Imputación de datos faltantes en caudales de fuentes hídricas del departamento de Antioquia: Un análisis comparativo de métodos tradicionales y basados en aprendizaje automático.

Tomás Londoño García

Monografía presentada para optar al título de Especialista en Manejo y Gestión del Agua

Asesor
Juan Camilo Parra Toro, Doctor (PhD) en Ingeniería Hidráulica y Medio Ambiente

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Manejo y Gestión del Agua
Medellín, Antioquia, Colombia
2024

Cita	(Londoño García, 2024)
Referencia **Estilo APA 7 (2020)**	Londoño García, T. (2024). *Imputación de datos faltantes en caudales de fuentes hídricas del departamento de Antioquia: Un análisis de métodos tradicionales y basados en aprendizaje automático* [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

creative commons CC BY NC SA

Especialización en Manejo y Gestión del Agua, Cohorte XXII.

Asesor de trabajo de grado: Juan Camilo Parra Toro
Asesor científico de datos: Álvaro Ramirez Cardona

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

CENDOI

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

Universidad de Antioquia - www.udea.edu.co

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

A mis padres, hermano y amigos que son parte fundamental para disfrutar y sortear la vida.

Agradecimientos

A la familia Ramirez Arango, porque sin ellos nada de esto hubiese sido posible. A mi asesor Juan Camilo Parra Toro, por ser un profesor muy comprensivo, y por alentarme a culminar satisfactoriamente este proceso. A la UdeA por abrirme las puertas para continuar con mi formación académica.

Tabla de contenido

Lista de tablas

Lista de ilustraciones

Lista de gráficas

Siglas, acrónimos y abreviaturas

MCAR	Missing Completely At Random
MAR	Missing At Random
MNAR	Missing Not At Random
KNN	K-Nearest Neighbors
MF	MissForest
RF	RandomForest
MSE	Error cuadrático medio
MAPE	Error medio absoluto porcentual
SIM	Similitud
CV	Coeficiente de variación
IQR	Rango intercuartílico
NSS	Nivel subsiguiente

Resumen

Se evaluaron cinco técnicas de imputación (interpolación lineal, media móvil, spline cúbico, KNN y MissForest) en 9 series de tiempo de caudales horarios de estaciones hidrológicas del departamento de Antioquia, con un 30%, 40% y 50% de datos faltantes simulados. Utilizando seis métricas de evaluación (MSE, similitud, R^2, correlación de Spearman, coeficiente de variación y rango intercuartílico), se encontró que la interpolación lineal fue la técnica más precisa. El spline, si bien puede capturar tendencias a largo plazo, presentó dificultades para modelar picos y cambios bruscos en los caudales horarios. La media móvil, por otro lado, fue efectiva para suavizar la serie y capturar la tendencia general, pero no se adaptó bien a variaciones bruscas en los datos. Los métodos basados en modelos de aprendizaje automático (KNN y MissForest) presentaron un mayor costo computacional y, en general, resultados menos precisos. Estos resultados sugieren que, para series de tiempo de caudales horarios, con las características analizadas, la interpolación lineal podría ser una opción sólida y eficiente para imputar datos faltantes. Sin embargo, se recomienda evaluar otras técnicas en futuros estudios con diferentes características de los datos (como el uso de variables meteorológicas) y con ventanas de valores faltantes del orden de semanas.

Palabras clave: Datos nulos, valores faltantes, imputación de datos, métodos determinísticos y aprendizaje automático.

Abstract

Five imputation techniques (linear interpolation, moving average, cubic spline, KNN, and MissForest) were evaluated on 9 hourly flow time series from hydrological stations in the Antioquia department, with 30%, 40%, and 50% simulated missing data. Using six evaluation metrics (MSE, similarity, R^2, Spearman correlation, coefficient of variation, and interquartile range), it was found that linear interpolation was the most accurate technique. While the spline could capture long-term trends, it had difficulties modeling peaks and abrupt changes in hourly flows. On the other hand, the moving average was effective in smoothing the series and capturing the overall trend but did not adapt well to sudden variations in the data. Machine learning-based methods (KNN and MissForest) had higher computational costs and generally less accurate results. These results suggest that, for hourly flow time series with the analyzed characteristics, linear interpolation could be a solid and efficient option for imputing missing data. However, it is recommended to evaluate other techniques in future studies with different data characteristics (such as the use of meteorological variables) and with windows of missing values on the order of weeks

Keywords: Missing data, data imputation, deterministic methods, machine learning.

1. Introducción

La gestión eficiente del agua es crucial para el desarrollo sostenible y la seguridad hídrica de cualquier región. En este contexto, las series de tiempo de caudales en estaciones hidrológicas juegan un papel fundamental al proporcionar datos esenciales para la planificación y toma de decisiones en la gestión de los recursos hídricos. Sin embargo, es común enfrentarse a la problemática de datos faltantes en estas series de tiempo debido a fallos en los equipos de medición, condiciones meteorológicas adversas o errores humanos. La presencia de datos faltantes puede afectar significativamente la calidad y precisión de los análisis y modelos hidrológicos, comprometiendo la capacidad de realizar predicciones fiables y tomar decisiones informadas.

Para mitigar estos problemas, se han desarrollado diversos métodos de imputación de datos, los cuales buscan estimar y completar los valores faltantes en las series de tiempo. La elección del método de imputación adecuado es de vital importancia, ya que un método inapropiado puede introducir sesgos y errores en los datos imputados, afectando negativamente el análisis subsiguiente. Por tanto, evaluar y comparar diferentes técnicas de imputación es esencial para identificar aquellas que ofrezcan mejores resultados en términos de precisión y eficiencia.

En este estudio, se analizan cinco técnicas de imputación (interpolación lineal, media móvil, spline cúbico, KNN y MissForest) aplicadas a series de tiempo de caudales horarios en estaciones hidrológicas del departamento de Antioquia. Se pretende proporcionar una visión detallada de la efectividad de cada método en la recuperación de datos faltantes y su impacto en la calidad de las series de tiempo, contribuyendo así a mejorar la gestión y análisis de los recursos hídricos.

2. Estado del arte

A continuación, se muestran algunos trabajos en donde se aplican algunas técnicas de imputación de datos para completar datos en series de tiempos hidrológicas y meteorológicas.

A Multi-imputation Method to Deal With Hydro-Meteorological Missing Values by Integrating Chain Equations and Random Forest (Jing et al., 2022)**.**

En este estudio se aborda un método de imputación múltiple para series hidrometeorológicas en tiempo real, mediante la integración de ecuaciones encadenadas y bosques aleatorios, denominado MICE-RF. Este método se compara con otros métodos como LINEAR, MEAN, MEDIAN, PCHIP, KNN, GAN y MICE-PMM para evaluar la robustez, fiabilidad y precisión del método MICE-RF. Según los resultados obtenidos, el método MICE-RF fue el que tuvo un mejor rendimiento de imputación para la mayoría de los escenarios de valores faltantes. Por otro lado, los métodos tradicionales LINEAR y PCHIP mostraron una mejor precisión, para los escenarios con una mayor cantidad de valores faltantes y baja longitud de valores faltantes, donde el coeficiente de correlación entre las variables es menor.

Imputation methods for recovering streamflow observation: A methodological review (Hamzah et al., 2020)**.**

En este artículo se resumen los patrones y mecanismos de los datos faltantes, además, se revisan varias técnicas de imputación de datos que son más convenientes para los análisis de series temporales en el caudal de ríos. Asimismo, se discutieron los enfoques más simples de imputación junto con técnicas más desarrolladas, como el método de imputación determinista basado en modelos y el método de aprendizaje automático. Entre los resultados obtenidos, se estableció que los métodos deterministas basados en modelos producen imputaciones más precisas en comparación con la técnica de eliminación y el método de imputación simple. Por otro lado, se menciona que los métodos que utilizan aprendizaje automático, superan las técnicas de imputación basadas en procedimientos estadísticos en la predicción de datos faltantes de caudal.

A Comparison of Multiple Imputation Methods for Recovering Missing Data in Hydrological Studies (Hamzah et al., 2021)**.**

Los autores compararon el rendimiento de tres (3) métodos de imputación en la predicción de la recurrencia en conjuntos de datos de caudal: imputación robusta por regresión aleatoria (RRRI), k-vecinos más cercanos (k-NN) y árbol de clasificación y regresión (CART). En este sentido, se utilizaron los datos históricos diarios de caudal desde el 2012 hasta el 2014, como conjunto de entrenamiento para evaluar la efectividad de los métodos de imputación en la

resolución de datos faltantes de caudal. Seguidamente, estos tres (3) anteriores métodos se implementaron junto con la regresión lineal múltiple (MLR) para restaurar los valores de caudal en el río objeto de estudio desde 1978 hasta 2016. Los resultados obtenidos, mostraron que la técnica RRRI junto con MLR (RRRI-MLR) tenían los valores más bajos de RMSE y MAPE, superando a todas las demás técnicas probadas para llenar los datos faltantes en conjuntos de datos diarios de caudal. Esto indica que RRRI-MLR es el mejor método para tratar los datos faltantes en conjuntos de datos de caudal.

Comparative assessment of univariate and multivariate imputation models for varying lengths of missing rainfall data in a humid tropical region: a case study of Kozhikode, Kerala, India (Kannegowda et al., 2024)**.**

Para este estudio se compararon modelos de imputación univariados y multivariados para diferentes longitudes de observaciones diarias de precipitación faltantes en una región tropical húmeda, mediante la utilización de 33 años de datos meteorológicos. Se utilizaron algunas medidas de precisión como el Error Cuadrático Medio (RMSE), el Error Absoluto Medio (MAE), la Eficiencia de Nash–Sutcliffe (NSE) y el Error Relativo Absoluto Medio (MARE). Los resultados obtenidos mostraron que los modelos conectados con Suavizado de Kalman (KS), como un modelo univariado representativo, tienen un desempeño excepcional al tratar con un pequeño número de observaciones faltantes. Por otro lado, los modelos multivariados como el Análisis de Componentes Principales (PCA) y el Bosque Aleatorio superaron a los modelos univariados para márgenes de brechas medianos a grandes.

Comparative Study on Univariate Forecasting Methods for Meteorological Time Series (Thi-Thu-Hong, 2018)**.**

Este articulo tiene como principal objetivo establecer un marco para la predicción de series temporales meteorológicas univariadas, para luego llevar a cabo una comparación de rendimiento de diferentes modelos univariados para la tarea de pronóstico en 18 meses. Se utilizaron seis (6) métodos para realizar el pronóstico de estos datos, tales como, Suavizado Exponencial Simple (SES), Naive Estacional (Snaive), ARIMA Estacional (SARIMA), Red Neuronal Feed-Forward (FFNN), Imputación Basada en Dynamic Time Warping (DTWBI) y Serie Temporal Estructural Bayesiana (BSTS). De acuerdo a los resultados obtenidos, mediante el análisis de las métricas,

tales como, similitud, NMAE, RMSE y FB, mostraron que el método FFNN se adapta bien para prever series temporales meteorológicas univariadas con estacionalidad y sin tendencia, en consideración de los índices de precisión, y el DTWBI es más adecuado al considerar la forma y dinámica de los valores pronosticados.

Novel Methods for Imputing Missing Values in Water Level Monitoring Data (Khampuengson & Wang, 2023)**.**

Los autores de este artículo proponen una metodología para la imputación de datos faltantes denominada Correspondencia de Subsecuencia Completa (FSM, por sus siglas en ingles). Esta técnica identifica la secuencia de valores faltantes y los reemplaza con algunos valores constantes para crear una secuencia completa ficticia. Luego, busca la subsecuencia más similar en los datos históricos de la serie de tiempo y esta subsecuencia se adapta para ajustarse a la parte faltante en función de su similitud. La comparación con otros métodos como la interpolación, k-NN, MissForest, y también con un método líder de aprendizaje profundo: la técnica de Memoria a Largo y Corto Plazo (LSTM), mostraron que la técnica FSM puede producir imputaciones más precisas, particularmente para aquellas con patrones periódicos fuertes.

Enhancing hydrological data completeness: A performance evaluation of various machine learning techniques using probabilistic fusion imputer with neural networks for streamflow data reconstruction (Arathy Nair et al., 2024)

Se realiza la estimación de datos faltantes de caudal mediante las técnicas de imputación de datos utilizando enfoques de aprendizaje automático, incluidos K-vecinos más cercanos (KNN), Emparejamiento Predictivo de la Media (PMM), Bosque Aleatorio (RF) y una técnica novedosa de Imputador de Fusión Probabilística con Redes Neuronales (PROFINN). Este estudio explora la eficacia de los modelos seleccionados en varias brechas de datos bajo diferentes escenarios hidrológicos, incluyendo diversas características de flujo (media, flujos altos y bajos) y longitudes de brechas (brechas largas y cortas, brechas continuas y discontinuas) y presenta un sistema de clasificación que evalúa el nivel de idoneidad de cada técnica para diversas brechas de datos. Entre los resultados obtenidos la técnica PROFFIN se destacó en todos los escenarios, logrando un RMSE promedio de 0,91 y un valor de promedio de Eficiencia de Nash-Sutcliffe (NSE) de 0,93 cuando se aplica a un sistema fluvial intermitente del río Pamba en el sur de Kerala, India. Por otro

lado, RF sigue a PROFINN en la imputación de flujos extremos, así como en escenarios de brechas largas y cortas. KNN sigue de cerca al método PROFINN para la imputación de escenarios de brechas continuas y discontinuas.

