Imputación de datos faltantes en series climáticas para la precipitación en el departamento del Meta

Ronaldo Hernández Mejia
Héctor Freddy Velasco Zapata

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2024

Imputación de datos faltantes en series climáticas para la precipitación en el departamento del Meta

Ronaldo Hernández Mejia
Héctor Freddy Velasco Zapata

Trabajo de grado presentado como requisito parcial para optar al título de:
Estadístico

Director:
Ph.D. Alvaro José Flórez Poveda
Codirector:
M.Sc. David Arango

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2024

Las grandes obras no son llevadas a cabo
por la fuerza, sino por la perseverancia.

— Samuel Johnson

Para que pueda surgir lo posible
es preciso intentar una y otra vez lo imposible

— Hermann Hesse

Agradecimientos

En primer lugar, a Dios, por ayudarme a entender este camino y crecimiento personal, que me hizo entender que hay un tiempo para todo y no me dejo desvanecerme, cuando no había más fuerza. A mi madre Adiela Mejia Arango, por brindarme su apoyo y enseñarme a luchar por mis sueños, que con gran valentía, esfuerzo y amor construyó todo en mi vida para formarme y permitirme desarrollarme profesionalmente. A mi Padre Jorge Hernandez, por apoyarme para poder culminar con esta etapa tan importante en mi vida y llevarme siempre en sus oraciones. Agradezco a mis maestros como el Profesor Roberto Behar Gutiérrez y Javier Olaya Ochoa que, con su gran vocación, humildad y amor por la educación, marcaron mi vida haciéndome creer en mí mismo, gracias a la Universidad Del Valle por brindarme una educación profesional y de calidad. Gracias a Sara Salinas, quien fue parte fundamental en esta última etapa académica, que, con amor y admiración, me llevo a plantear nuevas metas de vida, así como su apoyo incondicional deseándome siempre lo mejor.

Ronaldo Hernandez Mejia

Quiero expresar mi más profundo agradecimiento a todos aquellos que contribuyeron significativamente a la realización de esta tesis. En primer lugar a Dios, como fuente de sabiduría y guía, por concederme fortaleza y claridad mental durante este arduo proceso de investigación. Su divina inspiración ha iluminado mi camino y fortalecido mi espíritu en cada paso de esta jornada académica. A mis padres Emilse y Luis, cuyo amor y apoyo incondicional han sido mi fuente de inspiración y fortaleza a lo largo de este arduo proceso académico. Su constante aliento y sacrificio han sido cruciales para mi camino hacia el éxito en este trabajo. Agradezco profundamente a mis respetados profesores por su valiosa orientación, conocimientos compartidos y sabios consejos, que contribuyeron significativamente con su dedicación y compromiso a mi progreso académico, han dejado una marca indeleble en mi formación profesional. Además, agradezco a la Universidad del Valle por ofrecerme una educación de primer nivel y un ambiente académico enriquecedor. El compromiso de la institución con la excelencia académica y el fomento de la investigación me ha motivado a buscar constantemente formas de contribuir con el conocimiento adquirido al avance de la sociedad. Quiero expresar mi gratitud en particular a mi director Alvaro Flórez, cuya orientación experta, respaldo continuo fueron esenciales en cada etapa de este proyecto. Su dedicación y liderazgo fueron esenciales tanto para el logro de los objetivos establecidos como para el desarrollo de esta tesis.

Héctor Freddy Velasco Zapata

Resumen

El trabajo de grado se centra en la imputación de datos faltantes para la variable de precipitación en el Departamento del Meta, Colombia. El objetivo de la investigación fue evaluar la eficacia de varios modelos de aprendizaje automático, incluidos KNN, XGBoost y Random Forest, en la imputación de datos faltantes. Durante la experimentación y evaluación de los modelos, se utilizaron métricas básicas como MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), R^2(Coeficiente de determinación) y MAPE (Mean Absolute Percentage Error). Estas métricas se utilizaron para medir y comparar la precisión y el rendimiento de cada modelo en el proceso de imputación de datos faltantes de precipitación. Además, se utilizaron métodos sofisticados para maximizar el rendimiento del proceso de imputación. Con el objetivo de mejorar la eficiencia computacional y acelerar el tiempo de ejecución de los modelos de imputación en grandes conjuntos de datos, se examinaron técnicas de paralelización de procesos utilizando bibliotecas como Dask y Joblib de Python.

Palabras clave: precipitación, estación climatológica, datos faltantes, imputación, aprendizaje automático, paralelización de procesos.

Contenido

Lista de Figuras

Lista de Tablas

1. Introducción

1.1. Planteamiento del Problema

La precipitación es una variable climática importante en la hidrología, siendo esencial para el cálculo de balances hídricos y la identificación de alertas tempranas por riesgo de inundación o sequía. Además, su medición desempeña un papel relevante en el diseño de sistemas de acueducto y alcantarillado. En Colombia, donde la economía está fuertemente ligada a actividades agrícolas, el cambio climático tiene un impacto directo en la ciudadanía. En este sentido, el Departamento del Meta, con un área de gran extensión y con una economía basada en minas, agricultura y ganadería, experimenta influencias tropicales y fenómenos Océano-Atmosféricos como El Niño y La Niña (Instituto de Hidrología 2023).

Las mediciones de la precipitación son tomadas por la red de estaciones meteorológicas del Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM), a través de instrumentos como pluviómetro y su medida en milímetros (mm), entre otros, buscan registrar y comprender el comportamiento de la precipitación (Hurtado et al. 2001). La disponibilidad de datos completos y precisos de precipitación es fundamental para comprender su distribución espacial y temporal. En la escala temporal, los registros de precipitación se estudian como series de tiempo, ya sea diarias, mensuales o anuales. En cuanto a la escala espacial, la disponibilidad de datos del IDEAM permite analizar la precipitación a diferentes niveles geográficos, facilitando un entendimiento detallado de su variabilidad espacial (Prácticas Hidrológicas 1994).

El manejo de datos faltantes en los registros de precipitación de las estaciones meteorológicas es un desafío. Las circunstancias adversas, como fallas en las redes de medición en tierra, daños a equipos o robos, dificultan el seguimiento adecuado de esta variable (Medina Rivera 2008). Adicionalmente, la ubicación no homogénea de las estaciones meteorológicas, determinada principalmente por criterios de acceso y seguridad, afecta la representatividad de los datos obtenidos. La distribución y acceso resulta en zonas con un gran número de estaciones, mientras que en otras hay escasez o ausencia total de estaciones de monitoreo (Ortega & Rodriguez 2011).

Particularmente para el Departamento del Meta, se cuenta con datos diarios de 57 estaciones meteorológicas del IDEAM desde 1980 hasta 2016. Estos datos incluyen registros de precipitación, fecha, coordenadas (latitud y longitud), y una identificación única para cada

estación. En este extenso periodo de tiempo, se ha registrado un total de 656.436 datos. Cabe mencionar que un 8,27 % (64.312 datos) están ausentes, lo cual genera una serie de datos con valores faltantes.

Se presenta un desafío significativo debido a la gran cantidad de estaciones meteorológicas y la considerable cantidad de datos registrados en el Departamento del Meta, también enfrentamos la realidad de que un porcentaje de esta información presenta ausencias. Es importante abordar de manera eficiente la imputación de estos datos faltantes, dado el volumen de información disponible. En este sentido, la paralelización emerge como una estrategia fundamental. Esta técnica nos permite optimizar los procesos de imputación al realizar múltiples tareas simultáneamente, agilizando el tiempo de computación (Sabeena & Reddy 2017).

En este estudio, se propone abordar la problemática inherente a los datos faltantes en las series de precipitación del Departamento del Meta. La complejidad de este desafío requiere la comparación y evaluación de diferentes metodologías de imputación, considerando tanto la precisión en la imputación como la eficiencia computacional. Con el objetivo de mejorar la calidad de las imputaciones, se emplearán técnicas de machine learning, las cuales son seleccionadas debido a la complejidad de los datos y su capacidad para manejar patrones complejos y no lineales. Específicamente, utilizaremos métodos como el k-Nearest-Neighbor (KNN), Extreme Gradient Boosting (XGBoost) y Random Forest, los cuales se destacan por su capacidad de aprender y predecir con mayor exactitud en comparación con técnicas tradicionales (Hastie et al. 2001, Chen & Guestrin 2016*a*, Breiman 2001).

Para optimizar los cálculos, se incorporará la paralelización de procesos utilizando paquetes de Python. Esto último tiene como propósito optimizar la imputación al distribuir y ejecutar simultáneamente tareas computacionales, acelerando significativamente el proceso. Por lo tanto, no solo contribuirá a mejorar la calidad de las bases de datos climáticas, proporcionando información valiosa para la toma de decisiones y la gestión efectiva de los recursos hídricos en el Departamento del Meta, sino que también evaluará la eficacia de la integración de técnicas de machine learning y paralelización de procesos en la imputación de datos faltantes.

La selección de los métodos a comparar se llevó a cabo mediante la revisión bibliográfica (Fontecha Bello et al. 2020, Sahoo & Dillip Kumar 2022, Morales España et al. 2008, Sandín Carral 2015) de varias técnicas propuestas para la estimación de datos de este tipo. En esta, los autores concuerdan que la aplicación de metodologías avanzadas y la consideración de la eficiencia computacional son aspectos fundamentales para contribuir a la mejora de la calidad de la información.

1.2. Justificación

La variable climática de precipitación desempeña un papel fundamental en la definición del estado atmosférico y su disponibilidad afecta directamente diversas actividades económicas esenciales para la nación. En Montealegre (2009) se considera que la variabilidad en la precipitación, incluyendo eventos extremos como inundaciones y sequías, impacta no solo la economía sino también el medio ambiente y la sociedad en general.

En el ámbito agrícola, la medición precisa de la precipitación es vital para planificar siembras y cosechas, ya que afecta la humedad del suelo. El estudio realizado por González et al. (2020) analiza la relación entre la precipitación, el área sembrada y la producción de arroz en la región. Los resultados indican que una mayor área sembrada y una adecuada pluviosidad durante el ciclo productivo se traducen en una mayor producción de arroz, lo que resalta la utilidad de este estudio para mejorar la planificación y gestión agrícola en la región.

Dado que la agricultura es la actividad principal para el desarrollo del país y se destaca por su alta tecnificación, las entidades gubernamentales, como el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM[1]), y grupos de investigación como GIREH[2] e IREHISA[3], tienen conocimiento de la importancia de la variabilidad en el clima, especialmente en lo que respecta a la precipitación. Sin embargo, su conocimiento se traduce en acciones concretas, además de la comprensión teórica. Por ejemplo, vigilan de cerca las fluctuaciones climáticas y realizan análisis de componentes espaciales y temporales. Estas acciones son de gran utilidad en la prevención y orientación, en casos eventuales de catástrofes naturales o sequías que son generadas por las variaciones de la precipitación (Hurtado et al. 2001).

Aunque se cuenta con una red de estaciones meteorológicas que miden variables climáticas como la precipitación, la información disponible está incompleta debido a la falta de toma de datos ocasionada por daños eléctricos, pérdida o destrucción de registros, y errores en el proceso de medición (Medina Rivera 2008). Estas irregularidades en las estaciones de monitoreo pueden llevar a resultados erróneos en la calidad de los datos.

Además de las fallas técnicas, las estaciones meteorológicas del Departamento del Meta están ubicadas en una amplia zona con una variedad de relieves y condiciones ambientales. Debido a que algunas áreas tienen una concentración de estaciones mayor que otras, esta distribución espacial afecta la precisión de las mediciones recogidas. Para realizar una correcta imputación de los datos faltantes, es fundamental considerar la distribución espacial de las estaciones y aprovechar el volumen de información que proporcionan las estaciones cercanas. Por lo tanto, es fundamental comparar varios métodos de estimación bajo estas condiciones, ya que

[1] http://www.ideam.gov.co
[2] https://sites.google.com/a/unal.edu.co/dica/investigacion/gireh?authuser=0
[3] https://irehisa.univalle.edu.co/index.html

la efectividad de cada método puede variar según la distribución espacial de las estaciones. (Bello 2019).

El empleo inadecuado de técnicas de imputación en contextos diversos puede llevar a errores inferenciales significativos, incluyendo sesgos, distorsiones en la estructura de covarianzas y una incorrecta estimación de la varianza. Esto puede resultar en datos poco fiables y, por ende, en interpretaciones erróneas de los resultados (Gómez et al. 2006, García 2015).

La calidad de los datos de precipitación es fundamental para la validez de los estudios académicos. La precisión y confiabilidad de estos datos son fundamentales, especialmente para estudios que utilizan análisis de series de tiempo para estimar la precipitación y comprender su comportamiento a lo largo del tiempo (Dumedah & Coulibaly 2011). Sin datos precisos y confiables, los resultados de estos análisis pueden ser incorrectos o engañosos. Por lo tanto, es imperativo abordar con cuidado el problema de los datos faltantes, ya que la falta de información confiable puede afectar negativamente la efectividad y validez de estos estudios, comprometiendo la capacidad de obtener resultados útiles y exactos.

1.3. Pregunta de investigación

¿Qué tan adecuada es la imputación de datos faltantes en la precipitación mediante las estaciones de monitoreo del IDEAM en el departamento del Meta?

1.4. Objetivos

1.4.1. Objetivo general

Proponer una metodología para la imputación de datos faltantes para la precipitación diaria a través de las estaciones de monitoreo del IDEAM en el departamento del Meta.

1.4.2. Objetivos específicos

- Implementar modelos de imputación de datos faltantes en las estaciones de monitoreo para medir la variable precipitación en el departamento del Meta.
- Comparar el desempeño estadístico y computacional de los modelos de imputación de datos faltantes.

2. Antecedentes

En esta sección se presenta los antecedentes relacionados con estudios realizados de la precipitación, la integración de información de diferentes estaciones climáticas y satelitales, se realizó una revisión bibliográfica. A continuación, se presentan algunas de las investigaciones que se han realizado para el estudio de la precipitación y otras variables meteorológicas.

Sánchez Quiroga (2020) en su estudio tiene el objetivo de contrastar métodos de interpolación espacial en la imputación de datos faltantes de precipitación mensual acumulada en el departamento de Antioquia durante el periodo 2014-2018, con datos suministrados por el IDEAM. Para ello, en este trabajo se utilizó la distancia inversa ponderada (IDW), Spline de placa delgada (TPS) y kriging ordinario, como métodos de imputación y la evaluación se hace a través de la raíz del error cuadrático medio (RMSE). El Kriging ordinario fue efectivo cuando se tiene más del 10 % de los datos faltantes, el método de la distancia inversa ponderada fue el que arrojó mejores resultados cuando se tienen 5 % de los valores faltantes.

