender la influencia de cada variable independiente además de realizar pruebas de significancia estadística para validar los resultados.
* Comparar los resultados obtenidos con estudios previos o modelos alternativos. Realizar una revisión de literatura para contextualizar los hallazgos. Evaluar si el modelo desarrollado mejora la precisión en comparación con enfoques previos.

# Datos

## Datos originales

En el ámbito de las precipitaciones se tiene un dataset obtenido del sitio web de datos abiertos de Colombia ([www.datos.gov.co/](http://www.datos.gov.co/)).  El dataset llamado [Precipitación](https://www.datos.gov.co/en/Ambiente-y-Desarrollo-Sostenible/Precipitaci-n/s54a-sgyg/about_data) tiene datos de precipitaciones extraídas de sensores del IDEAM que miden la cantidad de lluvia en milímetros en distintos departamentos y municipios diariamente (el dataset es actualizado diariamente también) además contiene datos geográficos de longitud y latitud y los distintos sensores que capturan las medidas. Se han observado instancias desde el 2002 hasta el presente (mayo del 2024), aunque la calidad del dataset se vuelve significativo a partir del 2019, además el tamaño de los datos es 192 millones de registros a la fecha. La siguiente es una descripción de las columnas encontradas en el dataset original.

|  |  |  |
| --- | --- | --- |
| **Nombre de la columna** | **Descripción** | **Tipo** |
| CodigoEstacion | Corresponde al valor de identificación de la estación dentro del catálogo de estaciones | Texto |
| CodigoSensor | Código de identificación asignado al sensor | Texto |
| FechaObservacion | Fecha en la cual se realiza la medición | Fecha y hora |
| ValorObservado | Valor medido | Número |
| NombreEstacion | Corresponde a la identificación de la estación dentro del catálogo de estaciones | Texto |
| Departamento | Nombre del departamento donde se ubica la estación | Texto |
| Municipio | Nombre del Municipio donde se localiza la estación | Texto |
| ZonaHidrografica | Zona hidrográfica sobre la cual está ubicada la estación | Texto |
| Latitud | Corresponde a la latitud en la cual se ubica la estación | Número |
| Longitud | Corresponde a la longitud en la cual se ubica la estación | Número |
| DescripcionSensor | Precipitación | Texto |
| UnidadMedida | Milímetros | Texto |

Tabla 1. Descripción de columnas del dataset de precipitaciones.

Este dataset es accesible de manera pública y puede descargarse toda su información desde la página oficial directamente (no recomendado, la conexión es inestable por lo cual falla fácilmente), consumiendo su API (se encuentran muchas inconsistencias y demoras en las peticiones), la mejor manera es haciendo Queries con las herramientas de las que ellos disponen directamente en el portal, que permiten hacer los distintos filtros que sean necesarios para delimitar el problema y descargar el dataset. Los archivos pueden ser adquiridos en formato JSON o CSV.

Para el caso energético hay métricas disponibles desde el 2000 hasta la actualidad, que incluyen las variaciones de los precios de oferta y demanda, volumen útil de los embalses, vertimientos entre otros. Estos datos se encuentran en el sitio web de [XM](https://sinergox.xm.com.co/Paginas/Home.aspx) y [ido.xm.com.co/ido/SitePages/hidrologia.aspx?q=reservas](https://ido.xm.com.co/ido/SitePages/hidrologia.aspx?q=reservas) además de que cuentan con una librería de Python ([Pydataxm](https://pypi.org/project/pydataxm/)) la cual es bastante útil y permite obtener todas estas variables de una manera más fácil y rápida ya que no necesita ninguna capa de autenticación, solo algo de conocimiento básico de programación. La cantidad aproximada de registros para un periodo de 4 años es de 1460. A continuación, se listan las tablas que describen las variables que se van a consumir de estas API, las cuales corresponden al volumen útil y el precio.