Imputing Missing Data in Hydrology using Machine Learning Models (Vasker Sharma, 2021).

Se imputaron los valores faltantes de precipitación y temperatura en seis (6) estaciones meteorológicas ubicadas en el este de Butan utilizando varios modelos de aprendizaje automático, como el KNN y el modelo de árbol de decisión. Seguidamente, estos datos imputados se han utilizado como predictores para predecir los datos de caudal del río utilizando una Red Neuronal Artificial (ANN). Con los datos imputados por kNN como vector de entrada, se desarrolló un modelo de Red Neuronal Artificial (ANN) para predecir el caudal en las estaciones de Uzorong y Muktirap. El modelo se entrenó utilizando los datos de entrenamiento y, posteriormente, se probó con los datos de prueba. Basado en los datos de prueba, la desviación absoluta media para el flujo en Uzorong fue del 0.054%, mientras que para Muktirap fue del 0.088%. En consecuencia, la precisión del modelo fue del 94.53% y 91.11% para Uzorong y Muktirap, respectivamente.

Comparing Single and Multiple Imputation Approaches for Missing Values in Univariate and MultivariateWater Level Data (Umar & Gray, 2023)

Este artículo evalúa métodos de imputación simples y múltiples utilizados en datos mensuales univariantes y multivariantes del nivel del agua de cuatro estaciones de agua en los ríos Benue y Níger en Nigeria. Se consideraron los mecanismos de datos faltantes completamente al azar, faltantes al azar y faltantes no al azar. El mejor método de imputación se identificó utilizando dos métricas de error: el error cuadrático medio (RMSE) y el error porcentual absoluto medio (MAPE). Se encontró que para el método univariado de descomposición estacional es el mejor para imputar los tres mecanismos o tipos de datos faltantes, seguido del método de suavizado de Kalman, mientras que la imputación aleatoria mostró los peores resultados. Por otro lado, el método multivariado MissForest es la mejor técnica, seguida por el método KNN, para los tipos de datos faltantes completamente al azar y faltantes al azar. En cuando a los tipos de datos faltantes que no son al azar, el método el método KNN presentó mejores resultados que el MissForest. En cuanto a los otros métodos como RandomForest y predictive mean matching tienen un rendimiento

deficiente en termino de las dos métricas evaluadas. Por último, los resultados indicaron que el método de descomposición estacional, y los métodos missForest o k-nearest neighbour, pueden imputar datos faltantes univariantes y multivariantes del nivel del agua, respectivamente, con mayor precisión que los otros métodos considerados.

Estimación e imputación de datos faltantes mediante métodos de interpolación espacial para precipitación mensual acumulada en el departamento de Antioquia durante el periodo 2014-2018 (Sánchez, 2020)**.**

En este trabajo se comparan diferentes métodos de interpolación espacial para valores faltantes de precipitación mensual acumulada en el departamento de Antioquia durante los años 2014-2018, provenientes del IDEAM. Los métodos utilizados fueron la distancia inversa ponderada (IDW, por sus siglas en ingles), Spline de placa delgada (TPS, por sus siglas en ingles) y kriging ordinario. Estos métodos se evaluaron mediante el error cuadrático medio (MSE, por sus siglas en inglés). Los resultados encontrados mostraron que el método kriging funciona bien cuando se trata de un porcentaje de datos faltantes mayores al 10% y el método IDW cuando se tiene 5% de datos faltantes.

Metodología para la imputación de datos faltantes en meteorología (Urrutia et al., 2010)**.**

Los autores de este artículo presentaron una metodología para realizar la imputación de datos faltantes en series de tiempo de precipitación y/o temperatura, mediante el uso de correlaciones parciales, modelos de regresión, ajuste de datos por medio del método de doble masa y verificación de la tendencia mediante el test de Kendal. Los resultados evidenciaron que esta metodología funciona bien para un 20% de datos faltantes. Sin embargo, para datos faltantes mayores al 20% se recomienda utilizar otros modelos AMI bivariada para la imputación de los mismos, ya que los modelos de regresión lineal no son robustos para estos casos.

Imputation of spatial air quality data using gis-spline and the index of agreement in sparse urban monitoring networks(Londoño-Ciro & Cañón-Barriga, 2015)**.**

Este estudio establece una metodología para realizar la imputación de datos de calidad de aire en zonas urbanas para tres contaminantes atmosféricos (NO_2, PM10 y PST), mediante la

implementación de Sistemas de Información Geográfico (SIG) y dos (2) técnicas de interpolación IDW y spline. El procedimiento de imputación de datos propuesto está constituido por la técnica de interpolación espacial, un proceso de validación cruzada con el índice de (IOA), y por último el análisis de la densidad de muestreo y del coeficiente de variación con diferentes estadísticos de error. Los resultados mostraron que la técnica spline obtuvo un mejor rendimiento respecto a la técnica IDW para interpolar variables de calidad del aire dentro del dominio y espacio establecido. Asimismo, la técnica spline también tuvo ventajas sobre la técnica IDW, cuando se trataba de interpolar datos en las estaciones ubicadas en el borde del área de interpolación. Este artículo también propone que los mapas de interpolación creados durante la aplicación de la metodología se pueden complementar con los mapas de gradiente y gradiente direccional con el fin de establecer puntos críticos de muestreo de calidad de aire, que permitan ampliar las redes de monitoreo de calidad de aire en ciertos puntos de interés.

Imputación de Datos en Series de Precipitación Diaria Caso de Estudio Cuenca del Río Quindío (García Reinoso, 2015)**.**

Se llevó a cabo la imputación de datos de precipitación diaria mediante la implementación de cinco (5) métodos de interpolación ponderados en ocho estaciones ubicadas en la cuenca del río Quindío. La comparación de los estadísticos de la serie de precipitación original y la imputada demostraron que los métodos de la Media Estadística Ponderada obtuvieron un mejor desempeño y ventajas respecto a las demás técnicas, ya que conserva adecuadamente las medidas de tendencia central de la serie temporal de precipitación diaria con datos faltantes.

Técnicas de imputación para datos de precipitación máxima mensual en la zona central de Boyacá (Bello et al., 2019).

Se realizó la imputación de datos máximos de precipitación en la región de Boyacá para el periodo de 1974-2013. Inicialmente, se evaluó el desempeño de los mecanismos de imputación de perdida de datos MCAR, MAR o MNAR, y cada uno de estos se implementó la imputación de datos mediante tres (3) técnicas: imputación múltiple, KNN y suavizado de Kalman. Los resultados obtenidos al realizar algunas comparaciones como los estadísticos descriptivos de la serie original y la serie imputadas, se logró concluir que la mejor técnica fue la imputación múltipleizado previamente y que guardan una relación histórica con el tema de investigación actual.

3. Pregunta de investigación

¿Cuál de los métodos de imputación univariados y multivariados evaluados presenta un mejor desempeño en la reconstrucción de series de tiempo hidrológicas con datos faltantes, en términos de precisión y eficiencia computacional?

4. Objetivo general

Evaluar y comparar la efectividad de algunas técnicas de imputación de datos en series de tiempo hidrológicas, con el fin de reducir la incertidumbre en la estimación de caudales y mejorar la toma de decisiones.

4.1. Objetivos específicos

- Seleccionar las estaciones más adecuadas para simular los escenarios de valores faltantes en la región de estudio.
- Implementar 5 técnicas de imputación de datos en series de tiempo hidrológicas.
- Comparar el desempeño de estas técnicas y evaluar su capacidad para imputar los datos de acuerdo con la naturaleza de las series de tiempo.
- Establecer las ventajas y desventajas de las diferentes técnicas de imputación de datos.
- Determinar la técnica de imputación de datos más eficiente para datos.
- Indagar sobre posibles trabajos futuros en el campo de la imputación de datos hidrológicos

5. Marco teórico

Durante el análisis de datos contenidos en series de tiempo, es común que por diversos motivos se presente perdida de información, lo que puede dificultar la utilización de dichos datos y la toma de decisiones en determinado momento. Una forma de abordar este problema es rellenar o imputar los datos faltantes mediante diferentes técnicas, de tal forma, que se pueda tener continuidad en la información. Estos métodos también pueden ser utilizados para pronosticar datos futuros, sin embargo, en el alcance de este documento solo se abordará para completar datos faltantes en series de tiempo hidrológicas.

5.1. Tipos de datos faltantes.

El primer paso para llevar a cabo la imputación de datos faltantes en una serie de tiempo, nunca debe ser la imputación directa, inicialmente se debe indagar sobre la naturaleza de los datos faltantes, es decir, cual es la tipología de acuerdo a la causa y distribución de estos datos faltantes. De manera general, los datos faltantes pueden ser de tres (3) tipos:

5.1.1. MCAR (Missing Completely At Random)

Este tipo de datos faltantes hace referencia a que los valores ausentes en las series de tiempo se encuentran distribuidos completamente al azar y no dependen o se pueden explicar por ninguna de las variables observadas o no observadas en el conjunto de datos. En este sentido, cualquier observación tiene la misma probabilidad de perderse (Trejos Hernández & Villada Lizarazo, 2019).

De acuerdo con (Hamzah et al., 2020), en series de tiempo univariadas como el caso de datos de flujo de una corriente de agua o caudal, no existen otras variables excepto el tiempo como variables implícitas.

$$P\left(m \middle| Y_{(observado),}\, Y_{(faltante)}\right) = P(m)$$

Un ejemplo de este tipo de datos faltantes sucede cuando hay perdida de registros de medición debido a fallos de un sensor que ocurren aleatoriamente, sin relación con las condiciones medidas ni con otras variables del estudio. Para este tipo de datos faltantes la imputación es más sencilla y cualquier método de imputación estándar puede ser aplicado sin introducir sesgos significativos.

Basándonos en la definición de MCAR propuesta por (Rubin & Little, 2002) el valor faltante en el estudio de caudales, se determina como MCAR ya que el episodio de ausencia de datos en el flujo de corriente de un área no está influenciado por los datos en esa área ni en ninguna otra área.

5.1.2. MAR (Missing At Random)

De manera similar a MCAR, los datos faltantes también son al azar, pero su ausencia está relacionada con otras variables observadas en el conjunto de datos. Para series temporales univariadas donde no hay otras variables excepto el tiempo (implícitamente dado), la probabilidad de que una observación falte en MAR depende del momento o estado del tiempo de dicha observación en la serie (Hamzah et al., 2020).

$$P\left(m\middle|Y_{(observado),}\, Y_{(faltante)}\right) = P(m|Y_{(observado)})$$

Por ejemplo, si las mediciones de caudal faltan más frecuentemente durante los días con lluvia intensa, pero estos no dependen de variables no observadas. Para este caso, se pueden aplicar métodos de imputación más complejos, utilizando la información de las variables observadas relacionadas para reducir el sesgo.

5.1.3. MNAR (Missing Not At Random)

En relación a este último caso, los datos faltantes no son al azar y tampoco es predecible de otras variables del conjunto de datos, es decir, su ausencia está relacionada con la propia variable faltante o con variables no observadas (Trejos Hernández & Villada Lizarazo, 2019).

$$P\left(m\middle|Y_{(observado),}\, Y_{(faltante)}\right) = P(m|Y_{(observado)}, Y_{faltante})$$

Pensando en un escenario de caudales extremos (altos o bajo), los sensores tienden a fallar de una forma más frecuente, y la probabilidad de falta de datos depende del mismo caudal. Este corresponde, al caso más complicado de tratar. Las técnicas estándar de imputación pueden introducir sesgos significativos y puede ser necesario utilizar métodos especializados o recopilar más datos para modelar adecuadamente los valores faltantes.

5.2. Técnicas de imputación de datos.

De manera general, las técnicas de imputación de datos pueden ser univariadas o multivariadas. La primera se refiere a la estimación de datos faltantes a partir de la misma serie de datos y la segunda realiza la imputación de datos mediante otras variables que se encuentran relacionadas con esta serie de datos (Kannegowda et al., 2024).

Otra forma para clasificar las técnicas de imputación de datos según (García Reinoso, 2015) son los 1) métodos determinísticos, lo cuales consideran un modelo matemático que produce una

respuesta única de acuerdo a las condiciones iniciales ingresadas. 2) métodos estocásticos, que estiman los valores mediante distribuciones probabilísticas. 3) métodos de inteligencia artificial, que se basan en modelos de aprendizaje automático, los cuales se entrenan para para predecir la variable dependiente optimizando una métrica objetivo.