Fontecha Bello et al. (2020) en su investigación, abordan el problema de los datos faltantes para la variable precipitación, a un conjunto de datos recopilados en la parte central del departamento de Boyacá - Colombia para el período 1974-2013. Ellos evalúan el desempeño de los mecanismos de imputación de datos faltantes, cada uno de estos se realizó usando una imputación múltiple con un enfoque aleatorio, una asignación por el método de K-vecinos más cercanos (KNN) con enfoque espacial y una imputación por el método de suavizado de Kalman con enfoque temporal. Luego miden la convergencia de los estadísticos descriptivos del valor imputado y el valor original y se realizó la comparación de los ajustes gráficos y sus distribuciones de probabilidad, sugiriendo un mejor ajuste usando la herramienta de imputación múltiple *Amelia* del software R, en conjunto con un ajuste a una distribución gamma para los datos faltantes en el conjunto de datos de referencia.

Reinoso (2015) presentan los resultados obtenidos por la aplicación de cinco técnicas de imputación de datos en series de precipitación diaria para ocho estaciones que tienen influencia sobre la cuenca del río Quindío, localizada en la zona central colombiana. Las técnicas elegidas fueron la distancia inversa ponderada, el coeficiente de correlación ponderado, el exponencial inverso ponderado, la medida estadística ponderada y el radio normal ponderado, como técnicas de interpolación para la imputación de datos faltantes. La aplicación de estos métodos requiere, además de la información de precipitación, la

localización geográfica de las estaciones y la distancia que existe entre ellas. Con el propósito de preservar la generación de valores de precipitación igual a cero, se consideró el cálculo de probabilidades empíricas a partir de una cadena de Markov de primer orden. Los datos imputados por la técnica de distancia estadística ponderada conservan adecuadamente las medidas de tendencia central de la serie temporal de precipitación diaria con datos faltantes.

Sahoo & Dillip Kumar (2022) realizan un estudio en la cuenca de Cachar, estado de Assam (India), para la imputación de datos faltantes para la variable precipitación, considerando diecinueve estaciones pluviométricas, y utilizan KNN, mapas auto organizados (SOM), bosque aleatorio (RF) y red neuronal de retroalimentación (FNN). Los autores utilizan varios índices de rendimiento como el RMSE, el coeficiente de determinación (R^2) y el error absoluto medio (MAE). Los índices de rendimiento del modelo indican que la técnica FNN, demostró su efectividad en la imputación de datos faltantes de precipitación en comparación con KNN, SOM y RF, especialmente en regiones con un número extremo de datos faltantes. Los resultados de esta investigación demostraron ser muy imprescindibles a la hora de seleccionar técnicas preeminentes para estimar datos de precipitación y reducir las lagunas de datos en cuencas complejas como Cachar. Estas lagunas de datos son períodos o áreas donde no hay datos sobre la precipitación en la cuenca de Cachar. Para todas las estaciones, los índices de rendimiento de los modelos propuestos se encontraban dentro del rango estándar de modelización hidrológica, que demuestra que los modelos están generando resultados que son cohesivos y aceptables de acuerdo con los criterios establecidos por la comunidad científica para este tipo de aplicaciones. Basándose en los indicadores de rendimiento del modelo, los autores consideran que estas técnicas pueden utilizarse para la gestión de recursos hídricos y la modelización hidrológica.

Yozgatligil et al. (2013) tienen como objetivo comparar diferentes técnicas de imputación para completar los valores ausentes en series temporales meteorológicas espacio temporales. Para lograr esto, se evalúan seis métodos de imputación en función de una serie de criterios que incluyen exactitud, solidez, precisión y eficiencia para datos faltantes creados artificialmente de series mensuales de precipitación total y temperatura media del Servicio Meteorológico Estatal de Turquía. Los métodos más simples como el promedio aritmético, la relación normal (NR) y el NR ponderado con correlaciones más básicas o relaciones lineales directas entre variables., mientras que la estrategia de imputación múltiple utilizada por la cadena de Markov de Monte Carlo basada en maximizar el valor esperado (EM-MCMC) y la red neuronal de tipo perceptrón multicapa son computacionalmente intensivas. Se descubrió que, aunque requiere mucho trabajo computacional, el método EM-MCMC resultó ser más beneficioso para la imputación de series de tiempo meteorológicas. Para evaluar el rendimiento, se propuso una modificación en esta técnica y el uso de la medida de la dimensión de correlación, que evalúa las interdependencias espacio-temporales dentro de los datos, para evaluar el rendimiento. El estudio encontró que el uso del EM-MCMC antes de los análisis estadísticos reduce la incertidumbre y da resultados más sólidos. Además,

sugirió la dimensión de correlación como una medida efectiva para evaluar la imputación, especialmente con métodos computacionales.

Membreño & Zavala (2015) proponen en este trabajo utilizar métodos estadísticos que permiten obtener un análisis de la relación entre las variables climáticas y otras variables; utilizan los Modelos de Regresión; ellos analizan la precipitación como variable dependiente y las variables independientes latitud, longitud y elevación (usando curvas de nivel); y realizaron un análisis de serie temporal para una estación meteorológica en Chinandega y analizaron el comportamiento de la variable precipitación. Luego realizaron varias cartografías continuas de la precipitación promedio mensual entre 1984 y 2005 utilizando el método de interpolación IDW, para analizar la calidad de cada una de las cartografías continuas.

Ferrari & Ozaki (2014) explican en este articulo cómo se utilizaron las técnicas de imputación estadística y control de calidad para una base de datos de precipitación diaria de estaciones meteorológicas en el Estado de Paraná, Brasil. Después de la imputación, los datos se sometieron a un proceso de control de calidad para evitar errores. Estos errores incluyeron precipitaciones idénticas durante siete días consecutivos y valores de precipitación muy diferentes de los de las estaciones meteorológicas cercanas. Luego, modelaron la sequía agrícola utilizando la teoría de los valores extremos, considerando el número máximo de días consecutivos con precipitación inferior a 7 mm en las principales regiones agrícolas sojeras del Estado de Paraná para el período comprendido entre enero y febrero.

Campozano et al. (2014) analizan cómo funcionan 17 métodos deterministas para recopilar datos diarios sobre la precipitación y la temperatura en la cuenca del río Paute en los Andes Ecuatorianos. En este trabajo, se dio preferencia a los métodos determinísticos por su robustez, facilidad de implementación y eficiencia computacional. Esto a pesar de la existencia de métodos de relleno más sofisticados, como los métodos estocásticos o los métodos de inteligencia artificial, los hallazgos indican que para completar series temporales de precipitación diaria, el método de regresión lineal múltiple ponderada es el mejor. Para la temperatura, la media climatológica del día fue claramente el mejor método.

En la investigación de su tesis doctoral sobre la aceleración de algoritmos utilizando computación con múltiples núcleos para la clasificación supervisada como agrupamiento, se evidencio una significativa mejora en los tiempos de respuesta y aceleraciones por la paralelización, presento un enfoque de CPU y GPU, la Multi-core y many-core demostro ser eficaz, haciendo viables algoritmos que antes eran lentos en una arquitectura secuencial. Esto es especialmente relevante en aplicaciones que requieres repuestas en tiempo real o con grades volúmenes de datos Lorena (2022).

Sandín Carral (2015) abordan en su estudio de Análisis e implementación de algoritmos en nuevas tecnologías de paralelización la evaluación de técnicas como KNN y SVM

(Suport Vector Machine) el análisis se centra en la eficiencia de la computación paralelizada con distintos dispositivos entre CPU como procesadores de Smartphones, realizando comparaciones entre las versiones paralelas y secuenciales, en términos de tiempos de ejecución y consumo energético, el estudio concluye destacando las ventajas de la paralelización en rendimiento y eficiencia energética respecto a sus distintos hardware utilizados,enfocados en la optimización de procesos con dispositivos computacionales.

3. Marco Teórico

Este capítulo presenta el marco teórico que sustenta el progreso del proyecto. Se divide en dos partes, la primera sección trata sobre los conceptos fundamentales que ayudan a comprender adecuadamente los términos relacionados con la precipitación, La segunda sección explica las técnicas estadísticas y teoría relacionada con la imputación de datos faltantes en la precipitación.

3.1. Marco Conceptual

3.1.1. Geoestadística

La geoestadística es una rama de la estadística que estudia cómo las variables regionalizadas se comportan en el espacio. La geología de minas, la geología del petróleo, hidrogeología, oceanografía, sensores remotos, y meteorología entre otras ramas de la geociencias emplean métodos geoestadísticos para resolver sus problemas. El objetivo de la geoestadística es centrarse en la variabilidad espacial y la modelización de la dependencia espacial de los datos, haciendo hincapié en su aplicación a fenómenos geográficos y espaciales (Alperin 2013).

3.1.2. Precipitación

La precipitación en la atmósfera es un fenómeno de tipo hidrometeoro, definido como cualquier fenómeno natural observado sobre la superficie específicamente en su estado líquido o sólido, sin confundir con las formas de condensación, estas se forman a grandes alturas y entre más alta este la nube más fría será la lluvia hasta incluso poder cristalizarse y tener así forma de cristales de hielo, nieve, aguanieve, entre otras. Medida por un pluviómetro el cual es una probeta que mide en milímetros de altura la cálida del agua (DANE 2005).

El pluviómetro es un dispositivo que mide la cantidad de agua que ha caído en un lugar y en un momento específico. Es un cilindro de gran tamaño, similar a un embudo, desemboca en un tubo más estrecho donde se lleva a cabo la medición, se registra la cantidad de precipitación en milímetros (mm) o en litros por metro cuadrado (l/m2). Para que nos hagamos una idea: si precipitan 15 mm, queremos decir que la capa de agua sería de 15 mm si toda la lluvia se acumulara en un terreno acotado de 1 metro cuadrado plano sin escurrirse ni evaporarse (The Weather Channel 2017).

3.1.3. Estación Climatológica

Las estaciones climatológicas cumplen con unas reglamentaciones básicas ubicadas en zonas terrestres las cuales registran diferentes variables, como temperaturas, precipitación, contaminación del aire, humedad, etc. Las estaciones proporcionan las mediciones que están registradas en el tiempo exacto, contando así con un registro de gran volumen, por horarios dependiendo de la estación, con el fin de definir la climatología de la zona, los cuales fueron proporcionados por el Instituto de Hidrología, meteorología y Estudios Ambientales (IDEAM 2023).

3.1.4. Clima del departamento del Meta

El Departamento del Meta, es uno de los treinta y dos departamentos que, junto con Bogotá, Distrito Capital, componen el territorio de la República de Colombia. Se localiza en el Centro del país, más exactamente al este de la Cordillera Oriental, en la región de la Orinoquía colombiana. Cuenta con una superficie 85.635 Km^2 (Figura **3-2**), lo que representa el 7.49 % del territorio nacional. Su capital es la ciudad de Villavicencio y está dividido política y administrativamente en 29 municipios (TodaColombia.com 2023*a*).

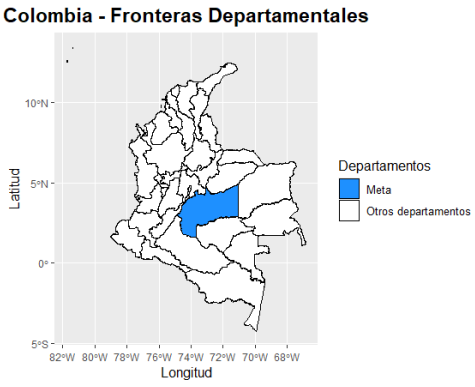

Figura 3-1.: Mapa de Colombia- Meta

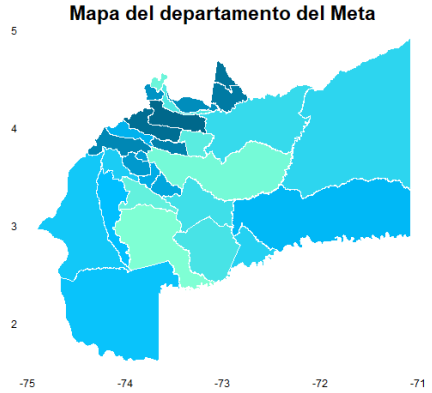

Figura 3-2.: Mapa municipios del meta

El Departamento del Meta está en la Zona de Confluencia Intertropical (ZCIT). Por tanto las precipitaciones varían desde 2.000 mm, en las partes altas de la cordillera, hasta los 6.000 mm o más por año, en cercanías de los municipios de El Castillo y Lejanías. Entre diciembre y marzo se presenta el período más seco, debido a que los vientos alisios del noreste son los dominantes en esta época del año y desplazan hacia el sur la ZCIT. El período de lluvias se extiende de marzo a noviembre, debido a que en esta época los vientos alisios del sureste empiezan a ser los dominantes, desplazando la ZCIT hacia el norte. El prolongado período de lluvias se debe al doble paso de la ZCIT por la alternancia de los vientos alisios dominantes. La temperatura del departamento varía desde un promedio de 6 °C, en el páramo, hasta temperaturas promedio de más de 24 °C en la llanura; en el piedemonte la temperatura

oscila entre 18 y 24 °C. De acuerdo con la variación de altura que hay en el departamento, los pisos térmicos presentes en su territorio son páramo (1,44 % del total), piso climático frío (4,47 %), medio (5,06 %), y cálido (89,03 %). La vegetación de la llanura está formada por pastos y pajonales con abundantes arbustos y árboles de baja altura. En las riberas de los ríos se encuentran los bosques de galería de gran variedad florística; en el occidente del departamento la vegetación es de bosque húmedo tropical, bosques andinos y páramo en las partes más altas (TodaColombia.com 2023*b*).

3.1.5. Programación paralelizada

La programación en paralelo es la ejecución de procesos que requieren un número elevado de cálculos, dividiendo la cantidad de tareas en subprocesos que pueden ser ejecutados de forma simultanea. Esto conlleva la creación de código que será diseñado para ejecutarse en paralelo, aprovechando el número de unidades de procesamiento disponibles en el hardware, este enfoque gestiona eficientemente la asignación y el uso de recursos compartidos como la CPU y la memoria RAM, en el cual se garantiza que los múltiples procesos en ejecución, no presenten interferencias entre ellos, esta metodología, ampliamente descrita por el Lawrence Livermore National Laboratory (Barney 2023), optimiza el rendimiento . El objetivo de la programación paralela es reducir el tiempo de ejecución total del programa. Este objetivo se vuelve fundamental para las aplicaciones desarrolladas que se vuelven cada vez más demandantes, con un gran enfoque en el rendimiento computacional. Aunque los costos de computadoras potentes están disminuyendo, la necesidad de aplicaciones en servidores que requieren una gran cantidad de procesamiento es obvia, ya sea por cálculos complejos, una gran cantidad de cálculos, o una gran demanda de usuarios, etc. Para una descripción más detallada sobre la programación paralelizada, ver Grama et al. (2003).