|  |  |  |
| --- | --- | --- |
| Nombre de Columna | Descripción | Tipo de Datos |
| MetricId | 119, 133 (Id de 2 tipos de métrica primera con filtro sistema y la segunda para filtrar embalses) | Número |
| MetricName | VoluUtilDiarEner (Esta es la cantidad en KWh del volumen útil diario de energía disponible diariamente por sistema o embalse) | Texto simple |
| Entity | Sistema, Embalse (Como se puede obtener el dataset segmentado por sistema o embalse) | Texto simple |
| MaxDays | 31 (días máximos por petición) | Número |
| Type | DailyEntities (Generalmente puede ser con base diaria o por horas) | Texto simple |
| Url | <http://servapibi.xm.com.co/daily> (URL a la cual se pueden hacer las peticiones directamente) | Texto simple |
| Filter | No aplica, Nombre Embalse (Filtro que se pueden aplicar dependiendo del MetricId o el Entity) | Texto simple |
| MetricUnits | kWh (Las unidades en las cuales se va a representar la variable principal, en este caso VoluUtilDiarEner) | Texto simple |
| MetricDescription | Volumen almacenado por encima del Nivel Mínimo útil | Texto simple |

Tabla 2. Variables para Volumen útil de embalses.

|  |  |  |
| --- | --- | --- |
| Nombre de Columna | Descripción | Tipo de Datos |
| MetricId | 160 (Id de la métrica) | Número |
| MetricName | PPPrecBolsNaci (Nombre de la métrica a obtener) | Texto simple |
| Entity | Sistema | Texto simple |
| MaxDays | 31 (días máximos por petición) | Número |
| Type | DailyEntities (Generalmente puede ser con base diaria o por horas) | Texto simple |
| Url | <http://servapibi.xm.com.co/daily> (URL a la cual se pueden hacer las peticiones directamente) | Texto simple |
| Filter | No aplica | Texto simple |
| MetricUnits | COP/kWh (Las unidades en las cuales se va a representar la variable principal, en este caso PPPrecBolsNaci) | Texto simple |
| MetricDescription | El Precio de Bolsa ponderado | Texto simple |

Tabla 3. Variables que describen el dataset, de precio de la energía en la bolsa nacional.

Como este dataset es interpretado inmediatamente por la librería [Pydataxm](https://pypi.org/project/pydataxm/) los datos son convertidos a un dataframe después de la petición por la misma librería, por lo cual pueden ser guardados en cualquier formato deseado posterior a la consulta, preferiblemente CSV por ser estándar y de tamaño compacto.

## Datsets

### DataSet Precipitaciones

Para crear el dataset objetivo, que permita correlacionar, graficar y manipular datos relacionados con precipitaciones, se realizó una segmentación inicial por departamento, seleccionando Cundinamarca por ser la región con mayor población del país[7]. Se utilizaron datos de los años 2019 a 2022, con el objetivo de reducir los tiempos de procesamiento ya que a pesar de haber aislado una sola región para el análisis exploratorio se sigue teniendo un dataset de un tamaño considerable (entre 900 MB y 1 GB cada uno). Estos archivos se obtuvieron del portal oficial de datos abiertos de Colombia mediante la herramienta de [Query Data](https://www.datos.gov.co/Ambiente-y-Desarrollo-Sostenible/Precipitaci-n/s54a-sgyg/explore) disponible en el sitio.

En el análisis exploratorio del dataset, no se encontraron datos nulos ni atípicos. Se eliminaron algunos datos duplicados y se transformaron las columnas de fechas al formato ‘date’, convirtiéndolas a la forma día-mes-año para agrupar los registros diarios. Esto facilitó la fusión de los datasets de volumen útil de embalse y precipitaciones, basándose en su correspondencia por fechas.

Los datos se escalaron utilizando la función MinMaxScaler de la librería scikit-learn, lo que no solo facilitó la correlación, sino también la visualización de las relaciones entre variables mediante gráficos. Para los siguientes pasos, en la búsqueda o ajuste de un modelo, se planea dividir el dataset en un 90% para entrenamiento y un 10% para pruebas o validaciones.