Las técnicas de imputación de datos que se abordaran en este documento corresponden a algunas técnicas determinísticas y otras que funcionan mediante la inteligencia artificial.

5.2.1. Técnicas determinísticas.

Los métodos determinísticos para la imputación de datos son aquellos que utilizan un modelo matemático y/o reglas fijas, que siempre genera la misma respuesta de acuerdo a la información inicial que es ingresada.

5.2.1.1. Interpolación lineal.

Este método de imputación de datos faltantes es una de las más rápida y sencillas, la cual consiste en trazar una línea recta que separa los valores observados antes y después del intervalo, para posteriormente estimar los datos faltantes mediante la interpolación (Hamzah et al., 2020).

De manera general, la fórmula utilizada para realizar la interpolación lineal de un valor M_2, es la siguiente:

N_1	M_1
N_2	M_2
N_3	M_3

Ilustración 1.
Ubicación del valor faltante a imputar y sus valores adyacentes.

$$M_2 = \frac{(N_2 - N)(M_3 - M)}{(N_3 - N_1)} + M_1$$

5.2.1.2. Interpolación polinómica (spline).

Esta es una técnica de interpolación polinómica, la cual utiliza funciones polinómicas a tramos de bajo grado que se ajusta a los valores observados y que son acordes al intervalo que representa, en lugar de usar una sola función general para todo el intervalo de puntos, como es el caso de la interpolación polinómica simple (Sánchez Quiroga, 2020). Este método de interpolación minimiza la curvatura general de la superficie a aproximar, resultando en una superficie suave que pasa exactamente por los puntos deseados (Chica Jimenez, 2018). Es importante mencionar, que esta técnica no es adecuada si hay grandes cambios en la superficie en una distancia corta, porque puede sobrestimar los valores imputados o generar sesgos no deseados (Ruelland et al., 2008).

La interpolación por spline suele preferirse a la interpolación polinómica, ya que intenta evitar el efecto de oscilación que puede observarse con funciones polinómicas de grados altos (Little & An, 2004). La función spline está restringida en puntos definidos (técnica local). Se debe considerar un número específico de valores vecinos; por lo tanto, el spline puede ajustarse a anomalías locales sin afectar los valores de interpolación en otros puntos de la región global. Las restricciones *r* están dadas por el grado *m* de la función polinómica (Hamzah et al., 2020)

- Si $r = 0$, no hay restricción.
- Si $r = 1$, la función debe ser continua.
- Cuando $r = m + 1$, la *m*-ésima derivada de la función, debe ser continua en todos los puntos.

Como se mencionó anteriormente, el spline puede considerar varias funciones polinómicas por tramos que se ajustan para formar una función continua. En este sentido, un spline para $m = 1$ se llama lineal para $m = 2$, es cuadrático, y para $m = 3$, es una función cúbica (Hamzah et al., 2020).

5.2.1.3. Media móvil (Moving Window).

El modelo de media móvil o ventana deslizante mantiene estadísticas sobre el flujo de datos acerca de los últimos *N* elementos o los elementos que llegan en las últimas *N* unidades de tiempo. El esquema de ventana deslizante combina las características del modelo de ventana deslizante y del esquema. Lo anterior, resuelve el problema de la memoria insuficiente dentro de un nivel aceptable de precisión. Al tomar el momento actual como el punto final, el boceto de ventana deslizante mantiene los datos que necesitan atención en una ventana de tamaño predefinido. Solo

los datos dentro de la ventana serán medidos mientras que los datos fuera de la ventana serán considerados expirados. Con el tiempo, nuevos datos entrarán en la ventana y los datos antiguos serán eliminados. Dependiendo de la dimensión de medición, el modelo de ventana deslizante se puede dividir en dos tipos: basado en conteo y basado en tiempo (Zeng et al., 2023).

5.2.2. Técnicas basadas en inteligencia artificial.

Los algoritmos de aprendizaje automático construyen modelos predictivos que, a partir de los datos existentes, generan estimaciones de los valores faltantes. Estos modelos se entrenan iterativamente con el objetivo de minimizar el error de predicción y obtener resultados cada vez más precisos

5.2.2.1. K-Nearest Neighbors (K-NN)

Este método también es conocido como imputación por vecinos más cercanos. Funciona bajo la suposición de que los puntos de datos vecinos pertenecen a la misma clase. Es decir, es más probable que un dato se estime con sus vecinos *k* más cercanos, que con un valor más distante en la serie de tiempo. El k-NN identifica los puntos vecinos a través de una medida de distancia, y los valores faltantes pueden ser estimados utilizando los valores de las observaciones vecinas (Khampuengson & Wang, 2023).

Algunas de las medidas de semejanza que utiliza esta técnica según (Bello et al., 2019) son las siguientes:

- ✓ **Distancia Euclidiana:**

$$d(x,y) = \sqrt{\sum i\ |x_i - y_i|^2}$$

- ✓ **Distancia de Manhattan:**

$$d(x,y) = i\ |x_i - y_i|^2$$

- ✓ **Distancia de Minkowski:**

$$d(x,y) = \left(\sum i|x_i - y_i|^p\right) \wedge (\frac{1}{p})$$

Entre los principales beneficios que conlleva la utilización de esta técnica es que es fácil de implementar y no requiere ninguna suposición sobre la distribución de los datos. Por otro lado, la principal desventaja radica en que es computacionalmente costoso, ya que se tiene que calcular la similitud entre todos los pares de registros del conjunto de datos. De igual forma, este método es muy sensible a los datos atípicos y al ruido, lo que puede aumentar el sesgo en la imputación de los datos faltantes (Freire Míguez, 2022).

La elección del *k* determina la cantidad de vecinos más cercanos que se utilizaran para la imputación de los datos faltantes. De manera general, los valores pequeños de *k* pueden llevar a imputaciones más localizadas pero más ruidosas, caso contrario sucede cuando se selecciona un *k* grande, lo que puede llevar a una imputación de datos más suaves, pero también pueden incluir vecinos que son menos relevantes (Arathy Nair et al., 2024b).

5.2.2.2. Miss Forest (MF).

Esta es una técnica multivariada de aprendizaje automático, la cual se basa en la técnica del Bosque Aleatorio (Random Forest o RF), que utiliza una combinación de árboles de decisión y muestreo aleatorio para hacer predicciones, mientras que la técnica Miss Forest (MF), es una implementación especifica de esta técnica para imputar datos faltantes (Kannegowda et al., 2024).

De manera general, el algoritmo de Random Forest se puede describir de la siguiente forma: i) Crea un conjunto de árboles de decisión basado en un subconjunto aleatorio de los datos, ii) Para cada árbol, muestrea aleatoriamente las variables predictoras y crea nodos de decisión basados en la mejor división entre las variables disponibles, iii) Utiliza los árboles de decisión resultantes para hacer predicciones sobre nuevos datos, tomando el promedio de las predicciones de todos los árboles (Stekhoven & Bühlmann, 2012).

Según (Zhou et al., 2023) el método de Miss Forest es el método mejorado de bosque aleatorio y se puede describir de manera general, tomando como ejemplo series de tiempo hidrológicas de la siguiente forma:

Para *n* estaciones de aforo con *m* series diarias de caudal, se podría producir una matriz de datos $X = X_{11}, \dots, X_{si}, \dots, X_{nm}$ $(i = 1 \sim m, s = 1 \sim n)$. Para una estación arbitraria X_s que incluye datos faltantes en días i_{mis}, X se separa en cuatro partes:

1. y_s^{obs}: los valores observados de caudal en la estación X_s.

2. y_s^{miss}: los valores faltantes o valores predichos en la estación X_s.
3. x_s^{obs}: el caudal observado en otra estación en los días $\{1, \dots, m\}/i_{mis}$
4. x_{mis}: el caudal faltante en otra estación en los días i_{mis}

Mediante el entrenamiento de un modelo de bosque aleatorio (RF) utilizando y_s^{obs} y X_s^{obs}, el valor y_s^{mis} se podría obtener aplicando el modelo RF entrenado a x_s^{mis}.

El primer paso en MF es ordenar las variables X_s según una cantidad descendente de valores faltantes, rellenándolos usando la media de y_s^{obs} o cualquier otro método de imputación basado en la conjetura inicial. Posteriormente, se inicia con la estación con menor cantidad de datos faltantes X_s, los valores faltantes y_s^{mis} se completan usando un modelo de RF entrenado para x_s^{mis}. Este procedimiento de imputación se itera hasta que la diferencia (Δ) entre la nueva matriz de datos imputados y la anterior aumente por primera vez (Stekhoven & Bühlmann, 2012).

5.2.3. Técnicas estocásticas.

Los métodos estocásticos para llevar a cabo la imputación de datos faltantes generalmente incorporan modelos estadísticos para estimar y reemplazar los valores ausentes en un conjunto de datos. Estos métodos se basan en la relación entre los datos observados y las propiedades estadísticas de estos mismos para general estimaciones razonables de los valores faltantes.

5.3. Métricas utilizadas para la evaluación de las técnicas de imputación de datos hidrológicos.

Para llevar a cabo la evaluación de la exactitud los diferentes métodos de imputación de datos faltantes respecto a la serie de tiempo original, se pueden emplear algunas métricas que comúnmente se usan para estos casos como lo son: error cuadrático medio (MSE), error promedio absoluto (MAE), similitud (Sim), coeficiente de variación (CV), rango intercuartílico (IQR) y el coeficiente de Spearman (ρ).

5.3.1. Error cuadrático medio (MSE).

El error cuadrático medio es la desviación estándar de los residuos (errores de predicción). Los residuos son una medida de qué tan lejos están los puntos de datos de la línea de regresión; MSE es una medida de la dispersión de estos residuos. En otras palabras, dice qué tan concentrados están los datos alrededor de la línea de mejor ajuste y permite cuantificar la magnitud de la desviación de los valores simulados respecto a los observados (Trejos Hernández & Villada Lizarazo, 2019). La técnica con un MSE más bajo sería la más precisa. El MSE se define de la siguiente forma:

$$MSE\ (\hat{y}, y) = \frac{1}{T}\sum_{i=1}^{T}(\hat{y}_i - y_i)^2$$

Donde *T,* es el número de valores faltantes.

5.3.2. Error medio absoluto porcentual (MAPE).

La métrica MAPE (Mean Absolute Percentage Error, por sus siglas en inglés) es una medida de precisión utilizada para evaluar la exactitud de los modelos de predicción. Se calcula como el promedio del valor absoluto del error porcentual entre los valores predichos y los valores reales. El MAPE se calcula de la siguiente forma:

$$MAPE\ (\hat{y}, y) = \frac{1}{T}\sum_{i=1}^{T}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$

Donde *T* es el número de observaciones; y_i el valor real; $\hat{y}_i$ el valor pronosticado.

5.3.3. Similitud (Sim).

Esta métrica define el porcentaje de valores similares entre los valores imputados $\hat{y}$ y los valores reales y.

$$Sim\ (\hat{y}, y) = \frac{1}{T}\sum_{i=1}^{n}\frac{1}{1 + \frac{|\hat{y}_i - y_i|}{\max(y) - \min\ (y)}}$$

Donde *T,* es el número de valores pronosticados. Una mayor similitud, es decir, un valor de Sim en el intervalo [0;1], resalta una mejor capacidad del método para la tarea de pronóstico (Thi-Thu-Hong, 2018).

5.3.4. Coeficiente de variación (CV).

Esta métrica evalúa la dispersión de los datos faltantes imputados al medir la dispersión relativa de los datos imputados respecto a los datos originales. En este sentido, este coeficiente se utiliza para evaluar la variabilidad de los datos en relación a los datos originales, mediante la siguiente expresión:

$$CV = \frac{\sigma}{\mu\,(^1/_2)}$$

Donde σ es la desviación estándar y μ es la mediana de los datos.

Calcular la relación del Coeficiente de Variación basado en la Mediana de los valores reales y los valores imputados proporciona una forma práctica de evaluar la consistencia de la imputación de datos en términos de variabilidad relativa y ayuda a identificar si el proceso de imputación ha alterado significativamente la dispersión de los datos.

- Cuando la relación > 1: Indica que la dispersión relativa de la serie original es mayor que la de la serie imputada. Es decir, los datos originales tienen una mayor variabilidad relativa a su mediana en comparación con las predicciones.
- Cuando la relación < 1: Indica que la dispersión relativa de la serie original es menor que la de la serie imputada. Esto sugiere que las predicciones son más variables en relación con su mediana en comparación con los datos originales.
- Cuando la relación = 1: Indica que la dispersión relativa de ambas series es igual. Esto significa que, en términos relativos, ambas series tienen la misma variabilidad con respecto a sus medianas.

5.3.5. Rango intercuartílico (IQR).

Es una métrica de dispersión que se utiliza para describir la variabilidad de los datos y es especialmente útil para comparar técnicas de imputación de datos faltantes. El IQR mide la

amplitud de la distribución de los datos y proporciona una indicación de la variabilidad de los datos sin ser afectado por valores atípicos. El IQR se calcula como la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1), de la siguiente forma:

$$IQR = Q3 - Q1$$

Donde, Q1 es el primer cuartil (el 25% de los datos están por debajo de este valor) y Q3 es el tercer cuartil (el 75% de los datos están por debajo de este valor).