3.2. Marco Teórico Estadístico

3.2.1. Datos Faltantes

En términos generales, el término de datos faltantes significa que hace falta algún tipo de información sobre los fenómenos que nos interesan (McKnight et al. 2007). En esencia, la ausencia de estos valores oculta información valiosa que podría influir en las conclusiones del análisis (Little & Rubin 2019).

Existen varias técnicas para abordar la imputación de datos faltantes, dependiendo de la estructura de los datos. Por ejemplo, en la década de los sesenta, los modelos Hot-Deck y Cold-Deck se utilizaban comúnmente para imputar datos en encuestas y censos. Sin embargo, Rubin en 1976 utilizó una metodología de inferencia estadística en la que se asumían los datos faltantes como variables aleatorias y los datos se imputaron sin alterar los modelos. Posteriormente, Rubin introdujo el término imputación múltiple en 1987, proponiendo que una serie de simulaciones que deberían usarse para imputar cada dato faltante (Medina Rivera 2008).

3.2.2. Aleatoriedad de los datos faltantes

La aleatoriedad de los datos faltantes se puede dividir en tres clases como proponen Little & Rubin (1987):

- Faltantes completamente al azar (MCAR): este es el nivel más alto de aleatoriedad. Ocurre cuando la probabilidad de que una instancia (caso) tenga un valor faltante para un atributo no depende ni de los valores conocidos ni de los datos faltantes. En este nivel de aleatoriedad, cualquier método de tratamiento de datos faltantes se puede aplicar sin riesgo de introducir sesgos en los datos.

- Faltantes al azar (MAR): cuando la probabilidad de que a una instancia le falte el valor de un atributo puede depender de los valores conocidos, pero no del valor de los datos faltantes en sí.

- No faltantes al azar (NMAR): cuando la probabilidad de que una instancia tenga un valor que falta para un atributo podría depender del valor de ese atributo.
  En la literatura se han propuesto varios métodos para tratar los datos faltantes. Muchos de estos métodos, como la sustitución de casos, se desarrollaron para hacer frente a la falta de datos en encuestas por muestreo, y tienen algunos inconvenientes cuando se aplican a la minería de datos. Otros métodos, como la sustitución de los valores perdidos por la media del atributo o moda, son muy sencillos y deben usarse con cuidado para evitar la inserción de sesgos.

Para profundizar estos conceptos mencionados previamente, se muestran las expresiones matemáticas en Verbeke & Molenberghs (2009); los autores definen el modelo estadístico de los mecanismos de datos faltantes teniendo en consideración la densidad total de datos como:

$$f(\overrightarrow{y_i}, \overrightarrow{r_i}|X_i, \overrightarrow{\theta}, \overrightarrow{\psi}) \tag{3-1}$$

Donde,$\overrightarrow{y_i}$ es un vector que contiene las observaciones de la variable de respuesta, $\overrightarrow{r_i}$ es un vector binario que indica la presencia o ausencia de datos en el vector $\overrightarrow{y_i}$, X_i representa la matriz de covariables, $\overrightarrow{\theta}$ son vectores que parametrizan la distribución conjunta, y $\overrightarrow{\psi}$ para describir los procesos de medición y datos faltantes, donde el proceso de medición se refiere a cómo se recolectaron los datos observados, mientras que el proceso de datos faltantes aborda cómo se determinaron los valores faltantes en los datos observados. La taxonomía propuesta por Rubin (1976), y desarrollada por Little & Rubin (1987) clasifica los mecanismos de datos faltantes en dos componentes: el proceso de medición y el proceso de faltantes. Esta clasificación se expresa mediante la factorización:

$$f(\overrightarrow{y_i}, \overrightarrow{r_i}|X_i, \overrightarrow{\theta}, \overrightarrow{\psi}) = f(\overrightarrow{y_i}|X_i, \overrightarrow{\theta})f(\overrightarrow{r_i}|\overrightarrow{y_i}, X_i, \overrightarrow{\psi}) \tag{3-2}$$

donde el primer factor es la densidad marginal del proceso de medición y el segundo es la densidad del proceso de datos faltantes, condicionado a los resultados. Es posible tener covariables adicionales en el modelo de faltantes, pero esto se elimina de la notación.

La factorización (3-2) constituye la base del modelado de selección, ya que el segundo factor corresponde a la (auto)selección de individuos en grupos observados y ausentes.

La taxonomía basada en el segundo factor de la factorización (3-2) es esencial para comprender estas clasificaciones. Se define como:

$$f(\overrightarrow{r_i}|\overrightarrow{y_i}, X_i, \overrightarrow{\psi}) = f(\overrightarrow{r_i}|\overrightarrow{y_i^o}, \overrightarrow{y_i^m}, X_i, \overrightarrow{\psi}) \tag{3-3}$$

- Donde $\overrightarrow{y_i^o}$ son las mediciones observadas y $\overrightarrow{y_i^m}$ son las observaciones no observadas (datos faltantes). Si (3-3) es independiente de las mediciones [es decir, cuando asume la forma $f(\overrightarrow{r_i}|X_i, \overrightarrow{\psi})$], entonces el proceso se denomina faltantes completamente al azar (MCAR).
- Si (3-3) es independiente de $\overrightarrow{y_i^m}$, pero depende de $\overrightarrow{y_i^o}$, asumiendo así la forma $f(\overrightarrow{r_i}|\overrightarrow{y_i^o}, X_i, \overrightarrow{\psi})$, entonces el proceso se denomina faltantes al azar (MAR).
- Finalmente, cuando (3-3) depende de los valores faltantes $\overrightarrow{y_i^m}$, el proceso se denomina no faltantes al azar (MNAR). Se permite que un proceso MNAR dependa de $\overrightarrow{y_i^o}$.

3.2.3. Métodos para tratar los datos faltantes

Los métodos de tratamiento de datos faltantes se pueden dividir en las siguientes tres categorías (Little & Rubin 1987):

- Ignorar y descartar datos: hay dos formas principales de descartar datos a los que le faltan valores. El primero se conoce como análisis completo del caso. Este método consiste en descartar todos los casos con datos faltantes.
  El segundo método es conocido como descartar casos y/o covariables. Este método consiste en determinar la extensión de los datos que faltan en cada caso y covariable, y eliminar los casos y/o covariables con altos niveles de datos faltantes.
  Ambos métodos se deben aplicar teniendo en cuenta con que tipo de mecanismo de datos faltantes cuenta y el fenómeno que se desea estudiar, puesto que si se realizan bajo el supuesto de MAR o MNAR se podrían obtener estimaciones sesgadas.

- Manejo de Datos Faltantes con Máxima Verosimilitud: los Métodos que utilizan la Máxima Verosimilitud, se emplean para estimar los parámetros de un modelo, tanto con la completitud de los datos o incluso variantes que permiten trabajar con datos faltantes, como lo son procedimientos que usan variantes del algoritmo Expectation-Maximization (EM)(Dempster & Rubin 1977) los cuales pueden manejar estimación de parámetros en presencia de datos faltantes, estos métodos emplean un modelo probabilístico que permiten estimar los valores faltantes, basándose en la función de verosimilitud de los datos observados, es particularmente útil en situaciones donde los datos faltantes siguen un patrón aleatorio, es decir, cuando los datos faltantes son MAR o MCAR.

- Imputación Múltiple: la imputación múltiple es una clase de procedimiento que tiene como objetivo completar los valores faltantes con los observados, teniendo en cuenta la información disponible que proporcionan mas variables que la variable que se desea completar. El objetivo es emplear relaciones conocidas que puedan ser identificadas en los valores válidos del conjunto de datos para ayudar a estimar los valores faltantes. Este método funciona para los datos faltantes en cualquier patrón, ya sea MCAR, MAR o MNAR. Sin embargo, es importante destacar que la eficacia de la imputación múltiple puede variar de acuerdo con el patrón particular de datos faltantes y la calidad de las relaciones encontradas entre las variables.

3.2.4. Medidas de distancia

Distancia Euclidiana

la distancia euclidiana, entre dos puntos x y y, en un espacio de uno, dos, tres o más dimensiones, viene dada por la siguiente fórmula:

$$d_e(\overrightarrow{x}, \overrightarrow{y}) = \left\| \sqrt{(\overrightarrow{y} - \overrightarrow{x})^T (\overrightarrow{y} - \overrightarrow{x})} \right\| \tag{3-4}$$

donde $\overrightarrow{x}$ y $\overrightarrow{y}$ son vectores que representan las coordenadas de los puntos en el espacio n-dimensional (Pang-Ning Tan et al. 2006*a*).

Distancia de Mahalanobis

Un tema relacionado es cómo calcular la distancia cuando hay correlación entre algunos de las variables, quizás además de las diferencias en los rangos de valores. Una generalización de la distancia euclidiana, es la distancia de Mahalanobis. Es útil cuando las variables están correlacionadas, tienen diferentes rangos de valores (diferentes varianzas), y la distribución de los datos es aproximadamente gaussiana (normal). Específicamente, la distancia de Mahalanobis entre dos vectores, $\overrightarrow{x}$ y $\overrightarrow{y}$ se definen como:

$$d_m(\overrightarrow{x}, \overrightarrow{y}) = \sqrt{(\overrightarrow{x} - \overrightarrow{y}) \Sigma^{-1} (\overrightarrow{x} - \overrightarrow{y})^T} \tag{3-5}$$

donde Σ^{-1} es la matriz de covarianza de los datos (Pang-Ning Tan et al. 2006*b*).

3.2.5. Métodos de imputación de datos faltantes

K-Vecinos Más Cercanos(k-Nearest-Neighbor)

La regla del vecino más cercano, que se basa en una medida de similitud o distancia para determinar el valor de un dato desconocido, es un método de aproximación simple no paramétrica (Cover & Hart 1967). Para predecir el valor de los nuevos datos, el método del vecino más cercano se puede extender utilizando no uno, sino un conjunto de datos más cercanos; estos se denominan k-vecinos más cercanos (K-NN). Al tener en cuenta varios vecinos, el cálculo de los valores considera mayor información disponible y se otorga protección contra el ruido (Moreno 2004).

La regresión de funciones con valores continuos se puede realizar fácilmente con el método de K-NN (Moreno 2004). El algoritmo asume que todos los datos pertenecen a R^p y, para aproximar una función $f : R^p \to R$, a partir de los k valores ya seleccionados, se determinan k datos más cercanos al nuevo dato x_q mediante una medida de distancia en ese espacio. Esta función es el promedio de los k valores más cercanos; si se considera el promedio aritmético (todos los datos del grupo tienen la misma importancia), la función aproximación tiene la forma que se muestra a continuación:

$$\hat{f}(x_q) \leftarrow \frac{1}{k} \sum_{i=1}^{k} f(x_i) \tag{3-6}$$

Para evitar que las características de los conjuntos de entrada con valores más altos dominen el cálculo de la distancia, todos los datos deben estar normalizados (Cover & Hart 1967).

XGBoost

En Chen & Guestrin (2016*a*) definen el extreme gradient boosting (XGBoost) como un algoritmo de aprendizaje supervisado que sirve tanto para clasificación como para regresión. El XGBoost está basado en arboles de decisión y es paralelizable y escalable. El algoritmo entrena secuencialmente el árbol de decisión en datos de entrenamiento. Para aumentar el valor de la función objetivo, el algoritmo agrega un nuevo árbol de decisión a cada iteración de los árboles de decisión anteriores. La función objetivo que se pretende minimizar consiste en el término de pérdida (l) y el término de regularización (Ω). La función objetivo de la t-ésima iteración (L_t) se define por la ecuación (3-7):

$$L^t = \sum_{i=1}^{n} [l(y_i, \hat{y_i}^{t-1} + f_t(x_i))] + \Omega(f_t) \tag{3-7}$$

Donde y_i es la etiqueta de clase real de la observación i (el valor verdadero que se espera predecir), $\hat{y}_i$ es la etiqueta de clase prevista de la observación i (el valor que el modelo predice), f_k es la función del árbol, n es el número de observaciones en el conjunto de entrenamiento y Ω es el término de regularización.

$\Omega(f_t)$ en(3-8) penaliza la complejidad del modelo para evitar el sobreajuste:

$$\Omega(f_t) = \gamma T_t + \frac{1}{2} \sum_{t=1}^{n} w_j^2 \tag{3-8}$$

Dónde γ y λ son los hiperparámetros, T es el número de hojas en el árbol, y w es el peso de cada hoja.

El valor de la función de pérdida en (3-9) se aproxima mediante la expansión de Taylor de segundo orden:

$$L^t \cong \sum_{i=1}^{n} [l(y_i, \hat{y_i}^{t-1}) + g_i f_t(x_i) + (1/2) h_i f_t^2(x_i)] + \Omega(f_t) \tag{3-9}$$

$$g_i = \partial_{\hat{y_i}^{t-1}} l(y_i, \hat{y_i}^{t-1}) \tag{3-10}$$

Donde g_i representa las estadísticas de gradiente de primer orden sobre la pérdida.

$$h_i = \partial^2_{\hat{y_i}^{t-1}} l(y_i, \hat{y_i}^{t-1}) \tag{3-11}$$

Donde h_i representa las estadísticas de gradiente de segundo orden sobre la pérdida.

Para una estructura de árbol fija, (3-12) y (3-13) indican el peso ideal de la hoja en el nodo de la hoja j y el valor ideal de la función objetivo, donde I_j es el conjunto de observaciones de la hoja j. La función de puntuación para evaluar la calidad de la estructura del árbol que es la ecuación (3-13).

$$w_j^* = -\frac{\sum_{i\varepsilon I_j} g_i}{\sum_{i\varepsilon I_j} h_i + \lambda} \tag{3-12}$$

$$L(q) = -\frac{1}{2}\sum_{j=1}^{T} \frac{(\sum_{i\varepsilon I_j} g_i)^2}{\sum_{i\varepsilon I_j} h_i + \lambda} + \gamma T \tag{3-13}$$

Sin embargo, para obtener la estructura de árbol ideal, se deben evaluar todas las estructuras de árbol posibles. Por lo tanto, debido al costo computacional es imposible considerar todas las estructuras de árbol posibles. En realidad, XGBoost utiliza un algoritmo iterativo para determinar la estructura de árbol ideal.