### DataSets precio energia y volumen útil de embalse

Con respecto a los datasets de precio de la energía y volumen útil de embalses, son directamente filtrados por la librería de Python, [Pydataxm](https://pypi.org/project/pydataxm/) la cual se alimenta de la API [servapibi.xm.com.co](https://servapibi.xm.com.co) perteneciente a XM, la empresa que gestiona el Sistema Interconectado y el Mercado de Energía Mayorista de Colombia[8]. Este conjunto de datos se analizó para el intervalo de años desde 2019 a 2022. El conjunto de datos para el precio de la energía de la bolsa nacional presenta registros diarios sin valores nulos o duplicados. Además, los tipos de datos son los correctos: los números son instancias de objetos tipo int o float, los textos son strings y las fechas están en formato date. Este dataset se exploró y se interpretó mediante gráficos y correlaciones en la siguiente sección.

En cuanto a los datasets de nivel de embalses, se segmentaron para las represas del departamento de Cundinamarca, específicamente CHUZA y AGREGADO BOGOTÁ[9]. Este dataset es un conjunto de registros diarios sin valores nulos o duplicados. El único filtro realizado fue una agrupación por fecha debido a que hay 2 embalses, esta agrupación permite que los datos sean compatibles o integrables con los datos de precipitación por fecha al momento de su fusión para su posterior correlación y análisis exploratorio.

Para ambos conjuntos de datos se aplicó la función MinMaxScaler de la librería scikit-learn. Además, en los siguientes pasos para la búsqueda o ajuste de un modelo, se planea dividir el dataset en un 90% para entrenamiento y un 10% para pruebas o validaciones.

## Analítica descriptiva

### precipitaciones

Para proporcionar un análisis descriptivo del conjunto de datos de precipitaciones, se calcularon las estadísticas clave utilizando la función describe de pandas:

|  |  |
| --- | --- |
| **métrica** | **Precipitación [mm]** |
| count | 1401 |
| mean | 0.218740 |
| min | 0.0 |
| 25% | 0.065753 |
| 50% | 0.177150 |
| 75% | 0.326431 |
| max | 1.742690 |
| STD | 0.082 |

Tabla 4. Estadística descriptiva del conjunto de datos de precipitaciones.

la precipitación media es de 0.218740 mm. La precipitación mínima registrada es de 0.0 mm, indicando días sin lluvia, mientras que la máxima es de 1.742690 mm. Los percentiles muestran que el 25% de las precipitaciones son menores a 0.065753 mm, el 50% (la mediana) es de 0.177150 mm, y el 75% de los valores están por debajo de 0.326431 mm. La desviación estándar es de 0.082 mm, lo que sugiere una variabilidad relativamente baja en las precipitaciones diarias.

A graph with blue dots and lines

Description automatically generated

Figura 1. ACF de precipitación.

El gráfico de la función de autocorrelación (ACF) de los datos de precipitaciones revela distintos patrones dentro de la serie temporal. A lo largo del eje horizontal, se muestran varios retardos de la serie, mientras que el eje vertical cuantifica el coeficiente de autocorrelación, que denota la relación entre los datos actuales y sus valores precedentes. Las barras verticales representan estos coeficientes para cada retardo, y su altura indica la magnitud de la correlación. La sombra azul que rodea la línea cero representa el intervalo de confianza, y las barras que se extienden más allá de esta sombra indican una autocorrelación significativa con esos retardos.

En este caso, se observa una tendencia notable: a medida que aumenta el retardo, la autocorrelación disminuye progresivamente, lo que indica una influencia cada vez menor de los valores pasados a lo largo del tiempo. No obstante, algunas barras que sobresalen persistentemente fuera del intervalo de confianza sugieren ciclos o patrones duraderos en los datos de precipitación.