Calcular la relación del Rango Intercuartílico (IQR) de los valores reales con el IQR de los valores imputados es otra forma práctica de evaluar la consistencia de la imputación de datos en términos de variabilidad en el IQR.

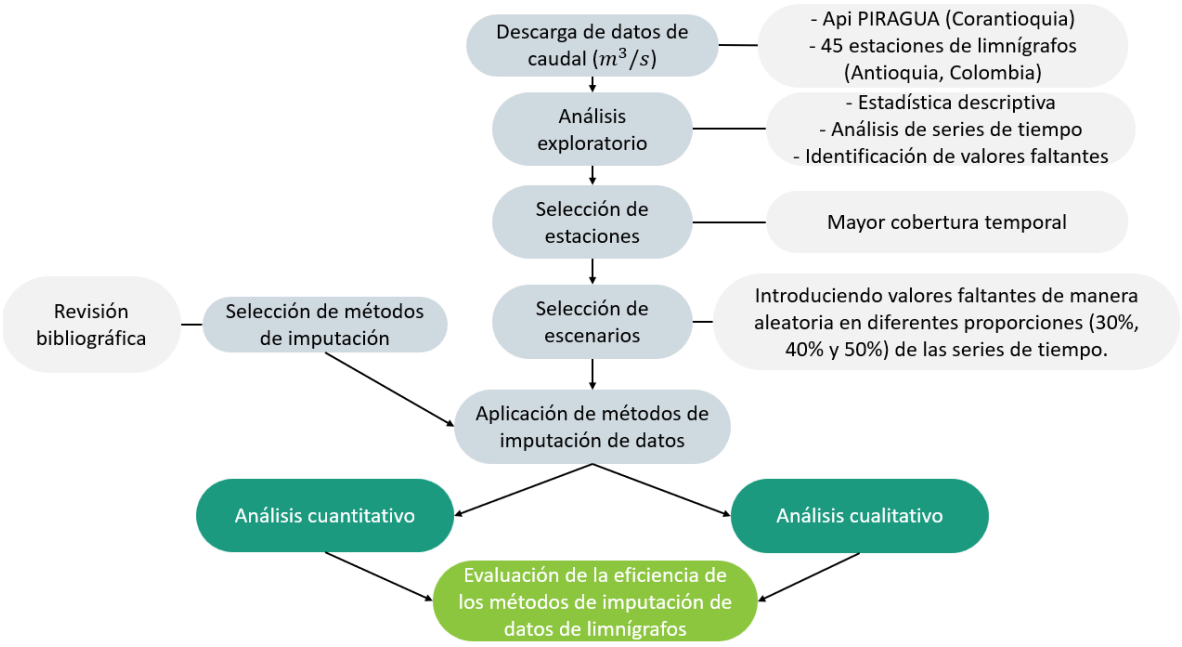
- Cuando la relación > 1: Indica que la dispersión central de los datos originales es mayor que los datos imputados. Los datos originales son más variables en su rango intercuartílico en comparación con las predicciones.
- Cuando la relación < 1: Indica que la dispersión central de los datos originales es menor que los datos imputados. Esto sugiere que las predicciones son más variables en su rango intercuartílico en comparación con los datos originales.
- Cuando la relación = 1: Indica que la dispersión central de ambas series es igual. Esto significa que ambas series tienen la misma variabilidad en su rango intercuartílico.

5.3.6. Coeficiente de Spearman.

Esta es una medida estadística que indica la correlación en términos de fuerza y dirección de la asociación entre dos variables en una escala ordinaria. Es decir, se utiliza para determinar si existe una relación monotónica entre dos conjuntos de datos, es decir, si uno tiende a aumentar a medida que el otro también lo hace, aunque no necesariamente de manera lineal. En cuanto a la interpretación de esta métrica, un valor de Spearman (ρ) igual a 1 indica que la técnica de imputación de datos conserva las relaciones monótonas entre las variables, sugiriendo que es efectiva

6. Metodología

Esta sección detalla las bases de datos utilizadas en esta investigación y el tratamiento aplicado para explorar los diferentes métodos de imputación empleados en esta monografía. De manera general, la metodología propuesta se puede resumir en la Gráfica 1.



Gráfica 1.
Metodología propuesta para la imputación de datos faltantes.

6.1. Recolección de datos.

Los datos hidrológicos (limnígrafos) se obtuvieron desde la plataforma oficial de PIRAGUA, una iniciativa de monitoreo hidrometeorológico gestionada por Corantioquia, a través de su API. Se descargó todo el historial disponible de las 45 estaciones hidrológicas del departamento de Antioquia.

6.2. Procesamiento de Datos

Para el procesamiento de datos se utilizó el lenguaje de programación Python, empleando librerías como numpy, pandas, scipy, scikit-learn y geopandas. Estas herramientas fueron esenciales para la descarga, procesamiento y visualización de los datos.

6.3. Análisis Exploratorio de datos

Se realizó un análisis descriptivo de las series de tiempo para evaluar la disponibilidad de datos y aspectos fundamentales de las series, como el número total de datos, media, desviación estándar, mínimos y máximos. La disponibilidad de datos se evaluó identificando las fechas más recientes y antiguas disponibles para cada estación, y cuantificando el porcentaje de valores faltantes. Esta información fue crucial para seleccionar las estaciones con mayor disponibilidad de datos y definir los periodos de análisis. Además, permitió identificar las particularidades de cada afluente, facilitando la creación de escenarios de imputación tanto univariados (basados en la información de un solo afluente) como multivariados (considerando la información de múltiples afluentes.

La metodología empleada en este análisis es transparente y fácilmente replicable. Se detalla el proceso completo, desde la selección del conjunto de datos y la generación de escenarios de datos faltantes (utilizando una función de eliminación aleatoria con semilla fija) hasta la configuración de los parámetros de cada método (Tabla 1, columna 6).

6.4. Generación de Escenarios Realistas para la Imputación de Datos

Se generaron escenarios realistas introduciendo valores faltantes en distintas proporciones (30%, 40% y 50%) de las series de tiempo seleccionadas. Para esto, se diseñó una función en Python que permite introducir de manera aleatoria y controlada un porcentaje determinado de valores faltantes en las series de tiempo, simulando así la pérdida de información que comúnmente ocurre en valores faltantes tipo MCAR. Se destacan las siguientes características:

- **Aleatoriedad:** La selección de los índices a eliminar se realiza de manera aleatoria, asegurando que los valores faltantes se distribuyan de forma no sistemática a lo largo de la serie de tiempo. Esto se logra mediante la generación de índices aleatorios de los datos a eliminar.
- **Equidad:** Al no alterar el orden de los valores faltantes, se garantiza que los patrones temporales y estacionales de la serie original se mantengan en los escenarios simulados.
- **Porcentaje Variable:** La función permite especificar diferentes porcentajes de datos faltantes (como 30%, 40% y 50%). Esto permite evaluar la robustez de los métodos de imputación bajo diversas condiciones de pérdida de datos, proporcionando un análisis más completo y exhaustivo de la efectividad de los métodos en diferentes escenarios de datos faltantes.
- **Semilla (flag):** Este valor inicializa el generador de números aleatorios, asegurando que los mismos datos sean eliminados en cada ejecución para simular los valores faltantes. De esta manera, se puede replicar el análisis y verificar los resultados obtenidos.

6.5. Selección de métodos de imputación de Datos

Los métodos de imputación se seleccionaron basándose en la literatura consultada, procurando incluir tanto métodos determinísticos, como la regresión lineal, como métodos más robustos asociados al machine learning, como MissForest y k-NN. Estos métodos abarcan enfoques univariados y multivariados, y cuentan con un sólido respaldo en librerías de Python, garantizando la confiabilidad de los análisis.

Tabla 1.
Resumen de los métodos de imputación implementados.

Método	Descripción	Enfoque	Costo Computacional	Librerías Usadas	Parametrizaciones del método
Regresión lineal	Método clásico basado en la relación lineal entre la variable a imputar y otras variables	Univariado	Bajo	scikit-learn	Se garantiza que toda la serie quede completa, sin ningún valor faltante, mediante la combinación de interpolación lineal

					y relleno hacia adelante y hacia atrás.
Media móvil	Método simple que utiliza la media de un conjunto de observaciones adyacentes para estimar valores faltantes	Univariado	Bajo	scipy, pandas	Se aplicó una media móvil simple de 72 horas para imputar los datos faltantes.
K-Nearest Neighbors (KNN)	Método basado en la distancia entre las observaciones, donde los valores faltantes se estiman a partir de los valores de las k observaciones más cercanas	Univariado	Alto	scikit-learn	Se entrenó un modelo de KNN utilizando 5 vecinos más cercanos. Se calculó la distancia euclidiana entre los registros, considerando tanto el caudal como la fecha. Previamente, se estandarizaron los datos.
MissForest	Método de imputación múltiple basado en bosques aleatorios, que captura relaciones no lineales y heterogeneidades en los datos	Multivariado	Muy Alto	missingpy,	Se entrenó un modelo Random Forest con un mínimo de una variable independiente para realizar las predicciones, limitando el proceso a 5 iteraciones.
Splines	Método de interpolación que ajusta una curva suave a los datos observados, permitiendo capturar patrones complejos	Univariado	Medio	scipy, pandas	Se utilizó un spline de orden 4.

El enfoque multivariado, como MissForest, aprovechan la información de múltiples estaciones para imputar valores faltantes, capturando relaciones complejas y no lineales entre las

variables a través de técnicas de *random forest*. Esta capacidad de modelar patrones complejos los hace especialmente adecuados para datos hidrológicos, donde la interdependencia entre estaciones puede presentarse. Por otro lado, los métodos univariados, como los basados en regresión lineal y media móvil, se limitan a analizar la historia de cada serie de tiempo individual, utilizando técnicas de suavizado y modelado para estimar los valores faltantes. Aunque más simples, estos métodos pueden ser efectivos cuando la información de otras estaciones no está disponible.

6.6. Evaluación de la Eficiencia de los Métodos de Imputación

Los resultados se evaluaron utilizando métricas de regresión como MSE, R2, similitud, MAPE y correlación de Spearman, así como métricas de dispersión como CV_RATIO e IQR_RATIO. Estas métricas permitieron evaluar la eficiencia de los métodos desde diferentes perspectivas.

6.7. Análisis de Resultados

Para evaluar los métodos de imputación, se llevaron a cabo dos tipos de análisis: visual y estadístico. A través de gráficas comparativas, se evaluó cualitativamente la capacidad de los métodos en capturar tendencias, estacionalidad y eventos extremos. Complementariamente, se realizó un análisis cuantitativo de las métricas obtenidas para cada método y estación, utilizando boxplots para visualizar la distribución de los errores y comprender el comportamiento general de cada método.

7. Análisis de resultados

7.1. Caso de estudio.

Los datos de limnígrafos que se utilizaron en este documento fueron extraídos del programa de PIRAGUA, el cual está constituido por la red de monitoreo del recurso hídrico más grande del país, es un programa de gestión socioambiental que desde el año 2011 crea con las comunidades una cultura de la información a través de redes sociales de monitoreo; promueve y desarrolla sistemas de información del agua, construidos, implementados y operados por las comunidades de los 80 municipios de la jurisdicción de Corantioquia.

Las estaciones seleccionadas correspondieron a el río Porce (1004) cuyo nivel subsiguiente (NSS) es el río Guadalupe y Medio Porce, se encuentra ubicada en la territorial Tahamíes en el municipio de Gómez Plata. La estación del río Cacerí (1005) tiene como NSS el mismo río Cacerí y se encuentra ubicada en la territorial Panzenú en el municipio de Caucasia. El río Nus (1013) fue otra de las estaciones utilizadas, y se encuentra ubicada en la territorial Zenufaná en el municipio de Cisneros. La estación del río Tarazá (1014), está ubicada en la territorial Panzenú en el municipio de Tarazá, su NSS esta conformado por el río Tarazá y otros directos al río Cauca. La única quebrada que se seleccionó en este estudio fue la quebrada Doña María (1015), la cual se encuentra ubicada en el municipio de Itagüí, su NNS es el río Aburrá y pertenece a la territorial Aburra Sur. Otra de las estaciones seleccionadas fue el río Medellín (1017), a la altura del municipio de Copacabana, perteneciendo a la territorial de Aburrá Norte, su NSS corresponde al río Aburrá como en el caso anterior. Asimismo, se seleccionó la estación del río Cauca (1019), cuyo NSS es el río Tarazá y otros directos al Cauca, y se encuentra localizada en la territorial Panzenú en el municipio de Cáceres. El río Bolívar (1021) fue otra de las estaciones seleccionadas, su NSS es el río San Juan y se encuentra ubicada en la territorial Citará en el municipio de Ciudad Bolívar. La última estación seleccionada fue el río Chico (1023), su NSS corresponde a río Grande y se ubica en la territorial Tahamíes en el municipio de Belmira.