El XGBoost maneja los datos faltantes internamente sin usar métodos de imputación o eliminación. Como se muestra en la Figura **3-3**, clasifica una observación con una característica faltante en la dirección predeterminada en cada nodo. La derecha y la izquierda son las dos opciones, y se selecciona la dirección que maximiza la ganancia del conjunto de entrenamiento. La ganancia del conjunto de entrenamiento se elige como dirección predecible.

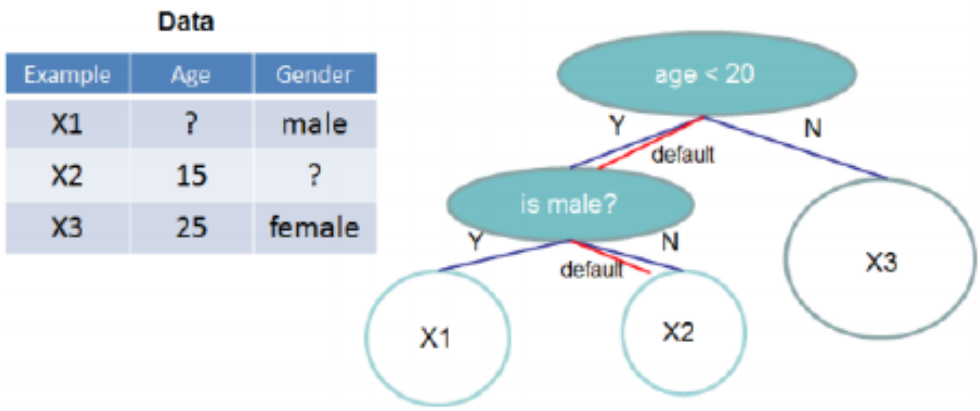

Figura 3-3.: Ejemplo de cómo XGboost maneja los datos faltantes (Chen & Guestrin 2016*a*).

Random Forest

Los bosques aleatorios (Random Forest), creados por Breiman (2001) Ibarra-Berastegi et al. (2011), se basan en árboles de decisión y son extremadamente adaptables y poderosos

clasificadores de conjuntos. Además, el marco proporciona una comprensión de la capacidad del Random Forest para predecir en términos de la fuerza de los predictores individuales y sus correlaciones entre ellos. El Random Forest se puede usar para clasificación, regresión y aprendizaje no supervisado Yang et al. (2017).

Sea $X = (X_1, X_2, ...X_p)$ una matriz de datos de dimensión $n \times p$ y X_s como una variable arbitraria con valores faltantes en las entradas $i^s_{mis} \subseteq \{1, ...n\}$ el conjunto de datos se puede separar en dos categorías:

- los valores observados de la variable X_s denotada por y^s_{obs}.
- los valores faltantes de la variable X_s denotada por y^s_{mis}.

Una estimación inicial del valor faltante en X puede ser determinado utilizando la media u otro método de imputación. Las variables X_s, $s = 1, ...p$, Se clasifican de acuerdo con el número total de valores faltantes, comenzando con el número más pequeño. Para cada variable X_s, los valores faltantes son imputados por un Random Forest adecuado con respuesta y^s_{obs} y predictores x^s_{obs}; y prediciendo los valores faltantes de y^s_{mis}, aplicando el Random Forest entrenado a x^s_{mis}. Luego se repite el procedimiento de imputación hasta que se detiene se cumple el criterio. En la Figura **3-4** se muestra la estructura del Random Forest.

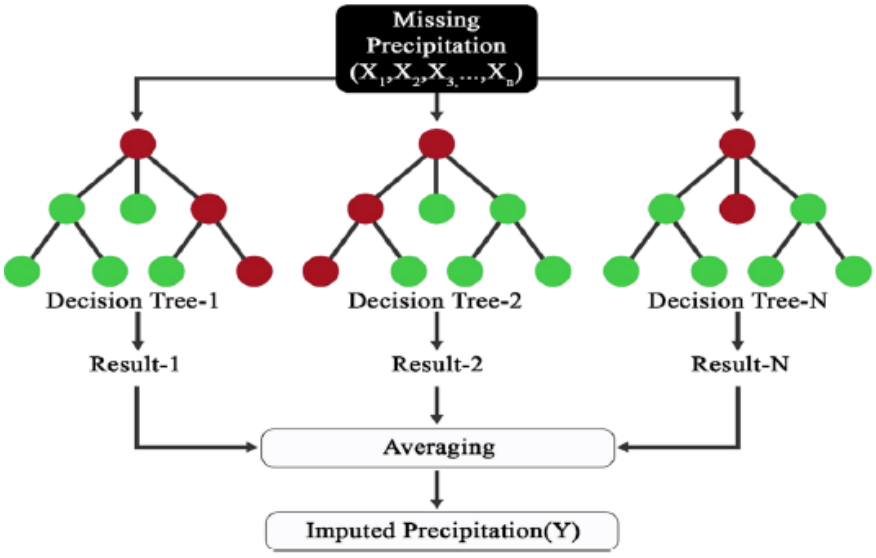

Figura 3-4.: Esquema del Random Forest (Sahoo & Dillip Kumar 2022).

3.2.6. Indicadores de desempeño

Los siguientes son los indicadores de desempeño utilizados para comparar la eficiencia de las metodologías:

RMSE

La raíz del error cuadrático medio (RMSE), Se define como la raíz cuadrada de la media de los errores cuadráticos entre los valores reales Y_i y los valores predichos $\hat{Y}_i$:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{Y}_i - Y_i)^2}{n}} \tag{3-14}$$

Donde n es el número de observaciones. El valor RMSE calculado más bajo se obtiene de la metodología más efectiva (Sahoo, Samantaray & Ghose (2021*a*), Sahoo, Samantaray & Paul (2021)).

MAE

El error absoluto medio (MAE) se define como la media de las diferencias absolutas entre los valores reales (Y_i) y los valores predichos ($\hat{Y}_i$):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\hat{Y}_i - Y_i)| \tag{3-15}$$

Un valor de MAE más bajo indica una mayor precisión en las predicciones (Willmott et al. 2009, Agnihotri et al. 2022).

Coeficiente de determinación

El coeficiente de determinación (R^2) da una varianza proporcional de una variable predicha a partir de otra variable, esta métrica se calcula de la siguiente forma:

$$R^2 = (\frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}})^2 \tag{3-16}$$

Donde $\bar{Y}$ y $\bar{Y}_i$ son los valores medios de los datos observados y estimados respectivamente. Los valores del coeficiente de determinación (R^2) oscilan entre 0 y 1. El valor R^2 igual a 1 indica una combinación ideal de datos observados y estimados, cuando el R^2 esta cercano a 0, los métodos son extremadamente ineficaces en comparación con los datos observados. El valor R^2 cercano a 1 indica mayor precisión (Samantaray et al. (2021), Sahoo, Samantaray & Ghose (2021*b*)).

MAPE

El error porcentual absoluto medio (MAPE) esta definido como:

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \tag{3-17}$$

Donde A_t y F_t denotan los valores reales y previstos en el punto de datos t, respectivamente, y n es el numero de puntos de datos. Sin embargo, MAPE tiene la importante desventaja de que produce valores infinitos o indefinidos para valores reales cero o cercanos a cero (Kim & Kim 2016).

4. Metodología

En este capítulo se presenta la propuesta metodológica de este proyecto, que incluye los métodos de estimación utilizados para los datos faltantes en la serie de precipitación diaria del departamento de Meta. Se describen también los escenarios relevantes, teniendo en cuenta las distancias y los porcentajes de datos faltantes.

Este estudio se centra en la evaluación de diversas metodologías de imputación de datos faltantes y la comparación de sus métricas de desempeño, así como en la exploración de los tiempos de cálculo utilizando técnicas de computación paralela. La ubicación geográfica y las características temporales influyen en las mediciones diarias de precipitación. Por lo tanto, es esencial considerar estos factores al evaluar la efectividad de las metodologías de imputación.

También se plantean escenarios de porcentajes de datos faltantes con umbrales del 10 %, 15 % y 20 % para examinar la dependencia entre el porcentaje de datos faltantes y la precisión de las imputaciones. Se destaca que, en el periodo analizado, el porcentaje real de datos faltantes fue del 8.27 %. La metodología propuesta incluye una comparación exhaustiva de todos los métodos de imputación seleccionados y sus respectivos escenarios de simulación, con el fin de verificar cómo se comporta el desempeño de estas metodologías en diferentes escenarios de datos faltantes.

4.1. Datos de estudio

Se cuenta con datos diarios de la precipitación desde 1980 hasta 2016, contando con 57 estaciones de meteorología del IDEAM, ubicadas en el departamento del Meta. En cada estación se registra la temperatura, altura, fecha, así como sus respectivas coordenadas en latitud y longitud. También algunas estaciones cuentan con medidores adicionales como la cantidad de dióxido de carbono (CO_2) , humedad relativa, radiación, entre otras. Las estaciones forman parte de la red nacional de estaciones meteorológicas del IDEAM y se encuentran en longitudes y latitudes (-71.34075,-74.3537222222) y (2.1761111111,4.5744166667) respectivamente.

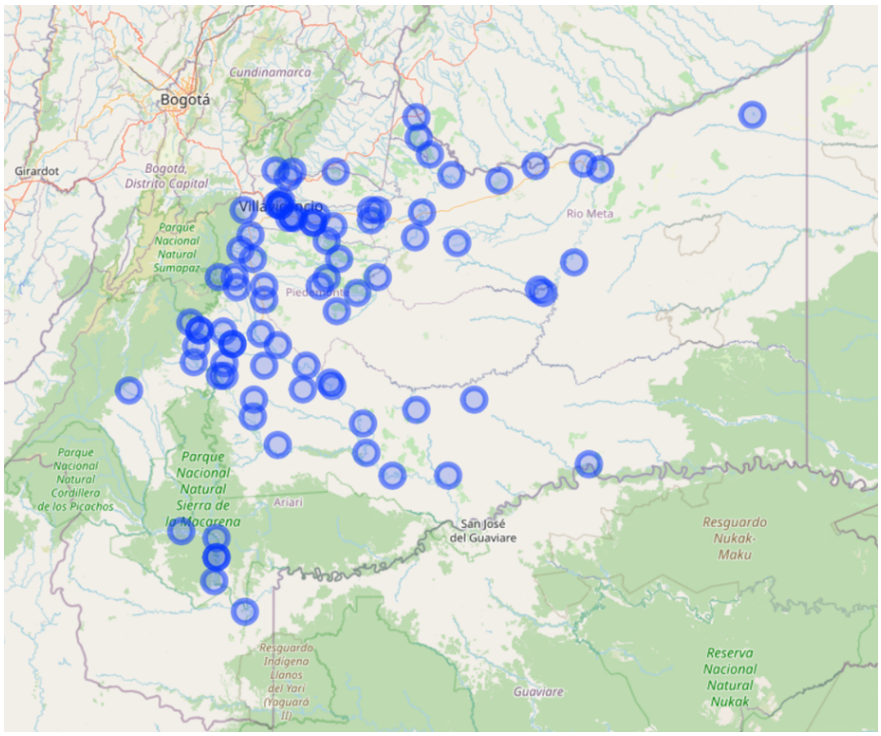

Figura 4-1.: Ubicación geográfica de las estaciones. La intensidad de los puntos describen el volumen de precipitación

Del Gráfico **4-1** También se observa las delimitaciones que presenta con Bogotá, como parte central y sur del país. La distribución geográfica de las estaciones meteorológicas no solo refleja una concentración en ciertos municipios, sino también una variabilidad en términos de la cantidad de precipitaciones registradas, esto se presenta mediante una escala de colores: cuanto más intensos son los puntos, mayor es la cantidad de precipitación. Este fenómeno puede atribuirse a diversos factores climáticos y geográficos que caracterizan a la región. También se destaca la intensidad en los puntos más cercanos a la capital y en zonas como el Parque Natural Sierra de la Macarena, donde se registra una mayor cantidad de precipitación

registrada, además, esta información es complementaria con los mapas brindados por la figura **4-2**, del DANE (2005).
La distribución espacial de las estaciones, así como la intensidad de la precipitación registrada, son indicativos de la diversidad climática del Departamento del Meta, así descrita en el capitulo 3.1.4, este análisis detallado permite una mayor comprensión de los patrones asociados al comportamiento de la precipitación, el cual es fundamental para la toma de decisiones informadas en sectores como la agricultura, gestión del riesgo, planificación territorial y el impacto en las metodologías de imputación de datos faltantes utilizadas en la literatura como en la presente investigación.

El proceso de recolección de los datos se puede realizar por medio de las paginas oficiales del IDEAM[1], donde corresponden a las series históricas de la precipitación diaria obtenidas por las estaciones de monitoreo, sin embargo cuando se necesita un volumen alto de información, se debe solicitar por medio de los canales oficiales.

4.2. Control de Calidad de los datos de estudio

La calidad de los datos es esencial para identificar registros incorrectos o no verificados (Vicente-Serrano et al. 2010). Los datos para este estudio se obtuvieron solicitando información al IDEAM mediante correo electrónico (IDEAM 2023). La información de todas las estaciones del departamento del Meta, incluyendo variables climáticas, registros espaciales y temporales, se compiló para su análisis posterior debido al volumen significativo de datos.

El software estadístico R, versión 4.3.2 (2023-10-31), y el entorno de desarrollo RStudio (RStudio Team 2012) se emplearon para modificar y transformar estos datos. La lectura y el procesamiento se realizaron mediante la recopilación de archivos en formato txt, a través de la información de cada una de las estaciones, se adjuntó en un formato que resuma por el identificador de la estación, lo que culminó en un archivo en formato .RData, que contiene la compilación de datos presentados como una variable en datos estructurados, facilitando el análisis a través de diversas estaciones y periodos temporales. Esto incluyó información esencial como la fecha, registros de variables hidrológicas entre otras variables, como la identificación de la estación meteorológica, latitud y longitud. La organización de los datos utilizadas en los modelos se efectuó en un formato estructurado de tipo plano, abarcando todos los registros, incluso aquellos con datos faltantes. Los detalles espaciales y temporales se conservaron para permitir un análisis completo, además busca detectar errores, valores inusuales o sesgos que puedan tener un impacto en la validez de los análisis posteriores (Figueras & Gargallo 2003). Dada la naturaleza de la variable objeto de estudio, no puede contener valores negativos así como valores extremos. Este enfoque garantiza que, incluso frente a los desafíos de los datos faltantes, la integridad espacial y temporal de la información se mantiene, facilitando una evaluación precisa del estudio.