A graph of a graph

Description automatically generated

Figura 2 Grafica serie temporal datos de precipitación

Por ultimo para la precipitacion se grafican los datos crudos con un suavizado de una ventana de 30 días, los datos crudos de este dataset contienen una gran cantidad de variabilidad diaria, lo que puede dificultar la identificación de patrones y tendencias significativas. Esta variabilidad, o "ruido", puede ocultar la verdadera naturaleza de los datos, haciendo que las conclusiones basadas únicamente en estos datos sean menos precisas y confiables. Aunque se usa un suavizado, que reduce este ruido, para el dataset de precipitacion que se esta evaluando no se observan patrones especificos que conduzcan a la existencia de periodisidad de los datos, esto tal vez debido a que existe una aleatoriedad de la posicion de las estaciones que miden la precipitacion.

### Volumen útil

Para la descripción básica del dataset de la energía almacenada en los embalses se tiene que la media de los volúmenes es de aproximadamente 1.2e+09 kWh, con un rango que oscila entre un mínimo de 3.0e+08 kWh y un máximo de 2.6e+09 kWh. La dispersión de los datos se refleja en una desviación estándar de alrededor de 6.3e+08 kWh.

|  |  |
| --- | --- |
| **Métrica** | **Volumen [kWh]** |
| count | 2922 |
| mean | 1.226130e+09 |
| min | 3.006187e+08 |
| 25% | 6.510397e+08 |
| 50% | 1.164612e+09 |
| 75% | 1.710071e+09 |
| max | 2.592348e+09 |
| STD | 6.293976e+08 |

Tabla 5. Estadística descriptiva de datos de energía almacenada embalses Cundinamarca.

Al trazar los datos *Figura 3* que representan el volumen útil a lo largo de cuatro años, resulta evidente una estacionalidad anual visualmente coherente. Esta uniformidad en la estacionalidad anual simplifica la tarea de predecir futuros volúmenes de energía almacenada. Ofrece un patrón fiable y predecible que puede modelarse y extrapolarse eficazmente para predecir los próximos niveles de almacenamiento.

A graph with blue line

Description automatically generated

Figura 3. Energía almacenada de embalses Cundinamarca.

A blue graph with numbers

Description automatically generated

Figura 4. ACF de volumen de embalses.

Por último, se realiza un ACF del volumen útil, donde se observa una fuerte autocorrelación positiva en los primeros lags, que gradualmente disminuye, pero permanece significativa a lo largo de muchos días. Este patrón sugiere que el volumen de los embalses tiene una memoria larga, es decir, los volúmenes actuales están fuertemente influenciados por los volúmenes en días anteriores, incluso hasta casi un año atrás. Esto es característico de los sistemas con grandes inercias, como los embalses, donde las variaciones en el volumen tienden a ser suaves y acumulativas debido a la capacidad de almacenamiento y las tasas de entrada y salida de agua relativamente constantes.

### Precio

Con 1461 observaciones, el precio promedio de energía es de 214 COP, con un valor mínimo de 82 COP y un máximo de 686 COP. El análisis de los percentiles muestra que el 25% de los precios se encuentran por debajo de 122 COP, el 50% (la mediana) está en 189 COP, y el 75% por debajo de 279 COP, lo que indica una asimetría hacia precios más altos. Además, la desviación estándar de 110 COP refleja una considerable variabilidad en los precios.

|  |  |
| --- | --- |
| **métrica** | **Precio [COP]** |
| count | 1461 |
| mean | 214 |
| min | 82 |
| 25% | 122 |
| 50% | 189 |
| 75% | 279 |
| max | 686 |
| STD | 110 |

Tabla 6. Estadística descriptiva de los datos de precio de energía en la bolsa nacional.