Ilustración 2.
Mapa de la zona de estudio.

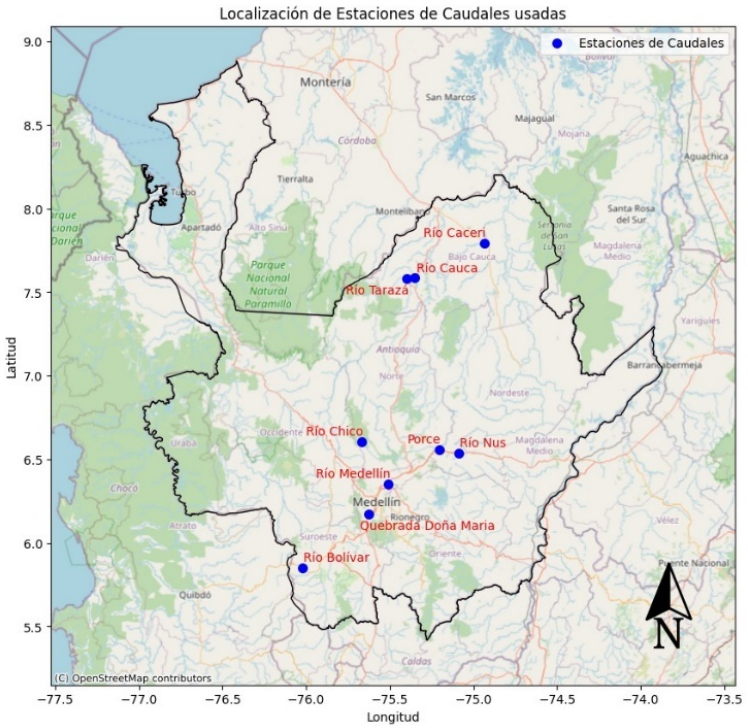


Ilustración 3.
Mapa de ubicación de las estaciones estudiadas.

7.2. Selección de estaciones y análisis exploratorio de las series de tiempo.

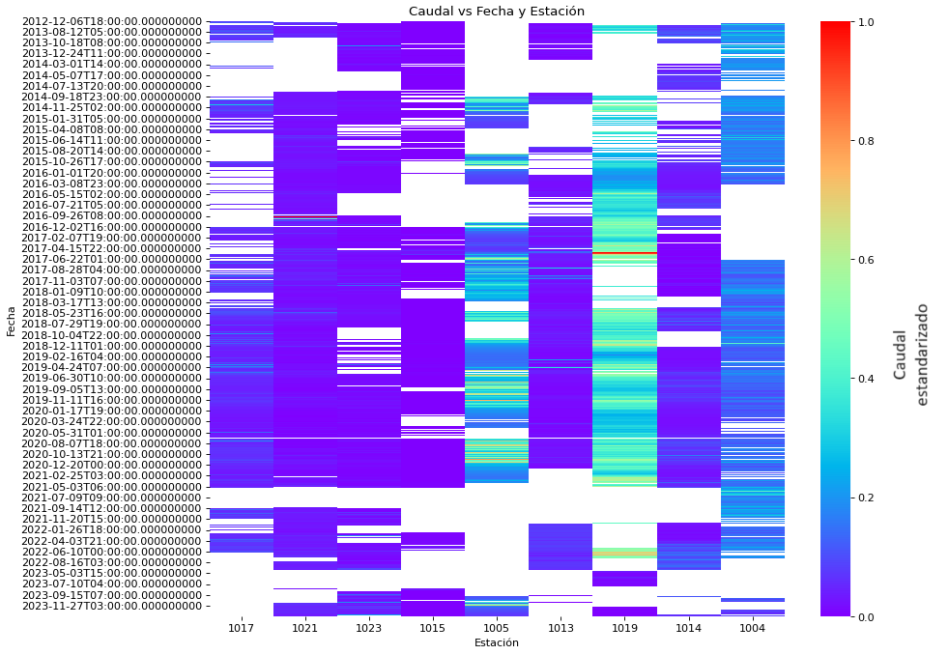
Una vez descargados los datos de las 45 series de tiempo horarias de caudales disponibles en el portal de PIRAGUA, se escogieron aquellas que contaran con una mayor cobertura temporal, seleccionando un total de nueve (9) estaciones. Entre estas estaciones escogidas, la que se encuentra ubicada en el río Medellin, fue la que tuvo una mayor cantidad de datos continuos con un total de 793 días de información horaria, para la fecha comprendida entre el 03/05/2018 y 05/07/2020. Respecto a la estación río Porce, esta fue la que contó con una menor cantidad de datos continuos, con un total de 270 días, para el periodo entre el 23/06/2017 y 21/03/2018. Es importante mencionar que los datos utilizados de estas estaciones no están comprendidos en el mismo periodo de tiempo, sin embargo, en su mayoría dichas estaciones contienen datos entre el año 2017 y 2020. Las nueve (9) estaciones de limnígrafos seleccionadas fueron las siguientes:

Tabla 2.
Fuentes hídricas seleccionadas para la aplicación de las técnicas de imputación de datos faltantes.

Identificación de la estación	Nombre de la fuente hídrica	Territorial	Coordenadas (latitud; longitud; altitud)	Fecha de inicio	Fecha de finalización	Cantidad de días sin valores faltantes
1017	Río Medellín	Aburrá Norte	6.350167; -75.507278; 1420	3/05/2018	5/07/2020	793
1021	Río Bolívar	Citará	5.849889; -76.023947; 1192	11/09/2018	5/07/2020	662
1023	Río Chico	Tahamíes	6.607472; -75.667720; 2533	27/04/2017	18/08/2018	478
1015	Quebrada Doña Maria	Aburra Sur	6.171831; -75.627628; 1610	10/05/2018	22/08/2019	469
1005	Río Cacerí	Panzenú	7.792736; -74.93405; 58	20/02/2017	11/03/2018	383
1013	Río Nus	Zenufaná	6.537571; -75.089490;	26/11/2018	25/11/2019	364

			1038			
1019	Río Cauca	Panzenú	7.584083; -75.348667; 81	26/06/2019	17/06/2020	356
1014	Río Tarazá	Panzenú	7.582581; -75.399639; 96	26/07/2019	5/07/2020	344
1004	Río Porce	Tahamíes	6.559589; -75.205878; 1828	23/06/2017	21/03/2018	270

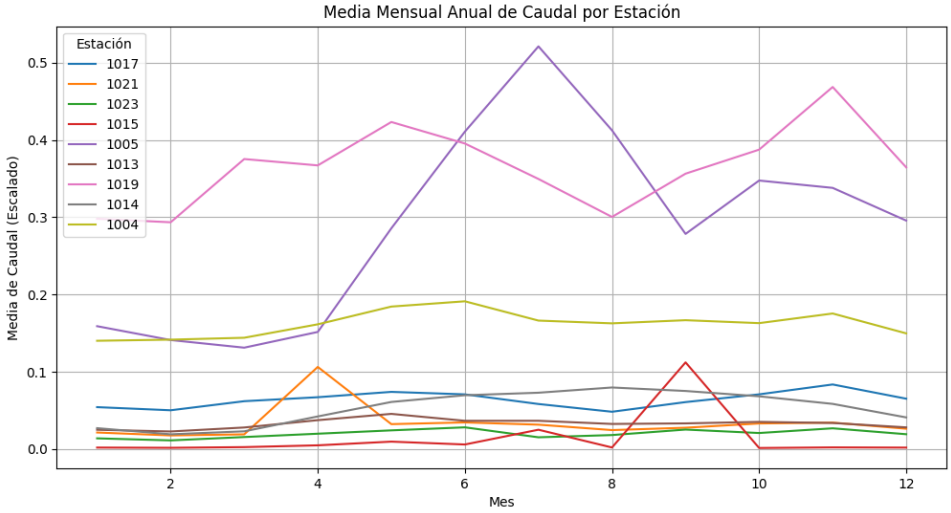
La Gráfica 2 presenta el caudal estandarizado mediante la técnica Min-Max Scaler y su relación con el tiempo, visualizado mediante un mapa de calor, lo cual permite visualizar de manera comparativa la disponibilidad de datos de las diferentes estaciones. Al escalar los datos entre 0 y 1, se resalta la posición de cada observación dentro del rango de valores de cada serie, independientemente de la magnitud absoluta del caudal. Este grafico evidencia la existencia de períodos críticos de escasez de información, especialmente antes de 2016 y después de 2021. Estas lagunas en los registros pueden comprometer la calidad y la representatividad de los análisis hidrológicos y limitar la comprensión de la dinámica hídrica de la región.



Gráfica 2

Mapa de calor de la distribución temporal de los datos de los limnígrafos.

La Gráfica 3, que muestra la media mensual multianual del caudal estandarizado, complementa el análisis al proporcionar una visión general de la variabilidad estacional de las diferentes estacionesen término de sus máximos y mínimos. Esta gráfica es útil para identificar patrones estacionales y detectar posibles anomalías en el comportamiento hidrológico.



Gráfica 3
Media mensual del caudal escalado por estación.

Con el objetivo de caracterizar de manera inicial las series de tiempo de caudal y obtener una primera aproximación a su distribución, se calcularon estadísticos descriptivos como la media, la desviación estándar y los percentiles 25%, 50% y 75%. Estos estadísticos proporcionan información valiosa sobre la tendencia central, la dispersión y la asimetría de los datos. Esta información se muestra en la Tabla 3.

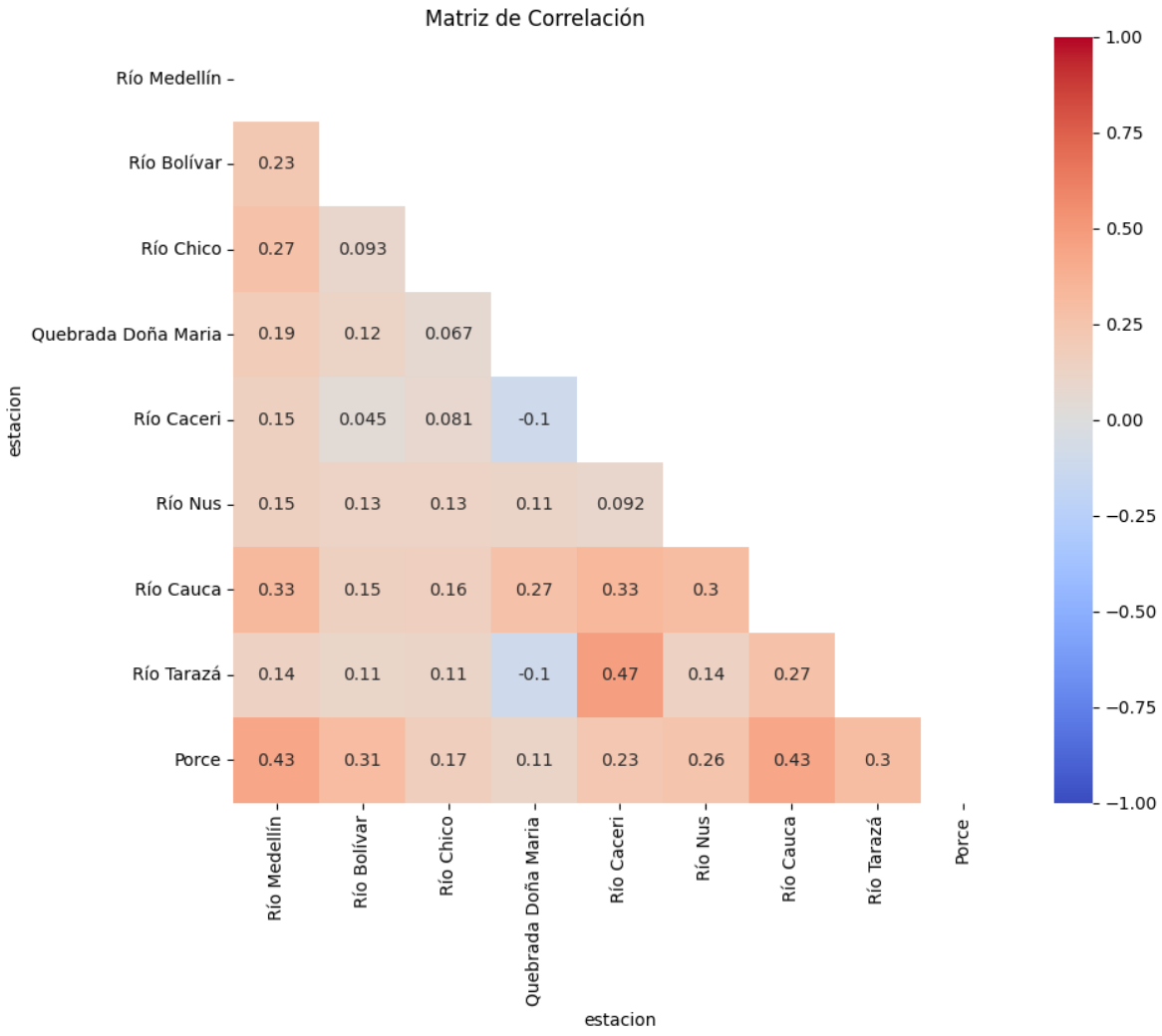
Tabla 3.
Estadísticos de las fuentes hídricas.