[1] http://www.ideam.gov.co/web/tiempo-y-clima

4.3. Análisis exploratorio

Antes de cualquier modelado estadístico, se requiere el Análisis Exploratorio de Datos (AED). El objetivo es comprender las estructuras básicas de las variables, así como las relaciones potenciales entre ellas y su calidad, que para el caso de la precipitación se valida con la información publicada en el IDEAM en los mapas mensuales de la precipitación máxima en 24 horas en un día, estas prácticas permiten conocer información preliminar que puede ser útil para el entendimiento del comportamiento de los modelos. Haciendo alusión a esta sección se muestra el resumen del análisis exploratorio del archivo de datos seleccionando 57 estaciones meteorológicas con medición de precipitación diaria comprendidas entre el año 1980 hasta el 2016, ubicadas en el departamento del Meta. Además, consultando con las fuentes de información verificadas y publicadas por el IDEAM en sus mapas mensuales de precipitaciones máximas absolutas en 24 horas, preparado por el grupo Climatología, Agroclimatología y Meteorología Marina de la Subdirección de Meteorología, el mapa presentado en la figura **4-2** es tomado del último documento publicado en el año 2023 de datos abiertos IDEAM (2023), que contiene información correspondiente en un rango de tiempo entre 1971 a 2005, con alrededor de 1900 estaciones distribuidas en todo el país, que tiene como objetivo observar los comportamientos de los registros de las estaciones pluviográficas.

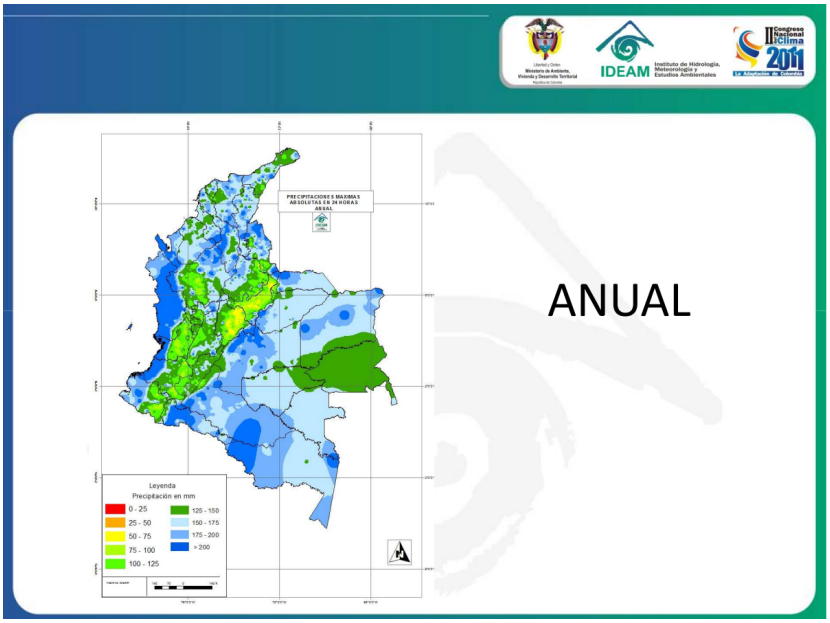

Figura 4-2.: Mapa Precipitación en Colombia - medida máxima en 24 horas - Anual

4.4. Procedimiento Aplicado en la investigación.

En esta sección se detallan los procedimientos adoptados para comparar los métodos de imputación utilizados en el estudio, por medio de la validación cruzada, se evaluarán los diferentes enfoques de imputación empleando un conjunto completo de registros, de modo que se presenta el total de las 56 estaciones meteorológicas con datos completos, contando así con un total de 592.124, por lo tanto será esta la condición con la se realizará la comparación de los modelos. Para cada zona, se han recolectado la información de la precipitación, incluyendo su respectiva longitud y latitud. Estos datos geoespaciales son cruciales, ya que permiten un análisis detallado y contextualizado de las condiciones climáticas en diferentes puntos geográficos.

La extracción de los datos se toma por medio de mecanismos aleatorios, para el propósito de esta investigación se optó por un mecanismo MCAR, asumiendo que los datos faltantes no están influenciados por factores específicos relacionados con la precipitación o con otras características geográficas. Esto considerando lo expuesto por Medina Rivera (2008), los datos faltantes se deben mayormente a factores externos, la falta de datos de la precipitación puede ser causada por agentes que no hacen parte de la naturaleza de la variable, como la falta de personal, daños eléctricos, pérdida o destrucción de registros, problemas con los instrumentos de medición y errores en el proceso de investigación. Se consideró la generación de escenarios del 10 % , 15 % y 20 % de datos faltantes Jacob Junior et al. (2024). Posteriormente, se procedió hacer la imputación de los registros según la información disponible de otras estaciones para el mismo día. Finalmente, se realizó la comparación de los métodos planteados y la eficiencia computacional presentada.

4.4.1. Métodos de Imputación de datos faltantes

K-vecinos más cercanos(KNN)

En la implementación de este método, se seleccionarán las estaciones más próximas a la estación objetivo, que requiere la estimación de un dato faltante. Concretamente, se escogerán las “k” estaciones más cercanas que dispongan de información relevante, donde este valor se determinará a través de un proceso de validación cruzada. Este enfoque permite evaluar distintas cantidades de estaciones cercanas para identificar el número óptimo que minimiza el error en la imputación, mejorando así la precisión del modelo.

Un aspecto destacable es la aplicación de la biblioteca scikit-learn Pedregosa et al. (2011), cuya función KNNImputer facilita la imputación de datos faltantes utilizando la distancia Euclidiana. Para este estudio, se adaptó dicha función para calcular la distancia de Mahalanobis, aprovechando además la biblioteca Joblib Varoquaux (2024) para habilitar la ejecución paralela del proceso.

En la fase subsiguiente, se realizó la estimación mediante el cálculo del promedio de los valores reportados por estas cinco estaciones cercanas. Este proceso tendrá en cuenta las coordenadas geográficas tanto de la estación que requiere la imputación de datos como de las estaciones seleccionadas, garantizando que la estimación se base en datos relevantes y contemporáneos a la variable y la fecha específicas. Este promedio constituirá la estimación de la variable para la estación en estudio. La estimación se efectuará diariamente, dado que el método opera en una base diaria, tomando siempre como referencia la localización de la estación (véase Figura **4-3**).

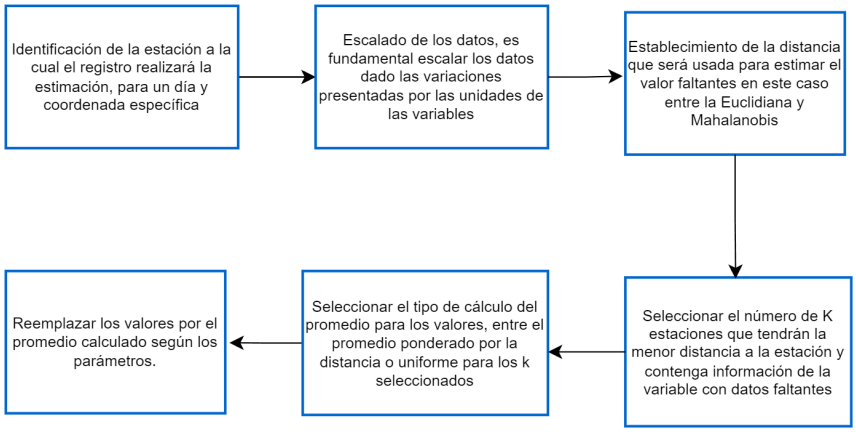

Figura 4-3.: Resumen Metodología K vecinos mas cercanos (Elaboración Propia).

XGBOOST

Consiste en un conjunto secuencial de árboles de decisión. El nombre de este conjunto es CART, que es el acrónimo de árboles de clasificación y regresión. Los árboles se agregan secuencialmente a fin de aprender del resultado de los árboles anteriores y corregir el error producido por los mismos, hasta que ya no se pueda corregir más dicho error. Este proceso se conoce como gradiente descendente. A continuación se enumeran los pasos para el funcionamiento del algoritmo para la imputación de datos faltantes:

- **Optimización de Hiperparámetros**: Antes de entrenar el modelo XGBoost, es fundamental optimizar sus hiperparámetros para maximizar su rendimiento. La tasa de aprendizaje (Learning rate), la profundidad máxima de los árboles (Max-depth), el número de estimadores (n- estimators) y el número de trabajos a realizar en paralelo (n-job), son los hiperparámetros clave. La técnica Grid Search se utiliza para optimizar estos hiperparámetros, para encontrar la combinación ideal que maximice el rendimiento del modelo en los datos de entrenamiento.
- **Identificación de datos faltantes**: Una vez optimizados los hiperparámetros, el primer paso es encontrar las variables en el conjunto de datos de precipitación en

cada estación, destacando las variables que carecen de valores. La precipitación es la única variable que presenta datos faltantes en este estudio y es la variable principal de interés. Por lo tanto, para garantizar la integridad del análisis, nos concentraremos en encontrar y tratar estos valores ausentes.

- **Preprocesamiento de datos**: Los datos deben prepararse antes de aplicar XGBoost. Esto implica definir las características (variables predictoras) y la variable objetivo, así como dividir el conjunto de datos en conjuntos de entrenamiento y prueba.
- **Entrenamiento del modelo XGBoost**: Utiliza las características disponibles y la variable objetivo (en este caso, variables temporales y geográficas). El modelo se entrena utilizando datos completos, excluyendo las observaciones en las que la variable de precipitación carece de datos.
- **Predicción e imputación de datos faltantes**: Una vez que el modelo XGBoost ha sido entrenado, se utiliza para predecir los valores faltantes en la variable de precipitación. Estas predicciones se realizan utilizando las mismas características empleadas durante el entrenamiento. Los valores faltantes de la variable de precipitación son luego reemplazados con los valores predichos por el modelo, completando así el proceso de imputación (Chen & Guestrin 2016*b*).

RANDOM FOREST

El algoritmo del Random Forest se ha adaptado para abordar la imputación de datos faltantes, incluidos los datos de precipitación en estaciones meteorológicas. En este contexto, se describen los pasos específicos para implementar el algoritmo de Random Forest para la imputación de datos de precipitación, basándose en la metodología propuesta por Cutler et al. (2012).

Los pasos para implementar el algoritmo de Random Forest para la imputación de datos de precipitación son los siguientes:

- **Optimización de Hiperparámetros**: Antes de comenzar el proceso de imputación, es esencial optimizar los hiperparámetros del modelo Random Forest. Los hiperparámetros clave incluyen el número de árboles en el bosque (n-estimators), la profundidad máxima de los árboles (Max-depth), el número mínimo de muestras requeridas para dividir un nodo (min-samples-split). La optimización de estos hiperparámetros se realiza mediante búsqueda exhaustiva (Grid Search) para encontrar la combinación óptima que maximice el rendimiento del modelo en el conjunto de datos de entrenamiento.
- **Selección del Conjunto de Entrenamiento:** Después de la optimización de los hiperparámetros, se selecciona un conjunto de entrenamiento S para cada uno de los j

árboles que componen el Random Forest. El conjunto de datos disponible que totaliza 592,124 datos, se divide en un 70 % para el conjunto de entrenamiento y un 30 % para el conjunto de prueba. Es decir, se utilizan aproximadamente 414,487 datos para el entrenamiento y 177,637 datos para la prueba. Esto ayuda en la imputación de datos faltantes para la precipitación, asegurando que los datos completos se incluyan en el conjunto de entrenamiento para ajustar el modelo y hacer predicciones para los datos faltantes.

- **Creación de Muestras Bootstrap:** Para cada árbol del Random Forest, se obtiene una muestra bootstrap de tamaño n del conjunto de entrenamiento S. Esta muestra se utiliza para construir el árbol individual. En python el backend de LightGBM se utiliza en "miceforest", que implementa técnicas de remuestreo y submuestreo internamente según las configuraciones. LightGBM puede realizar submuestreo en cada iteración por defecto, lo que se puede controlar mediante el parámetro subsample.

- **Partición del Conjunto de Datos:** Después de obtener la muestra de entrenamiento, se procede a la partición de los datos,este paso implica dividir el conjunto de datos de entrenamiento en subconjuntos más pequeños para entrenar cada árbol individual en el bosque aleatorio. Inicialmente, todas las observaciones se consideran en el nodo raíz del árbol. Luego, se realiza la partición del conjunto de datos siguiendo un criterio de parada y seleccionando un número de variables aleatorias a utilizar en cada partición.

  Basados en características (variables temporales y geográficas), las observaciones se dividen en subconjuntos. Para cada división de nodo, el Random Forest selecciona aleatoriamente un subconjunto de estas características, lo que ayuda a evitar el sobreajuste y a capturar mejor la variabilidad en los datos de precipitación.

- **Selección de la Partición Óptima:** Finalmente, se determina la partición óptima basándose en el criterio utilizado para realizar las divisiones en cada nodo del árbol. La selección de la partición óptima implica encontrar la mejor manera de dividir los datos en cada nodo del árbol en el contexto de la imputación de datos faltantes para la precipitación. Esto se logra evaluando varios criterios de división, como la ganancia de información o la reducción de la impureza, que miden cómo una división específica separa efectivamente observaciones con datos completos de observaciones con datos faltantes.

La predicción de la precipitación se realiza mediante la moda para los árboles de clasificación o la media de las observaciones en el nodo terminal correspondiente para los árboles de regresión. Este enfoque permite obtener estimaciones robustas y precisas de los valores de precipitación para las observaciones con datos faltantes. En este caso para realizar las predicciones, se utiliza el Random Forest por regresión, ya que la variable precipitación es continua, las predicciones se realizan utilizando la media de las observaciones en el nodo terminal correspondiente de cada árbol de decisión.

Es importante destacar que, aunque el algoritmo de Random Forest es una técnica avanzada y actualizada en 2023, su aplicación específica para la imputación de datos de precipitación en estaciones meteorológicas es fundamental. Esta aplicación particular permite entender cómo esta metodología se adapta a un caso específico, proporcionando perspectivas sobre su efectividad y adecuación para este tipo de datos. La figura **4-4** presenta un diagrama del proceso de imputación utilizando el paquete Miceforest White (2024), ilustrando cómo se aplica el algoritmo de Random Forest para la imputación de datos faltantes de la variable precipitación.