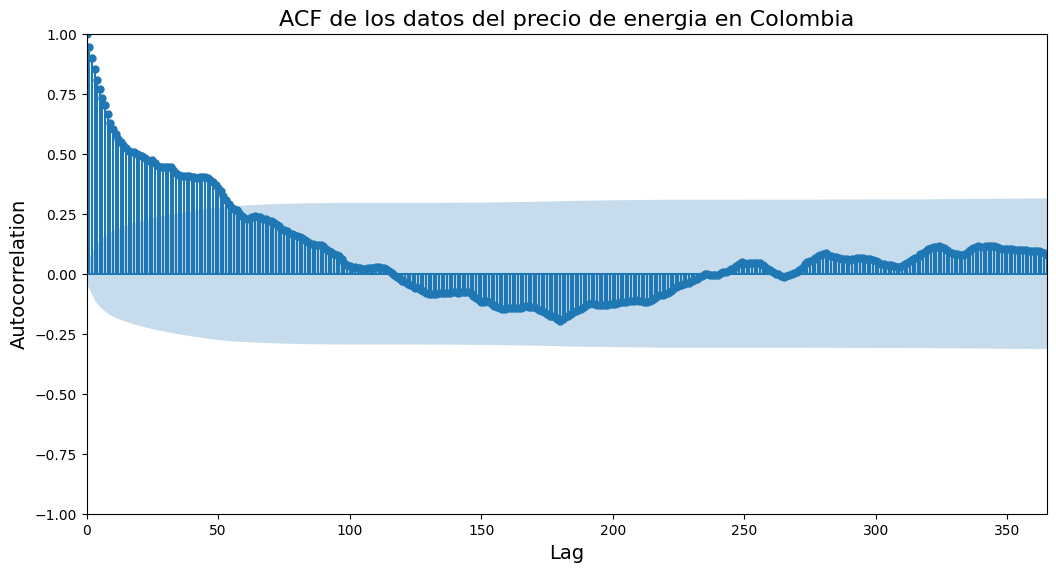


Figura 5. ACF precios de energía.

En el caso de la función de autocorrelación (ACF) de los precios de la energía, se observa una autocorrelación positiva prominente en los rezagos iniciales, que disminuye gradualmente y pasa a valores negativos en torno al rezago 120. Esta tendencia sugiere que los precios de la energía en Colombia están sustancialmente influenciados por valores pasados a corto plazo, con un impacto que disminuye con el tiempo. Las oscilaciones alrededor de cero en los rezagos más largos implican la presencia de posibles patrones estacionales o cíclicos en los precios, aunque no particularmente robustos.

### Correlaciones entre variables de precipitación y volumen útil

A graph with a line

Description automatically generated

Figura 6. Búsqueda de mejor correlación Precipitación Vs Vol. embalses

El proceso de aplicar un desfase temporal entre la precipitación y el volumen útil busca encontrar el retraso óptimo que maximiza la correlación entre ambas variables. Esto se debe a que, al ser series temporales, los efectos de la precipitación en el volumen no son inmediatos y muestran cierta inercia. El valor óptimo de este desfase resultó en un retraso de -46 días en las precipitaciones, lo que generó una correlación inversa entre las dos variables. Sin embargo, es importante señalar que este fenómeno no explica los aspectos físicos subyacentes que se estudian, ya que intuitivamente, a medida que aumenta la cantidad de lluvia, el volumen almacenado debería incrementarse.

A graph with red and blue lines

Description automatically generated

Figura 7. Superposición escalada de series de tiempo precipitación Vs Volumen embalse.

### Correlaciones entre volumen útil y precio

Al analizar la relación entre el conjunto de datos del volumen útil y los precios de la energía, se observa una asociación directa, a diferencia de la relación con las precipitaciones, que requiere transformaciones matemáticas para derivar los caudales de entrada y seleccionar los parámetros adecuados para una correlación precisa. Esta asociación directa entre el volumen almacenado y los precios de la energía conduce a una correlación significativamente mejor. A diferencia de las precipitaciones, que presentan complejidades inherentes a su transformación en caudales de entrada y posterior correlación con el volumen almacenado. Como tal, esta relación directa entre el volumen almacenado y los precios de la energía simplifica y mejora el poder predictivo de los modelos que pretenden pronosticar los precios.