Identificación de la estación	Nombre de la fuente hídrica	Territorial	Media del caudal (m^3/s)	Desviación estándar (m^3/s)	Percentil 25%	Percentil 50%	Percentil 75%
1017	Río Medellín	Aburrá Norte	27.07	17.28	18.07	23.36	29.81
1021	Río Bolívar	Citará	1.76	1.58	1.21	1.62	2.09

1023	Río Chico	Tahamíes	1.47	2.22	0.81	1.03	1.39
1015	Quebrada Doña Maria	Aburra Sur	1.05	3.90	0.03	0.34	1.05
1005	Río Cacerí	Panzenú	170.82	105.39	98.10	142.05	213.09
1013	Río Nus	Zenufaná	0.40	0.39	0.18	0.29	0.51
1019	Río Cauca	Panzenú	3604.08	1415.81	2907.70	3579.15	4440.92
1014	Río Tarazá	Panzenú	167.39	136.09	71.88	151.79	229.61
1004	Río Porce	Tahamíes	61.48	18.38	50.08	59.29	70.25

La anterior información, permite contrastar que las mayores magnitudes de las fuentes hídricas seleccionadas son: la estación 1005 – Río Cacerí y 1019 - Río Cauca que registran un caudal medio de 170.82 m^3/s y 3604.08 m^3/s, respectivamente.

El análisis de correlación de Pearson (Gráfica 4) entre las diferentes estaciones reveló una fuerte asociación entre caudales en estaciones geográficamente cercanas y con regímenes hidrológicos similares. Destacan los altos coeficientes de correlación entre el río Cacerí y el río Taraza (0.47), así como entre el río Cauca y los ríos Porce y Medellín (0.43), lo cual es esperable dada su proximidad geográfica y conexiones hidrológicas. En contraste, la quebrada Doña María mostró una correlación significativamente menor con las demás estaciones, sugiriendo una dinámica hidrológica más independiente



Gráfica 4.

Coeficiente de correlación de los caudales de las diferentes estaciones.

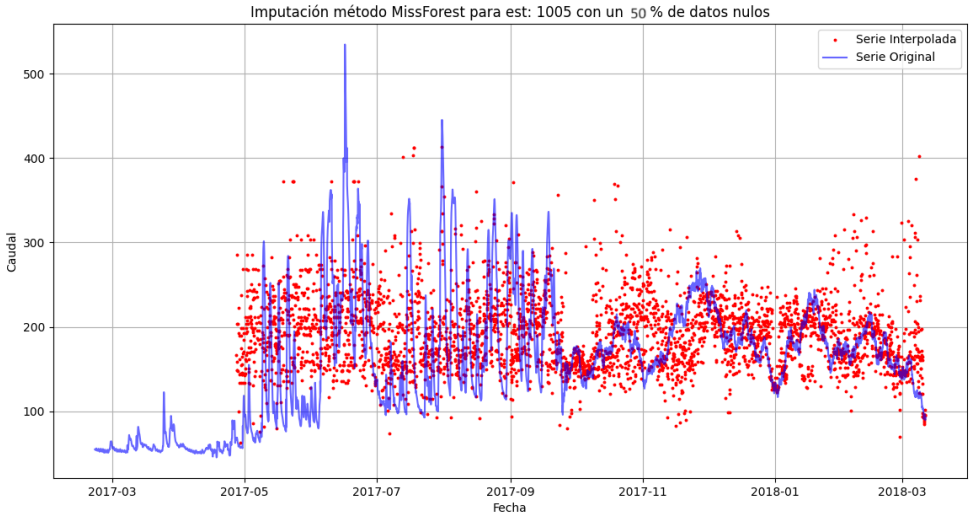
7.3. Análisis cualitativo de los resultados de las técnicas de imputación aplicadas.

Luego de seleccionar las estaciones con mayor cobertura temporal, realizar el análisis exploratorio de los datos y preparar los diferentes escenarios para la imputación de datos faltantes, como se especifica en el numeral 6.4 y 6.5, se implementaron las cinco técnicas de imputación de datos propuestas en este documento. Para ilustrar de manera cualitativa el desempeño de los diferentes métodos de imputación, se seleccionó un conjunto de series de tiempo con un 50% de datos faltantes, representativo del peor escenario analizado. A través de gráficas comparativas, se

visualizó el comportamiento de cada método en la reconstrucción de las series originales. Este análisis visual permitió identificar de forma general las fortalezas y debilidades de cada técnica de imputación

7.3.1. MissForest.

En la Gráfica 5, se observa el comportamiento de la imputación de datos con MissForest de la estación del río Cacerí (1005) para un escenario de datos faltantes del 50%. Al observar los valores imputados se revela un comportamiento errático de este método, ya que exhibe una alta variabilidad en los valores imputados, con una tendencia a dispersarse alrededor de la media y sin una adecuada representación de los valores extremos. Asimismo, en algunas ocasiones este método imputa de manera frecuente valores máximos o mínimos cuando la serie de tiempo original se comporta de manera opuesta.



Gráfica 5.
Imputación MissForest para la estación 1005 con un total de 50% de datos faltantes.

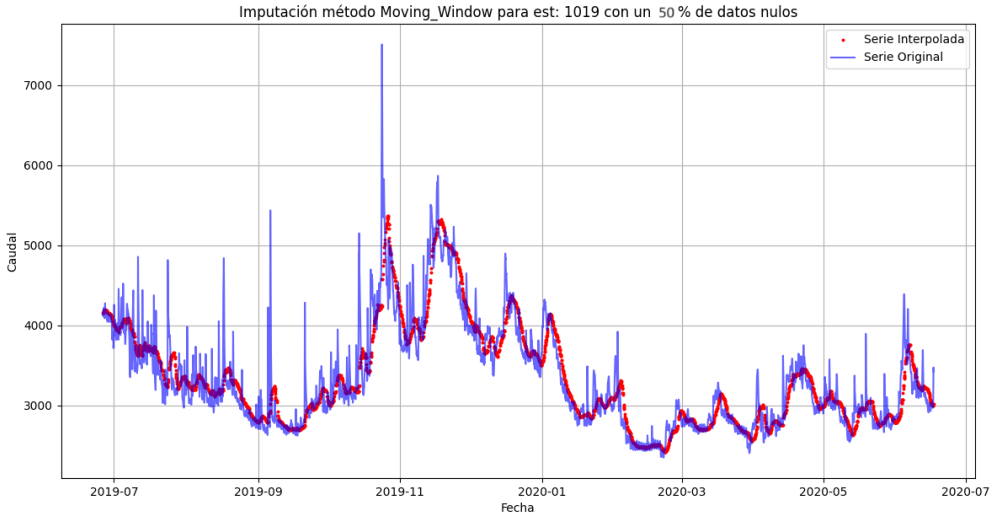
Es importante destacar que el método MissForest no pudo imputar los datos faltantes de los primeros meses de la serie del río Cacerí (1005). Esto se debe a la naturaleza multivariada del algoritmo, el cual requiere información de otras estaciones para realizar la imputación. Para este caso, ninguna de las otras estaciones cuenta con registros históricos de caudal tan recientes como

los de la estación 1005, lo que limita la disponibilidad de datos para alimentar el modelo en ese período específico.

7.3.2. Media móvil.

El método de media móvil o ventana deslizante se aplicó para una ventana temporal de 72 horas, cuya decisión se fundamenta en la necesidad de cubrir la laguna más extensa simulada en los datos, verificada de 70 horas. Una ventana de menor tamaño habría sido insuficiente para rellenar estos vacíos de información de manera efectiva conllevando a errores en la imputación de los datos.

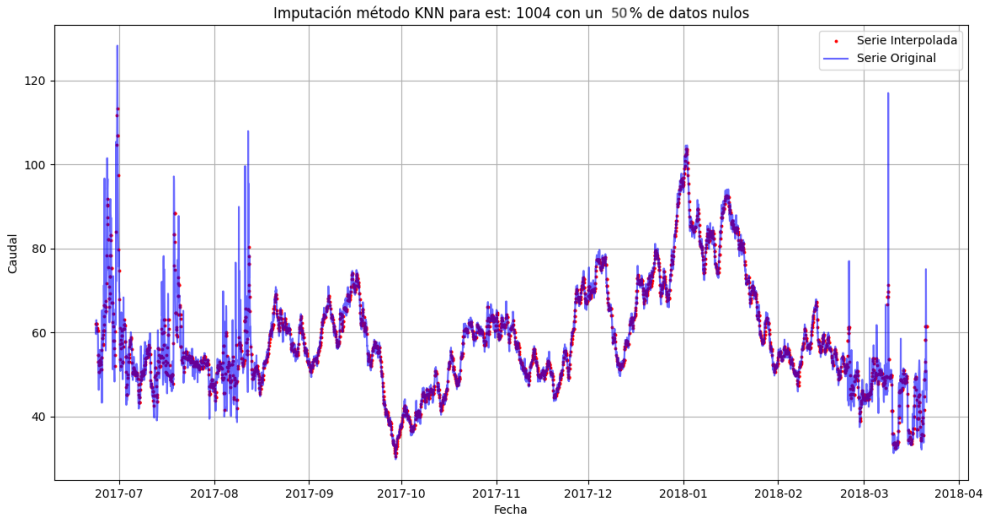
El método de media móvil, a pesar de su capacidad para capturar la tendencia general de los caudales, presenta algunas limitaciones. Su naturaleza suavizante puede subestimar los valores máximos y mínimos, lo que resulta en una representación menos precisa de los eventos extremos. Además, la ventana temporal de 72 horas utilizada para la imputación provoca un desplazamiento hacia la derecha de los valores estimados, como se observa en la Gráfica 6. A pesar de esta limitación, el método de media móvil proporciona una estimación razonable de la tendencia central de los datos.



Gráfica 6.
Imputación con media móvil para la estación 1019 con un total de 50% de datos faltantes.

7.3.3. KNN

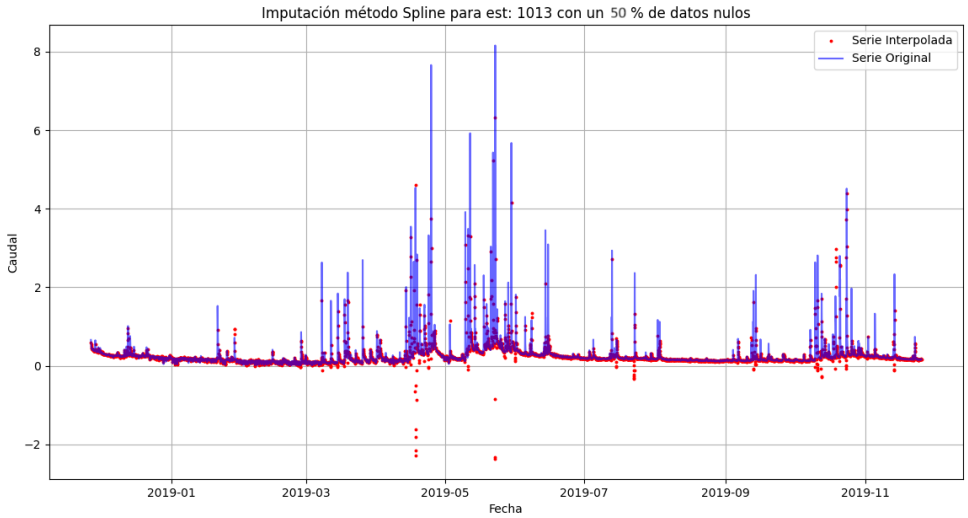
La Gráfica 7 revela que el método utilizado, aunque captura de forma adecuada la tendencia general de la serie temporal, presenta ciertas limitaciones. Si bien la imputación de los datos faltantes se ajusta razonablemente a la serie original, se observa una subestimación de los valores máximos de caudal. Además, es importante destacar que este método conlleva un costo computacional significativo debido a la complejidad de los cálculos involucrados, lo que puede limitar su aplicabilidad en grandes conjuntos de datos.



Gráfica 7.
Imputación KNN para la estación 1004 con un total de 50% de datos faltantes.