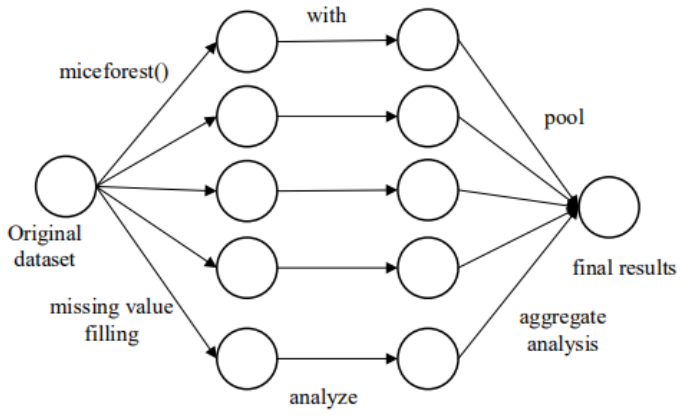

Figura 4-4.: Diagrama de funcionamiento miceforest tomado de Shao & Chen (2023)

4.4.2. Ejecución paralelizada

La paralelización aprovecha la arquitectura de los algoritmos para su ejecución, por lo que en la investigación este enfoque es particularmente relevante en métodos como K-Vecinos mas cercanos (KNN), donde se distribuyen tareas como el cálculo de distancias y la selección de vecinos entre varios núcleos, acelerando notablemente el proceso. Lo que permite ampliar las capacidades de investigación, como se muestra en la Figura **4-5**.

Los algoritmos como vienen por defecto en R Team (2024*b*) y Python Foundation (2024) el procesamiento es secuencial, por lo que los paquetes como Joblib Varoquaux (2024) y Dask Team (2024*a*) que brindan las herramientas para que el cálculo se efectúe en CPU por el número de núcleos, de forma paralela, donde el usuario defina cuantos va a usar, teniendo en cuenta las capacidades del hardware usado. La implementación de la paralelización no solo reduce el tiempo de procesamiento en el análisis de grandes volúmenes de datos, sino que también amplía las capacidades de investigación y desarrollo en campos que demandan alto poder computacional, superando así las barreras previas impuestas por las limitaciones de tiempo y recursos computacionales.

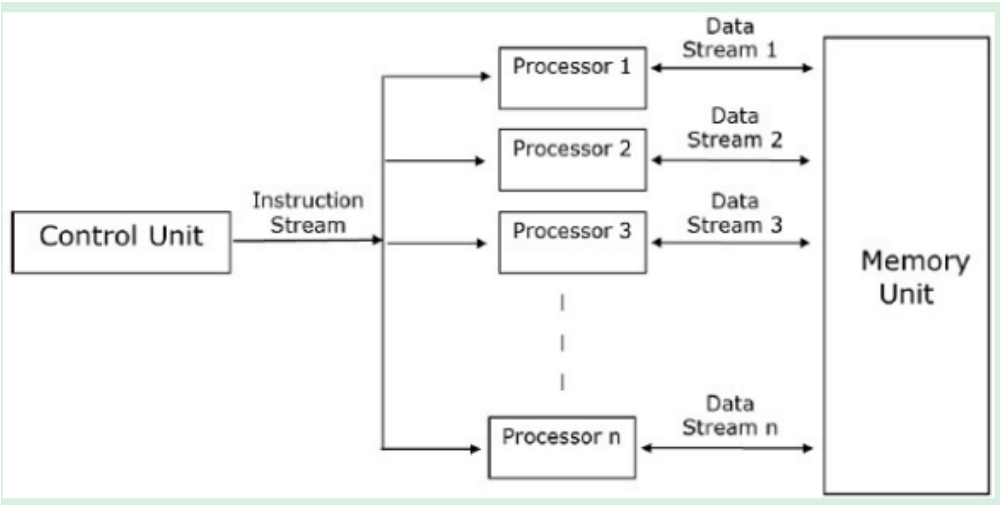

Figura 4-5.: Descripción de la paralelización - tomado de Aziz et al. (2021)

En el marco de esta investigación, donde se ha implementado la imputación de datos faltantes en la variable de precipitación, mediante el uso de algoritmos en su forma de optimización paralela, como KNN o XGBoost. Estos algoritmos se integran en el flujo de trabajo de procesamiento paralelo mediante la distribución eficiente de tareas computacionales entre los múltiples núcleos de la CPU, permitiendo la ejecución simultánea de cálculos en diferentes partes del conjunto de datos.

En el caso del KNN, se implementa dividiendo el conjunto de datos en subconjuntos que son asignados a diferentes núcleos, permitiendo que las distancias entre los puntos de datos se calculen de manera simultánea, lo que acelera significativamente el proceso de identificación de los vecinos más cercanos. En el XGBoost, que involucra un modelo de árboles de decisión optimizado, la paralelización se implementa durante la construcción de los árboles de decisión, dividiendo las tareas de búsqueda de los mejores puntos de partición y el cálculo de gradientes que se distribuyen entre los diferentes núcleos. El uso de herramientas como Joblib y Dask gestiona eficientemente estos algoritmos, favoreciéndose de la arquitectura multicore del hardware. Esta optimización no solo reduce el tiempo de ejecución, sino que también mejora la escalabilidad permitiendo trabajar con mayor volumen de información.

Los tiempos de ejecución fueron medidos utilizando la función de cronometraje integrada en Jupyter Jupyter (2024) de Visual Studio Code Corporation (2024), proporciona una evaluación del tiempo total de procesamiento, permitiendo realizar una comparación fiable de la eficiencia de computo.

4.4.3. Comparación de métodos de imputación

Debido a que cada método imputa los datos de manera diferente, se comparan para comprender cuál realiza la mejor estimación dentro de una realización.Como resultado, se crearon archivos de datos para cada método, incluyendo las estimaciones correspondientes a las regiones y el porcentaje de datos faltantes creados, y se calcularon métricas para comparar la mayoría de los métodos. La comparación de métodos se realiza para cada uno de los escenarios de datos faltantes, considerando el área geográfica de estudio. Se utilizan 4 métricas para comparar en el escenario anterior:

- RMSE - Raíz del Error cuadrático Medio.
- MAE - Error Medio Absoluto.
- MAPE - Error Porcentual Medio Absoluto.
- R^2 - Coeficiente de Determinación R cuadrado.

4.4.4. Hardware implementado

Las características con las que se ejecutó todas las metodologías, ya que es clave en cuanto al rendimiento total, son:

- CPU - Ryzen 5 - 3600.
- RAM - 16GB - 3600Hz.
- Board - Aorus - AM4.
- GPU - Nvidia RTX 3060.

5. Resultados

En este capitulo se presenta los resultados obtenidos sobre la estimación de los datos Hidrometeorológicos diarios de precipitación en el Departamento del Meta, completando los datos en el rango de tiempo entre 1980 y 2016, con un total de 56 estaciones climatológicas del IDEAM, analizando los resultados obtenidos por medio de los métodos de estimación K-Vecinos Cercanos en versiones de distancias como Euclidiana, Mahalanobis, ponderando los K vecinos, XGboost, Random Forest, teniendo en cuenta sus componentes espaciales y temporales de estaciones identificadas por su respectivo ID, cada uno de los métodos implementados serán evaluados por medio de medidas de desempeño como: Raíz del Error cuadrático, Error Medio Absoluto, Error Porcentual Medio , R^2 - Coeficiente de Determinación R cuadrado. Se realizó la comparación de los tiempos del cómputo en paralelo contra su versión secuencial original, así como efectos que tiene la computación en paralelo, El análisis se realiza por medio del software R, con su entorno Rstudio, Python 3.10.4 (Foundation 2024) con un entorno visual en Vscode (Corporation 2024), los modelos desde Jupyter Notebook (Jupyter 2024), todo el desarrollo del código fue elaborado y llevando su control de versión de manera rigurosa. se puede consultar todo el desarrolló del código en el siguiente enlace[1].

[1] https://github.com/Ronaldoh99/Proyecto_mod_precipitacion

5.1. Análisis descriptivo.

Con un total de 656,436 registros comprendidos sobre el rango de tiempo analizado se presentan 64,312 datos faltantes, lo que representa el 8.27 % , lo cual indica una sustancial fuente de información para los modelos de imputación de datos. Este porcentaje presenta variación en la cantidad de datos perdidos según la estación y periodo de tiempo observado. La tabla **5-1** presenta las medidas descriptivas del porcentaje de datos faltantes de cada estación estación meteorológica. Se observa que el rango de datos faltantes por estación es muy alto, donde el menor porcentaje de datos faltantes se presentó en la estación "Apto Vanguardia" ubicada en la ciudad de Villavicencio con la mínima del 0.04 %, dado a la cantidad de estaciones, el conjunto completo de la tabla con las 56 estaciones y su respectivo porcentaje de datos faltantes se encuentra en el anexo **A-2**.

Tabla 5-1.: Porcentaje de Datos Faltantes en Estaciones Meteorológicas

Media	Varianza	Desviación Estándar	Coeficiente de Variación	Mínimo	Máximo
7.75 %	0.39 %	6.22 %	80.27 %	0.04 %	19.63 %

Se observa que entre las estaciones en promedio tienen un porcentaje de datos faltantes menor que el global. Sin embargo, se presenta una variabilidad considerable en los porcentajes de datos faltantes entre las estaciones, donde se ve reflejado en el coeficiente de variación, contando así con estaciones que en su máxima cantidad registran el 19.63 % de datos incompletos como la estación "Ojo de Agua" del municipio de Villavicencio.

Con el fin de observar algún patrón temporal en los datos faltantes, se organizaron cronológicamente con su respectiva información de cada estación meteorológica, como se observa en la figura **5-1** Aquí se ilustra el vector de completitud de datos en una serie temporal, para la variable precipitación.

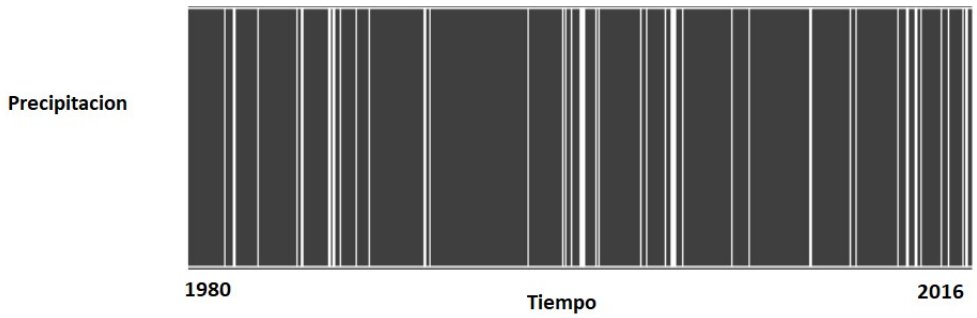

Figura 5-1.: Visualización de datos faltantes en la precipitación

Las causas de información perdida podría estar sujeta a errores o daños en la herramienta de medición, los cuales podrían persistir hasta su reparación o corrección, sin embargo no se identificó un patrón definido en los datos faltantes, tomando el registro temporal ordenados cronológicamente, también se destaca que se cuenta con el total de registros en sus componentes temporales y su respectiva información espacial.

Realizando el análisis sobre el comportamiento natural de la precipitación, se presenta la Tabla **5-2** con estadísticas descriptivas por cada estación meteorológica, donde se seleccionó aleatoriamente 10 estaciones para observar el comportamiento por estación, el registro completo se encuentra en el anexo **A-4** :

Tabla 5-2.: Estadísticas descriptivas por estaciones

ID ESTACIONES	NOMBRE	MUNICIPIO	PRECIPITACIÓN PROMEDIO	DESVIACIÓN ESTANDAR	COEFICIENTE VARIACIÓN
32020020	URIBE LA [32020020]	URIBE	10.499	18.272	174.0
32030020	RAUDAL UNO [32030020]	LA MACARENA	8.561	16.160	188.8
32035010	MACARENA LA [32035010]	LA MACARENA	7.206	14.132	1961
32060020	MESA DE YAMANES	EL CASTILLO	8.617	17.350	201.3
32060060	CALIME [32060060]	EL DORADO	13.914	24.247	174.3
32070010	CAMPO ALEGRE	VISTAHERMOSA	7.804	16.688	213.8
32070030	MICOS LOS [32070030]	SAN JUAN DE ARAMA	8.601	16.812	195.5
32070040	PINALITO [32070040]	VISTAHERMOSA	7.586	15.676	206.6
33035010	CARIMAGUA [33035010]	PUERTO GAITÁN	7.143	15.398	215.6
35035020	APTO VANGUARDIA[35035020]	VILLAVICENCIO	12.286	20.823	169.5

Los resultados de la Tabla **5-2** indican para los datos seleccionados una baja variabilidad entre las estaciones en cuanto al promedio, este comportamiento se presenta a lo largo de todas las estaciones, con una media y desviación estándar global de 9.02 y 2.89 respectivamente, variando según el comportamiento climatológico de la zona.

Los valores promedios de cada estación son muy similares entre sí, generando una consistencia en la cantidad de lluvia registrada en el departamento del Meta, el comportamiento general de las estaciones se puede observar en la Figura **5-2** presentado en una escala logarítmica, donde no se observa una gran variación entre el rango de los cuartiles de la precipitación en las estaciones. Respecto al coeficiente de variación de las estaciones, indica una dispersión significativa, el conjunto total cuenta con una oscilación entre 153 % y 216 %, lo que refleja una variabilidad notable en la precipitación a lo largo de las estaciones. Se destaca también las consistencias de los datos con la información reportada por el IDEAM, identificando los meses de mayor precipitación los cuales son entre Mayo y Junio, el listado completo de los meses y años con mayor precipitación se encuentran en el anexo **A-4**.