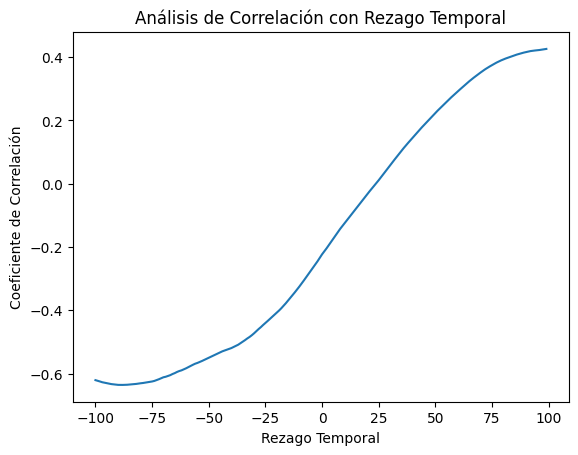


Figura 8. Búsqueda de mejor correlación Vol. embalses Vs Precio.

Al aplicar la función de desfase para determinar el plazo óptimo que correlaciona mejor los precios de la energía, se descubrió que un desfase de -89 días arroja un coeficiente de correlación de hasta 0,64. Este hallazgo subraya la importancia de la dinámica temporal para comprender la relación entre el volumen almacenado y los precios de la energía. Al identificar el desfase que maximiza la correlación, se pueden anticipar mejor las futuras fluctuaciones de los precios de la energía basándose en los datos históricos de los embalses.

A graph with red and blue lines

Description automatically generated

Figura 9. Superposición escalada de series de tiempo Vol. Embalse Vs Precio.

# Referencias

[1] “Precios de electricidad subieron 19,51% anual y la Costa Caribe es donde más crece.” https://www.larepublica.co/economia/precios-de-electricidad-subieron-19-51-anual-y-la-costa-caribe-es-donde-mas-crece-3566130/ (accessed May 27, 2024).

[2] Minenergía, “Colombia apuesta a la aplicación de un estándar mundial de sostenibilidad con el apoyo de la Cooperación Económica y Desarrollo (SECO) de Suiza y la Asociación Internacional de Energía hidroeléctrica (IHA),” *Minenergía*, 2023. https://www.minenergia.gov.co/es/sala-de-prensa/noticias-index/colombia-como-uno-de-los-líderes-latinoamericanos-en-energía-hidroeléctrica-le-apuesta-a-la-aplicación-de-un-estándar-mundial-de-sostenibilidad-con-el-apoyo-de-la-cooperación-económica-y-desar (accessed Apr. 03, 2024).

[3] “Chapter 3: Vector Autoregressive Methods — Time Series Analysis Handbook.” https://phdinds-aim.github.io/time\_series\_handbook/03\_VectorAutoregressiveModels/03\_VectorAutoregressiveMethods.html (accessed May 27, 2024).

[4] “XGBoost Documentation — xgboost 2.0.3 documentation.” https://xgboost.readthedocs.io/en/stable/ (accessed May 27, 2024).

[5] “Prophet | Forecasting at scale.” https://facebook.github.io/prophet/ (accessed May 27, 2024).

[6] J. L. Devore, J. L. Cárdenas, and J. M. T. Flores, *Probabilidad Y Estadística Para Ingeniería Y Ciencias*. CENGAGE Learning, 2015. [Online]. Available: https://books.google.com.co/books?id=vTAXvgAACAAJ

[7] “Geoportal del DANE - Geovisor CNPV 2018.” https://geoportal.dane.gov.co/geovisores/sociedad/cnpv-2018/ (accessed May 27, 2024).

[8] “¿Quiénes somos en XM ?” https://www.xm.com.co/nuestra-empresa/nosotros/quienes-somos (accessed May 27, 2024).

[9] “Embalses - XM.” https://www.xm.com.co/hidrologia/embalses (accessed May 27, 2024).