7.3.4. Spline

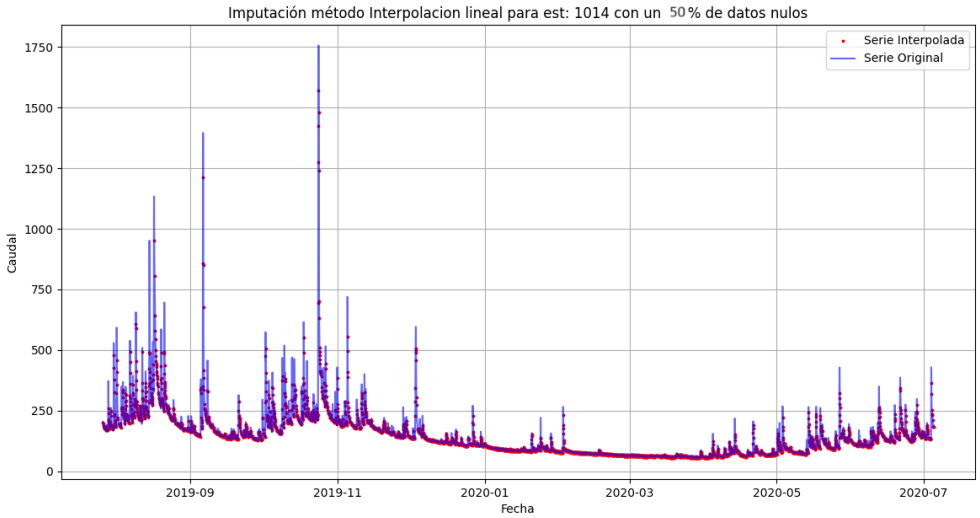
Si bien esta técnica demostró ser efectiva en la imputación de valores máximos y mínimos, presenta una limitación significativa: la generación de valores de caudal negativos en ocasiones que se relaciona directamente con la presencia de cambios bruscos en la serie de tiempo. Al intentar modelar estos cambios utilizando funciones polinómicas, el algoritmo puede extrapolar de manera inadecuada y generar valores no físicos. Esta limitación reduce la confiabilidad de los resultados obtenidos.



Gráfica 8.
Imputación Spline para la estación 1013 con un total de 50% de datos faltantes.

7.3.5. Interpolación lineal.

La evaluación gráfica de los resultados indica que este método de imputación supera a los métodos anteriores en términos de precisión. Al capturar de manera efectiva la tendencia, los valores extremos y hasta los atípicos de la serie, demuestra una buena capacidad para estimar los datos faltantes. La elección de la interpolación lineal resulta particularmente adecuada para series de caudal, caracterizadas por cambios graduales, ya que este método asume una variación lineal entre los puntos de datos, lo cual concuerda con la naturaleza de este tipo de series. Sin embargo, es importante destacar que este método puede no ser adecuado para series con patrones más complejos, como aquellos que presentan cambios bruscos o estacionalidad marcada.



Gráfica 9.
Imputación por Interpolación lineal para la estación 1014 con un total de 50% de datos faltantes.

7.4. Análisis cuantitativo de los resultados de las técnicas de imputación aplicadas.

Se llevó a cabo un análisis comparativo de diversos métodos de imputación, empleando un conjunto de métricas diseñadas para evaluar tanto la precisión cuantitativa (MSE, R^2, similitud) como la calidad cualitativa (IQR, CV y correlación de Spearman) de las imputaciones. Los resultados obtenidos revelan que la elección del método de imputación tiene un impacto significativo en el desempeño, especialmente en escenarios con altos porcentajes de datos faltantes. Si bien todos los métodos experimentaron una degradación en su desempeño a medida que aumentaba la cantidad de datos faltantes, algunos mostraron una mayor robustez que otros, destacando la importancia de seleccionar el método más adecuado en función de las características específicas del conjunto de datos y del problema en cuestión.

Los resultados, visualizados mediante boxplots, evidenciaron un desempeño superior de los métodos de interpolación lineal y spline. Estos métodos mostraron una mayor precisión en las imputaciones, reflejada en medianas de error más bajas y rangos intercuartílicos más estrechos. La interpolación lineal, al asumir una relación lineal entre los puntos de datos, resultó particularmente eficiente en escenarios donde esta suposición era válida. Por su parte, la interpolación spline, al

ajustar funciones suaves a los datos, demostró ser flexible para capturar patrones más complejos. No obstante, se observó que el método spline, aunque preciso, puede generar valores imputados negativos en algunas ocasiones, lo que limita su aplicabilidad en este contexto. Esto puede deberse a que los polinomios de orden 4 pueden tomar cualquier valor, incluyendo valores negativos; si los datos de entrada tienen alguna tendencia o patrón que sugiera una curvatura pronunciada, el spline puede intentar ajustarse a esta curvatura de manera tan precisa que exceda los límites de los datos originales, produciendo valores negativos en las zonas de interpolación.

El método KNN, si bien ofrece buenos resultados, presenta un mayor costo computacional debido a su naturaleza basada en distancias vectoriales y aprendizaje automático. Por otro lado, MissForest, a pesar de su robustez y capacidad para manejar múltiples variables, mostró un desempeño inferior, especialmente cuando el porcentaje de datos faltantes era del 50%. Esto sugiere que la complejidad inherente a este método puede no ser necesaria para los datos hidrológicos y, además, la disponibilidad de caudales de otras estaciones no mejoró significativamente los resultados. Así mismo, al evaluar las métricas IQR y CV, se observa que hay una cierta tendencia en subestimar los valores del caudal en estos dos métodos.

Los resultados obtenidos indican que los métodos univariados, especialmente la interpolación lineal, superaron a los métodos multivariados y al KNN en la tarea de imputación de datos faltantes en caudales. La interpolación lineal demostró ser la opción más adecuada, combinando precisión y eficiencia computacional. Si bien el método KNN y MissForest son robustos en diversas situaciones, en este caso específico, caracterizado por series temporales de caudales con un tamaño de muestra relativamente pequeño y una baja correlación entre las variables, no lograron capturar patrones complejos en los datos. Con un conjunto de datos más amplio y variables más relacionadas entre sí (como variables meteorológicas), es probable que estos modelos capturen patrones más complejos y ofrezcan predicciones más precisas.

La Tabla 4 muestra el resumen de las métricas promedio (en los tres escenarios), que fueron utilizadas para comparar los diferentes métodos de imputación de datos. De manera particular, se evidenció que para la estación del río Cauca (1019), se presentaron las mejores métricas para la mayoría de todas las técnicas de imputación aplicadas.

Tabla 4.

Promedio de las métricas obtenidas para los tres escenarios en las técnicas de imputación de datos.

Estación	Método	MSE	R^2	Similitud	CV	IQR	Spearman
1004	Interpolacion lineal	2.76	0.98	1.00	1.00	1.01	0.99
	KNN	5.48	0.97	0.99	1.02	1.03	0.98
	MissForest	73.03	0.58	0.96	1.17	1.16	0.75
	Moving_Window	15.74	0.91	0.98	1.03	1.06	0.95
	Spline	16.03	0.91	1.00	0.97	1.00	0.99
1005	Interpolacion lineal	3.35	1.00	1.00	1.00	1.00	1.00
	KNN	25.46	1.00	0.99	1.01	1.01	1.00
	MissForest	2116.70	0.58	0.96	1.12	1.12	0.76
	Moving_Window	755.65	0.86	0.98	1.05	1.02	0.94
	Spline	4.18	1.00	1.00	1.00	1.00	1.00
1013	Interpolacion lineal	0.01	0.88	1.00	1.04	1.00	0.99
	KNN	0.04	0.61	0.99	1.28	0.95	0.96
	MissForest	0.06	0.44	0.99	1.18	0.95	0.73
	Moving_Window	0.04	0.68	1.00	1.21	0.97	0.92
	Spline	0.02	0.82	1.00	0.97	1.00	0.98
1014	Interpolacion lineal	113.78	0.99	1.00	1.01	1.00	1.00
	KNN	744.34	0.93	0.99	1.07	0.99	0.99
	MissForest	1389.24	0.86	0.99	1.10	0.98	0.92
	Moving_Window	1853.50	0.81	0.99	1.08	0.99	0.98
	Spline	79.32	0.99	1.00	1.00	1.00	1.00
1015	Interpolacion lineal	1.07	0.64	1.00	1.26	0.99	0.99
	KNN	1.78	0.40	0.99	1.65	0.95	0.97
	MissForest	1.82	0.39	0.99	1.97	0.96	0.68
	Moving_Window	1.21	0.59	1.00	1.38	0.99	0.96
	Spline	1.33	0.55	1.00	1.06	0.99	0.99
1017	Interpolacion lineal	30.23	0.89	1.00	1.06	1.00	0.99
	KNN	92.85	0.66	0.98	1.25	0.97	0.95
	MissForest	84.22	0.69	0.99	1.21	1.11	0.88
	Moving_Window	84.08	0.69	0.99	1.21	0.99	0.91
	Spline	65.82	0.76	1.00	0.96	0.99	0.98
1019	Interpolacion lineal	648.25	1.00	1.00	1.00	1.00	1.00
	KNN	4911.11	0.99	0.99	1.01	1.00	1.00
	MissForest	44790.08	0.89	0.98	1.05	1.08	0.93
	Moving_Window	26854.55	0.94	0.99	1.02	1.00	0.97
	Spline	523.70	1.00	1.00	1.00	1.00	1.00
1021	Interpolacion lineal	0.00	0.98	1.00	1.00	1.00	1.00
	KNN	0.02	0.88	0.99	1.06	1.00	0.98
	MissForest	0.10	0.51	0.99	1.11	1.09	0.81
	Moving_Window	0.02	0.90	1.00	1.05	1.01	0.96
	Spline	0.01	0.95	1.00	0.98	1.00	0.99
1023	Interpolacion lineal	0.10	0.95	1.00	1.02	1.01	0.99
	KNN	0.75	0.65	0.99	1.18	0.99	0.97
	MissForest	1.35	0.35	0.99	1.17	0.42	0.55
	Moving_Window	0.52	0.75	1.00	1.12	0.92	0.92
	Spline	0.10	0.95	1.00	0.99	0.99	0.99

La Tabla 4 presenta un promedio de las métricas obtenidas para los métodos de imputación evaluados. En cuanto al error cuadrático medio (MSE), indicador de la precisión de las predicciones, los métodos MissForest y, en algunos casos, KNN y Ventana Móvil, mostraron los valores más elevados. Esto sugiere que estas técnicas tienden a subestimar o sobreestimar los valores faltantes en mayor medida.

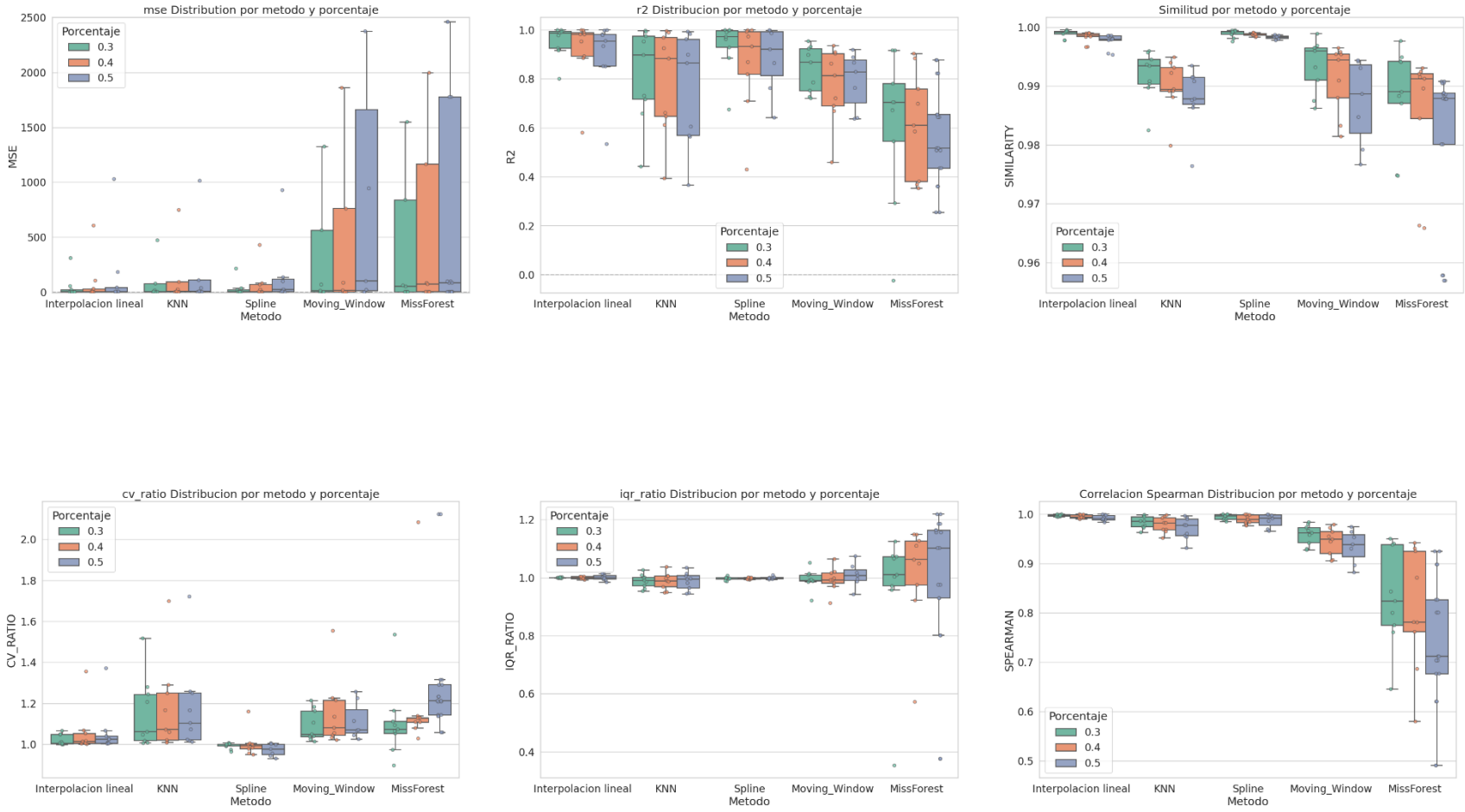
Por otro lado, el coeficiente de correlación de spearman reveló una alta correlación entre las series imputadas y originales al utilizar interpolación lineal y Spline, llegando incluso a valores de 1 en algunas estaciones. En contraste, MissForest presentó los coeficientes de correlación más

bajos, especialmente en las estaciones quebrada Doña María (1015) y río Chico (1023), indicando una menor correspondencia entre los datos originales y los imputados.