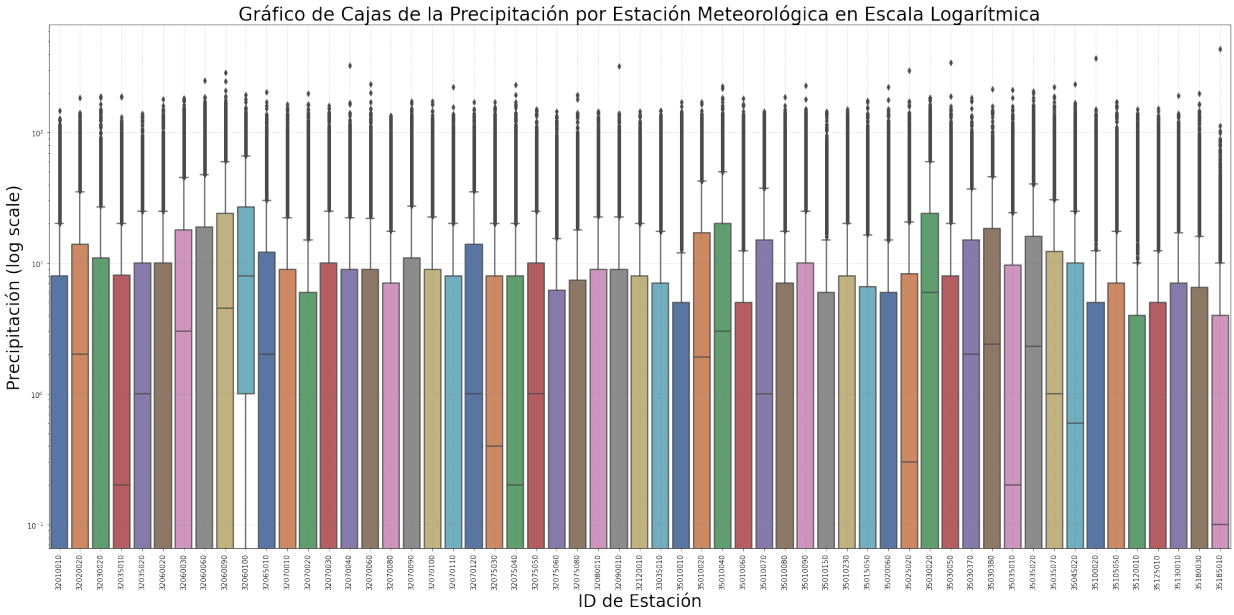

Figura 5-2.: Diagrama de cajas por estación en escala logarítmica - Precipitación

Para contrastar la información sobre la dispersión de los datos de precipitación para cada año, y con el objetivo de identificar posibles inconsistencias, como valores atípicos, se presenta la Figura **5-3**. En esta figura, que utiliza una escala logarítmica, se muestran diagramas de caja correspondientes a la precipitación anual, permitiendo observar los años en los que se registraron máximos históricos

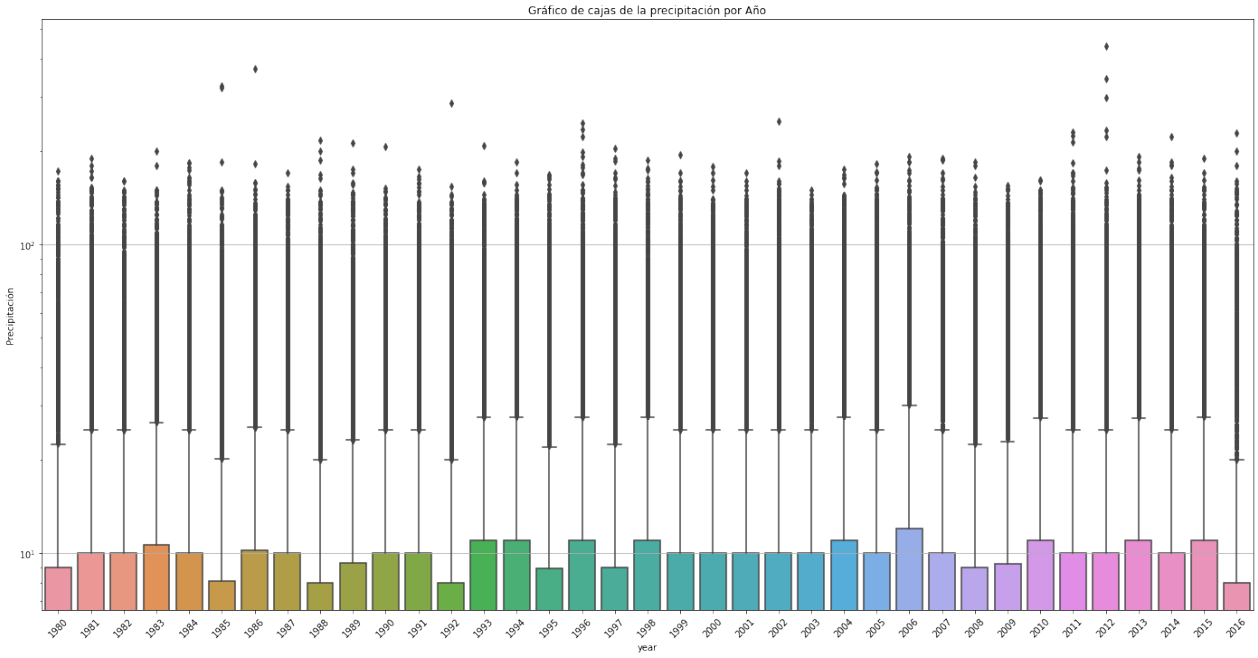

Figura 5-3.: Diagrama de cajas por año - Precipitación

En relación con la Figura **5-3**, se identifican zonas del departamento del Meta con alta precipitación, evidenciando registros que superan los 200 mm, como se observa en los valores atípicos indicados en los gráficos. Al combinar esta información con la Figura **4-2**, se puede inferir una consistencia en el comportamiento de la precipitación registrada. En el contexto de esta investigación, se subraya que la ausencia de registros en las series temporales puede afectar la precisión de las estimaciones. Según Niako (2023), una imputación adecuada de estos datos es crucial para la fiabilidad de las predicciones meteorológicas. La Figura **5-4** muestra las series de datos de precipitación correspondientes a 12 estaciones seleccionadas aleatoriamente, en las cuales se observan períodos sin registro en algunas estaciones, como San Ignacio y Mesetas.

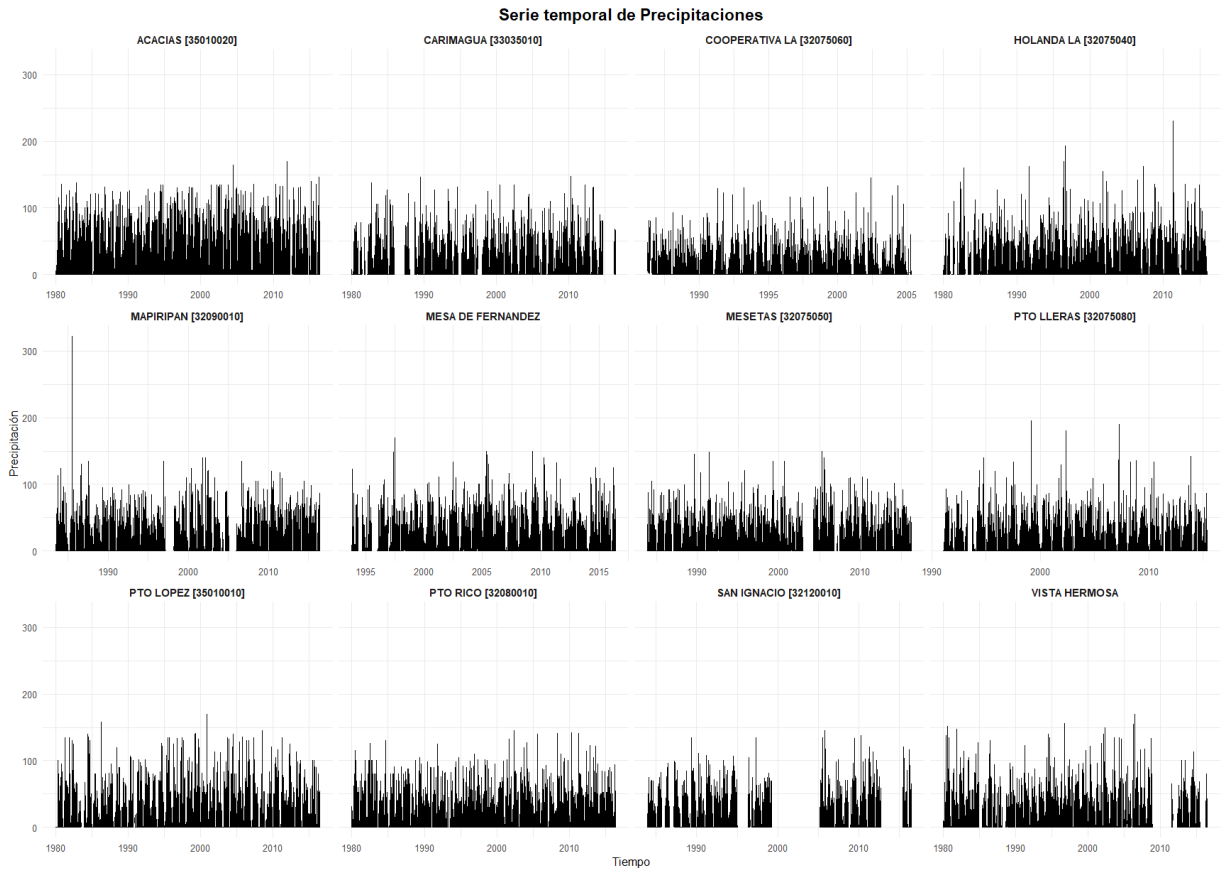

Figura 5-4.: Series de la Precipitación diaria por Estaciones

El adecuado uso de los datos disponibles por parte de los métodos de imputación es fundamental, en especial por la variación, donde la ubicación de la estación puede influir en la calidad de las estimaciones, esto por la ausencia de datos que presenten, dado que mediante el llenado de los datos, se busca mitigar las brechas en las series de datos y generar una integridad en la información.

El análisis de correlaciones entre estaciones permite observar el sentido de las relaciones y en qué medida, en la Figura **5-5** indica que todas las estaciones meteorológicas presentaron una correlación positiva, con algunas correlaciones superiores al 0.60 para ciertas estaciones, por lo que soporta el uso de modelos como el KNN, ya que permite aprovechar la información espacial y temporal disponible, en donde pueden mejorar la imputación de datos faltantes (Rachdi et al. 2021). El conjunto del total de los gráficos de correlación entre estaciones se encuentran en el Anexo **A-4** y **A-5**.

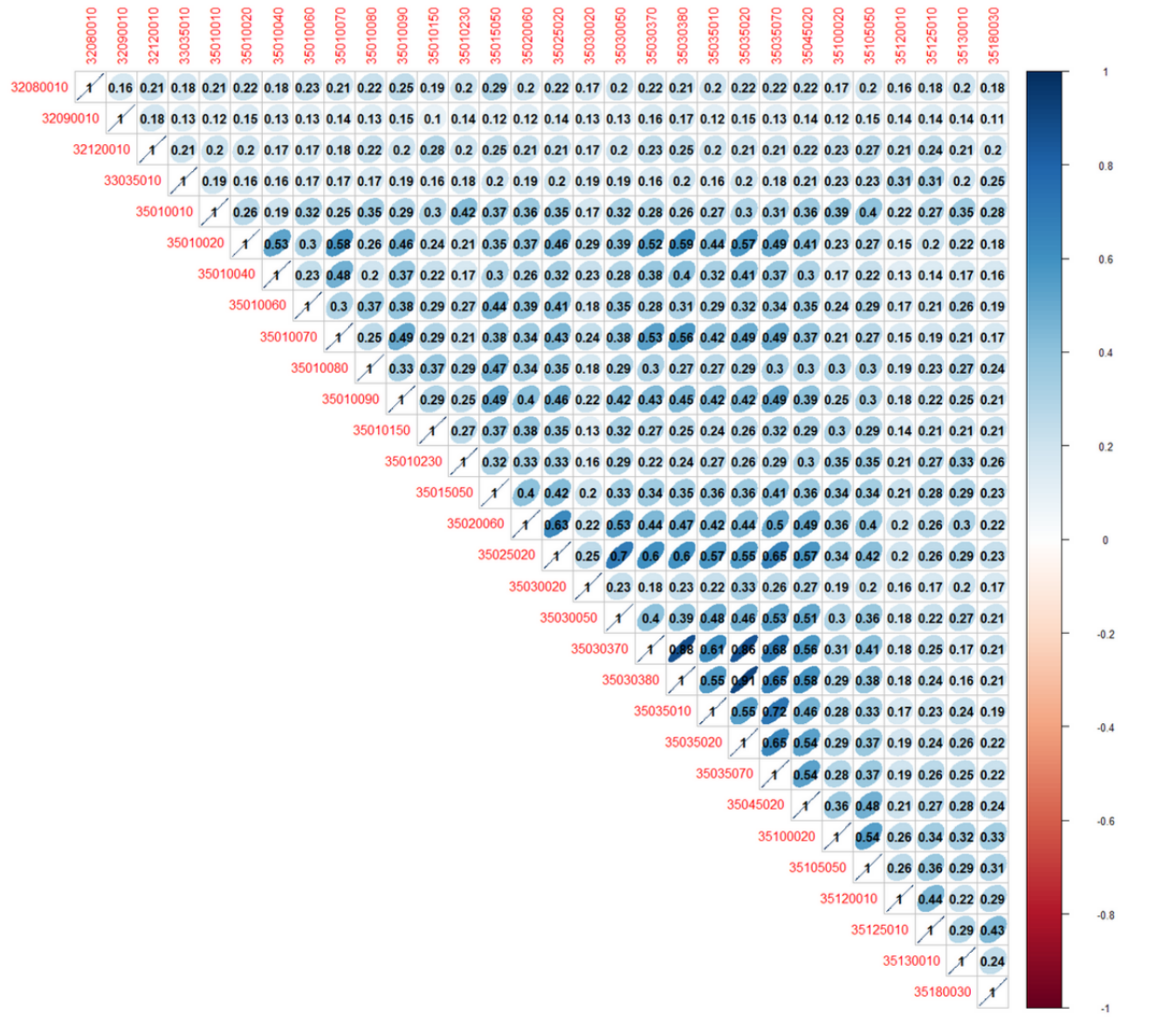

Figura 5-5.: Correlación para la precipitación entre estaciones

5.2. K-vecinos más cercanos (K Nearest Neighbors)

Para este estudio, se ha optado por un k igual a 5, basándose en que este número de estaciones cercanas resultó como uno de los que presenta un menor Error Cuadrático Medio (ECM) en los escenarios evaluados, esto por lo presentado en la figura **5-6**. Esto se realiza mediante la validación cruzada, que documenta cómo el Error Cuadrático Medio (ECM) varía al ajustar el número de estaciones cercanas incluidas en el análisis ('k'). El mismo resultado se presentó para diferentes niveles de datos faltantes, donde el ECM alcanza una estabilidad óptima cuando k se establece 5 estaciones cercanas para la estimación precisa de los datos faltantes.