La similitud entre las series también fue evaluada, y los resultados mostraron valores cercanos a 1 para la mayoría de los métodos y estaciones, indicando una buena correspondencia general. Sin embargo, MissForest presentó los valores más bajos en las estaciones río Porce (1004) y Río Cacerí (1005), lo que sugiere una menor similitud entre las series originales y las imputadas por este método.

El coeficiente de variación (CV), que mide la variabilidad de los datos, mostró valores más altos para KNN y MissForest en la estación quebrada Doña María (1015), indicando una mayor dispersión de los datos imputados en comparación con la serie original. Ahora bien, con respecto al rango intercuartílico (IQR), que mide la dispersión central de los datos, reveló que MissForest tendió a subestimar o sobreestimar la dispersión de los datos en varias estaciones. Por ejemplo, en la estación río Chico (1023), el IQR fue menor para MissForest, indicando una menor dispersión en los datos imputados.

En resumen, los resultados obtenidos indican que la interpolación lineal y Spline son métodos más precisos y confiables para la imputación de datos faltantes en series de caudales. Estos métodos preservan mejor las características estadísticas de los datos originales, como la media, la varianza y la correlación. Por otro lado, MissForest y KNN mostraron un desempeño menos consistente, especialmente en términos de precisión y capacidad para capturar la variabilidad de los datos.



Gráfica 10.

Diagramas boxplot de las métricas calculadas para comprar los métodos de imputación.

8. Conclusiones

- La interpolación lineal se destacó como el método más preciso entre los evaluados. El spline, aunque capaz de modelar tendencias a largo plazo, mostró limitaciones al representar picos y cambios abruptos en los caudales horarios. La media móvil, por su parte, suavizó efectivamente la serie, pero resultó menos adecuada para capturar variaciones bruscas.
- Si bien MissForest presentó un desempeño inferior en nuestro análisis, es importante destacar que la efectividad de este método está estrechamente ligada a la calidad y cantidad de los datos de entrenamiento. Al incluir información de estaciones altamente correlacionadas y variables explicativas relevantes, como variables meteorológicas, se puede mejorar significativamente su desempeño. Además, MissForest puede ser especialmente útil para imputar brechas de datos más extensas, como las de semanas o meses.
- La evaluación del algoritmo KNN reveló que, aunque no fue el más adecuado para nuestro caso específico, podría ser una opción interesante en otras situaciones. Sin embargo, su desempeño se ve afectado por factores como la dimensionalidad de los datos y la cantidad de datos faltantes. En nuestro análisis, la falta de un entrenamiento multivariado limitó su capacidad para capturar patrones complejos en los datos. No obstante, en escenarios con estaciones altamente correlacionadas y variables explicativas relevantes, KNN podría ofrecer una alternativa viable a otros métodos.
- La selección del método de imputación para series de tiempo de caudal debe ser personalizada según las características específicas de los datos y las limitaciones del estudio. En nuestro caso, la interpolación lineal demostró ser la técnica más adecuada debido a la naturaleza generalmente lineal de las variaciones en los caudales y su baja demanda computacional. Sin embargo, para series con patrones más complejos o gaps de datos más grandes, métodos como MissForest o splines podrían ser más apropiados. La elección final debe considerar un equilibrio entre precisión, eficiencia computacional y la capacidad de capturar las características particulares de la serie.

- Los resultados obtenidos sugieren que la robustez de los métodos de imputación ante diferentes tasas de datos faltantes puede variar considerablemente. Mientras que la interpolación lineal y spline mostraron un desempeño relativamente estable, métodos como KNN, media móvil y MissForest presentaron una mayor sensibilidad al incremento en el porcentaje de datos faltantes.

9. Investigación futura

Se recomienda realizar investigaciones adicionales para evaluar la efectividad de diferentes métodos de imputación en escenarios con *gaps* que contengan periodos de semanas y meses. Además, es importante explorar la integración de variables relacionadas geográficamente y temporalmente para mejorar el rendimiento de los modelos de aprendizaje automático en la imputación de series temporales de caudal.

Una de las principales limitaciones de este estudio fue la dificultad en obtener datos de precipitación de calidad y continuidad. La falta de estaciones meteorológicas cercanas a las estaciones limnimétricas y la discontinuidad en los registros disponibles en la plataforma PIRAGUA dificultaron la inclusión de esta variable en los modelos de imputación multivariados. Como consecuencia, los modelos desarrollados no capturan la influencia de la precipitación en el régimen hidrológico, lo que podría afectar la precisión de las imputaciones, especialmente durante eventos extremos. Para futuras investigaciones, se recomienda explorar el uso de productos de precipitación satelital y técnicas de reconstrucción de series de tiempo para mejorar la calidad de los datos y obtener una representación más realista de las relaciones hidrológicas.

Inicialmente, se exploraron los modelos ARIMA/SARIMA debido a su amplia aplicación en el pronóstico de series de tiempo. Sin embargo, estos modelos están diseñados principalmente para pronósticos y no están optimizados para la imputación de datos faltantes en puntos específicos de la serie. Además, se consideró entrenar un modelo ARIMA o SARIMA para cada segmento de la serie temporal disponible. No obstante, este enfoque resultaba extremadamente costoso desde el punto de vista computacional y complejo de programar. Por esta razón, se optó por explorar otras técnicas más adecuadas para nuestra aplicación. Sin embargo, valdría la pena explorar métodos alternativos para abordar este tipo de problemas mediante el uso de estos algoritmos, ya que estos modelos pueden ofrecer resultados precisos si se cumplen con los supuestos de estacionariedad en la serie temporal y con suficiente información histórica.

10. Referencias

Arathy Nair, G. R., Adarsh, S., El-Shafie, A., & Ahmed, A. N. (2024a). Enhancing hydrological data completeness: A performance evaluation of various machine learning techniques using probabilistic fusion imputer with neural networks for streamflow data reconstruction. Journal of Hydrology. https://doi.org/10.1016/j.jhydrol.2024.131583

Arathy Nair, G. R., Adarsh, S., El-Shafie, A., & Ahmed, A. N. (2024b). Enhancing hydrological data completeness: A performance evaluation of various machine learning techniques using probabilistic fusion imputer with neural networks for streamflow data reconstruction. Journal of Hydrology, 639. https://doi.org/10.1016/j.jhydrol.2024.131583

Bello, A. M., Andrés Cuta, J., & García, E. K. (2019). TÉCNICAS DE IMPUTACIÓN PARA DATOS DE PRECIPITACIÓN MÁXIMA MENSUAL EN LA ZONA CENTRAL DE BOYACÁ Imputation techniques applied in a maximum monthly precipitation data in the central zone of Boyacá. In Rev. Revista Ingeniería, Investigación y Desarrollo (Vol. 19, Issue 1).

Chica Jimenez, J. A. (2018). INTERPOLACIÓN SPLINE Y APLICACIÓN A LAS CURVAS DE NIVEL.

Freire Míguez, L. (2022). Tratamiento de falta de información en técnicas de minería de datos.

García Reinoso, P. L. (2015). Imputación de Datos en Series de Precipitación Diaria Caso de Estudio Cuenca del Río Quindío. Ingeniare, 18, 73–86. https://doi.org/10.18041/1909-2458/ingeniare.18.539

Hamzah, F. B., Hamzah, F. M., Razali, S. F. M., & Samad, H. (2021). A comparison of multiple imputation methods for recovering missing data in hydrological studies. Civil Engineering Journal (Iran), 7(9), 1608–1619. https://doi.org/10.28991/cej-2021-03091747

Hamzah, F. B., Mohd Hamzah, F., Mohd Razali, S. F., Jaafar, O., & Abdul Jamil, N. (2020). Imputation methods for recovering streamflow observation: A methodological review. Cogent Environmental Science, 6(1). https://doi.org/10.1080/23311843.2020.1745133

Jing, X., Luo, J., Wang, J., Zuo, G., & Wei, N. (2022). A Multi-imputation Method to Deal With Hydro-Meteorological Missing Values by Integrating Chain Equations and Random Forest. Water Resources Management, 36(4), 1159–1173. https://doi.org/10.1007/s11269-021-03037-5

Kannegowda, N., Udayar Pillai, S., Kommireddi, C. V. N. K., & Fousiya. (2024). Comparative assessment of univariate and multivariate imputation models for varying lengths of missing rainfall data in a humid tropical region: a case study of Kozhikode, Kerala, India. Acta Geophysica, 72(4), 2663–2678. https://doi.org/10.1007/s11600-023-01152-y

Khampuengson, T., & Wang, W. (2023). Novel Methods for Imputing Missing Values in Water Level Monitoring Data. Water Resources Management, 37(2), 851–878. https://doi.org/10.1007/s11269-022-03408-6

Little, R., & An, H. (2004). ROBUST LIKELIHOOD-BASED ANALYSIS OF MULTIVARIATE DATA WITH MISSING VALUES. In Statistica Sinica (Vol. 14).

Londoño-Ciro, L. A., & Cañón-Barriga, J. E. (2015). Imputation of spatial air quality data using gis-spline and the index of agreement in sparse urban monitoring networks. Revista Facultad de Ingenieria, 2015(76), 73–81. https://doi.org/10.17533/udea.redin.n76a09

Rubin, D., & Little, R. (2002). Statistical Analysis with Missing Data (2 nda).

Ruelland, D., Ardoin-Bardin, S., Billen, G., & Servat, E. (2008). Sensitivity of a lumped and semi-distributed hydrological model to several methods of rainfall interpolation on a large basin in West Africa. Journal of Hydrology, 361(1–2), 96–117. https://doi.org/10.1016/j.jhydrol.2008.07.049

Sánchez, L. (2020). Estimación e imputación de datos faltantes mediante métodos de interpolación espacial para precipitación mensual acumulada en el departamento de Antioquia durante el periodo 2014-2018.

Sánchez Quiroga, L. (2020). Estimación e imputación de datos faltantes mediante métodos de interpolación espacial para precipitación mensual acumulada en el departamento de Antioquia durante el periodo 2014-2018.

Stekhoven, D. J., & Bühlmann, P. (2012). Missforest-Non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Thi-Thu-Hong, P. (2018). Comparative Study on Univariate Forecasting Methods for Meteorological Time Series.

Trejos Hernández, L. J., & Villada Lizarazo, J. I. (2019). Comparación de algoritmos de imputación para el parámetro de la precipitación de modelos hidrológicos empleando técnicas de ciencia de datos y Big Data.

Umar, N., & Gray, A. (2023). Comparing Single and Multiple Imputation Approaches for Missing Values in Univariate and Multivariate Water Level Data. Water (Switzerland), 15(8). https://doi.org/10.3390/w15081519

Urrutia, J. A., Palomino, R., Sc, M., Investigador, M. P., Darío, H., Sc, S. M., & Profesor, C. (2010). Metodología para la imputación de datos faltantes en meteorología. Scientia et Technica Año XVII, 46.

Vasker Sharma. (2021). Imputing Missing Data in Hydrology using Machine Learning Models. International Journal of Engineering Research And, V10(01). https://doi.org/10.17577/IJERTV10IS010011

Zeng, Z., Cui, L., Qian, M., Zhang, Z., & Wei, K. (2023). A survey on sliding window sketch for network measurement. In Computer Networks (Vol. 226). Elsevier B.V. https://doi.org/10.1016/j.comnet.2023.109696

Zhou, Y., Tang, Q., & Zhao, G. (2023). Gap infilling of daily streamflow data using a machine learning algorithm (MissForest) for impact assessment of human activities. Journal of Hydrology, 627. https://doi.org/10.1016/j.jhydrol.2023.130404