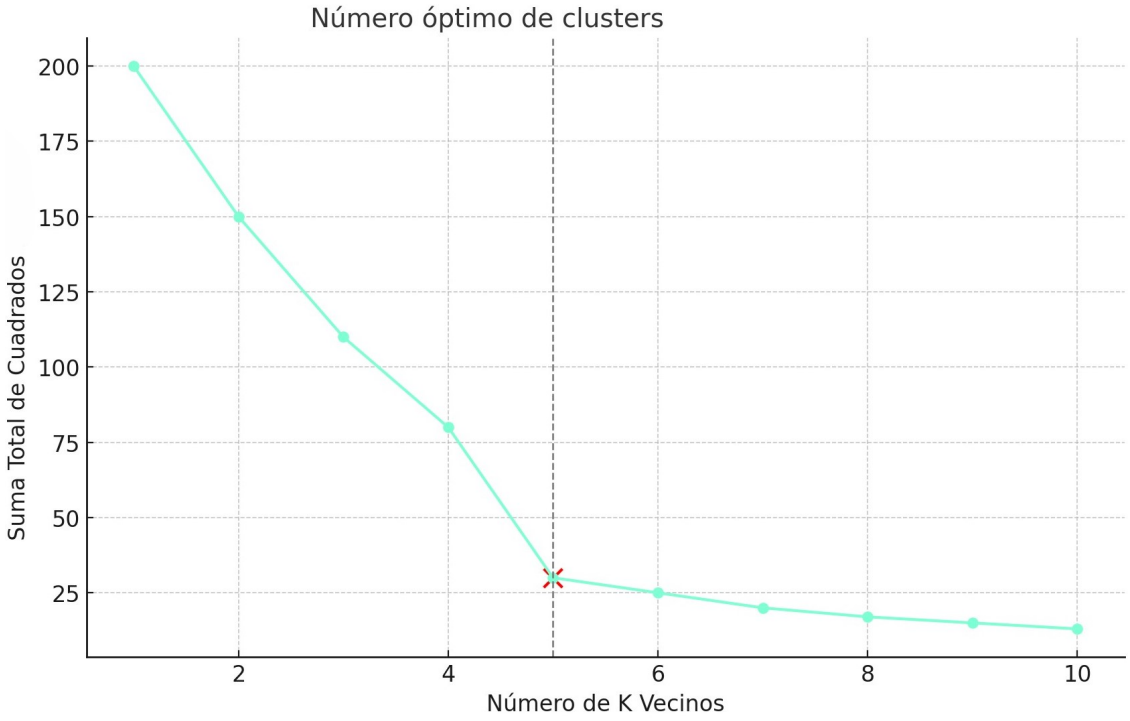

Figura 5-6.: Inflección para la Determinación Óptima K Vecinos

Los resultados de la evaluación de los diferentes tipos del método K Vecinos Más Cercanos (KNN), asumiendo diferentes ponderaciones según la distancia y el tipo de distancia utilizada, se exponen en la Tabla **5-3**. Los resultados muestran que los tres tipos de KNN presentan métricas de desempeño similares. Esta consistencia se produce incluso en los escenarios cuando aumenta el porcentaje de datos faltantes. Sin embargo, el método que utiliza la distancia de Mahalanobis, indicó ser mas eficiente en los resultados generales de las métricas, con valores poco mas bajos que sus versiones en distancia Euclidiana.

Tabla 5-3.: Métricas de rendimiento para el método KNN en cada escenario de valores ausentes

	10 %			
Métrica	**RMSE**	**MAE**	**MAPE**	**R2**
KNN Euclidiano Ponderado	5.28	1.05	25.013	0.856
KNN Euclidiano Uniforme	5.26	1.28	25.848	0.901
KNN Mahalanobis	5.17	1.06	24.91	0.9

	15 %			
Métrica	**RMSE**	**MAE**	**MAPE**	**R2**
KNN Euclidiano Ponderado	6.74	1.68	39.96	0.803
KNN Euclidiano Uniforme	6.81	1.72	40.66	0.854
KNN Mahalanobis	6.02	1.63	39.83	0.863

	20 %			
Métrica	**RMSE**	**MAE**	**MAPE**	**R2**
KNN Euclidiano Ponderado	7.43	2.31	51.93	0.803
KNN Euclidiano Uniforme	7.41	2.38	52.78	0.801
KNN Mahalanobis	7.16	2.43	51.56	0.806

En cuanto a la demanda computacional, este método se caracteriza por el elevado número de cálculos necesarios para la imputación de datos faltantes, por lo tanto, el volumen total de datos manejados, se traduce en un imponente total de 42 mil doscientos millones de cálculos, esto teniendo en cuenta el número de iteraciones que debe realizar encontrando los K vecinos con menor distancia y las características de los parámetros elegidos.

Dada la similitud en los resultados obtenidos mediante estas metodologías, el análisis del tiempo de cómputo se convierte en un clave para determinar la elección entre una u otra. Esta decisión se basa en la eficiencia computacional, por lo que se presenta en la tabla **5-4** los tiempos de procesamiento correspondientes a cada uno de los modelos evaluados.

Tabla 5-4.: Comparación de tiempos de ejecución para métodos KNN

Método -%NA\Tipo de Ejecución	**Secuencial (Por Defecto)**	**Ejecución Paralelizada**
KNN - Euclidiano Ponderado - 10 %	48.24 Minutos	4.31 Minutos
KNN - Euclidiano Uniforme - 10 %	36.46 Minutos	3.33 Minutos
KNN - Mahalanobis - 10 %	32.46 Minutos	3.65 Minutos
KNN - Euclidiano Ponderado - 15 %	59.37 Minutos	6.13 Minutos
KNN - Euclidiano Uniforme - 15 %	49.36 Minutos	5.39 Minutos
KNN - Mahalanobis - 15 %	40.85 Minutos	4.21 Minutos
KNN - Euclidiano Ponderado - 20 %	113.16 Minutos	8.78 Minutos
KNN - Euclidiano Uniforme - 20 %	107.36 Minutos	7.94 Minutos
KNN - Mahalanobis - 20 %	58.31 Minutos	5.53 Minutos

Los datos de la Tabla **5-4** reflejan una mejora sustancial en la eficiencia temporal al implementar técnicas de ejecución paralelizada en comparación con el enfoque secuencial estándar en métodos KNN. Se observa, por ejemplo, que el método KNN - Euclidiano Ponderado experimenta una reducción significativa en el tiempo de procesamiento, con una disminución de 10 veces el tiempo de ejecución con la aplicación de paralelización. Patrones similares emergen en las variantes Euclidiano Uniforme y Mahalanobis. Esta disminución en los tiempos de ejecución resalta la superioridad del procesamiento paralelo en entornos de análisis de datos a gran escala, resaltando su capacidad de optimizar la eficiencia computacional, sin sacrificar la integridad de los resultados analíticos. La distancia Euclidiana, aunque computacionalmente es menos exigente, contrasta con la complejidad que implica el uso de la matriz de covarianzas en la distancia de Mahalanobis, la cual requiere el cálculo de la matriz inversa por cada iteración para un conjunto de datos de gran tamaño.

Mientras que la ejecución secuencial del cálculo de matriz inversa de covarianzas presentó alto tiempo, la implementación de un enfoque paralelizado reduce significativamente este tiempo de cálculo. Esta disminución demuestra que en el procesamiento paralelo en operaciones computacionalmente intensivas, es una herramienta valiosa para la eficiencia en rendimiento para análisis estadísticos. (Zaharia et al. 2016).

5.3. Paralelización

La representación de la carga computacional implica un costo, esto debido a la exigencia del procesamiento en paralelo, por lo que el monitoreo de recursos de Windows 10 presentado en la figura **5-7**, indica el porcentaje de uso de cada CPU.

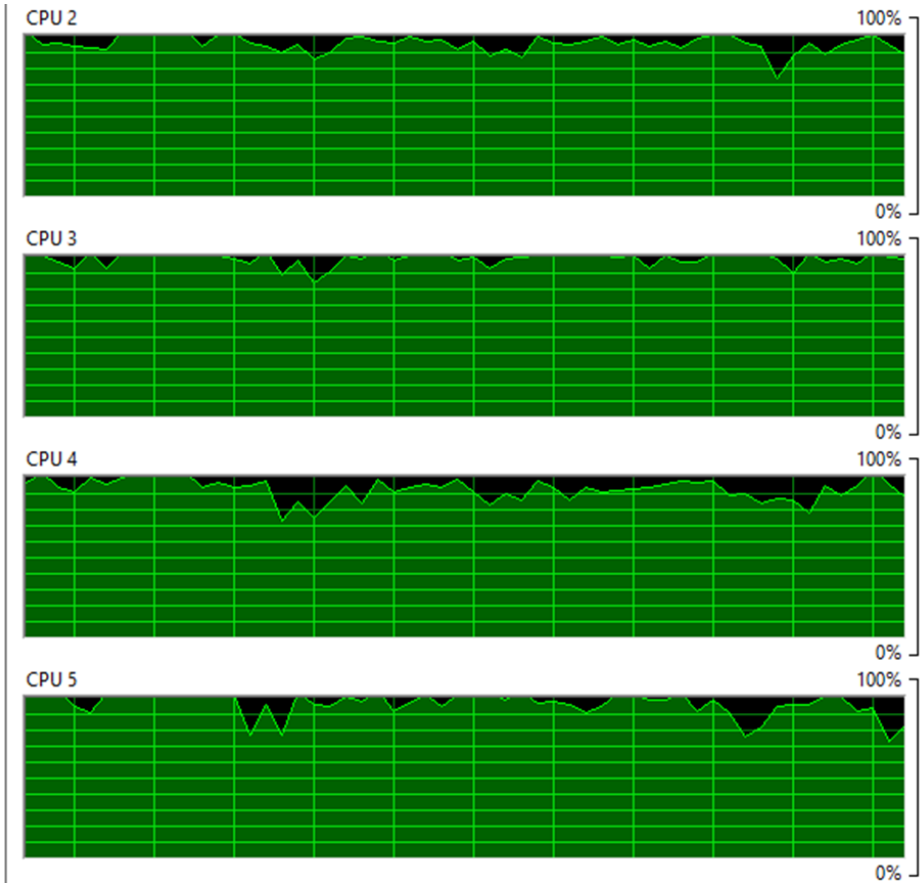

Figura 5-7.: Monitorio de Núcleos CPU

En la función para realizar la paralelización de las metodologías, se seleccionó el total de los núcleos disponibles en la CPU, lo que se ve reflejado en el alto porcentaje de uso de cada uno de los núcleos, lo que demuestra un alto rigor computacional que es ejecutado por el CPU.

Las velocidades de cómputo dependerán del hardware con el que se disponga, por lo que los tiempos y capacidad de procesado dependerán de que CPU o GPU que se este implementando, también es importante el resto de sus componentes dado que influyen en las velocidades de transmisión de información como la memoria RAM o la ROM para el almacenamiento de datos.

La paralelización de los proceso produjo una optimización al tiempo del cómputo total, además, como en los hallazgos de la investigación de Sandín Carral (2015) se alinean con los presentados en la Figura **5-8**

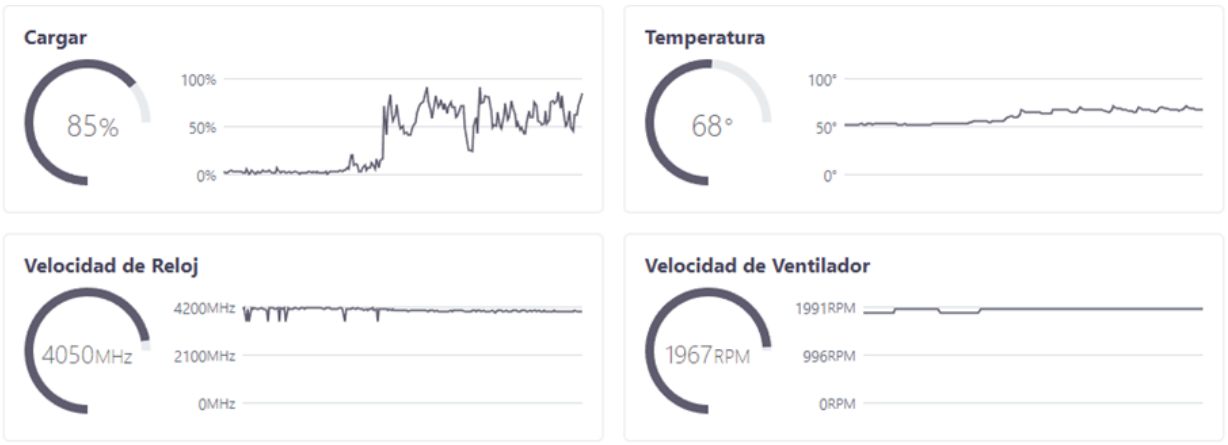

Figura 5-8.: Impacto en comportamiento CPU

Por medio de la aplicación NZXT *CAM* (2023) se evidencia un alto impacto en el consumo de recursos, ya que coloca toda su capacidad de cómputo para realizar los cálculos, esto tiene implicaciones en el consumo energético como la elevación de temperaturas experimentadas por el CPU, así como la velocidad del reloj, lo que implica mayor consumo energético en el momento de la ejecución paralelizada, a pesar de ello, es mas eficiente ya que la ejecución secuencial, aunque no exige la maquina al máximo, si tendrá que ejecutar el mismo número de cálculos, por lo que el prolongado tiempo de ejecución se traduce en mayor consumo energético, resaltando así las mejoras de la implementación de las metodologías en su ejecución en paralelo.

5.4. XGBoost y Random Forest

Antes de ejecutar el modelo XGBoost se realiza la optimización de hiperparametros mediante validación cruzada. Los hiperparámetros son configuraciones ajustables que afectan el comportamiento y rendimiento del modelo, en la tabla **5-5** se muestran los hiperparametros que tuvieron configuración, porque los demás tiene valores por defecto.

Tabla 5-5.: Parámetros Óptimos XGBOOST

HIPERPARÁMETRO	RESULTADO
MIN-SAMPLES-SPLIT	0.3
MAX DEPTH	7
n-ESTIMATORS	100
n-JOB	-1

Para obtener estos hiperparametros se utiliza la función GridSearchCV de la librería Scikit-learn learn developers (2014), que realiza una búsqueda exhaustiva en un espacio de hiperparámetros predefinido. Se define un conjunto de posibles valores para cada hiperparámetro.

Para cada combinación de hiperparámetros, el modelo XGBoost es entrenado utilizando validación cruzada. Esto significa que el conjunto de datos se divide en varias partes, en este caso no se especifica esa división, por defecto se utiliza un valor de 5, lo que significa que los datos se dividen en 5 partes y el modelo se entrena en 4 partes, mientras que se valida en la parte restante. Este proceso se repite varias veces para asegurarse de que los resultados no dependan de una sola división de los datos.

El rendimiento del modelo se evalúa en cada combinación de hiperparámetros y se guarda el conjunto de hiperparámetros que produce el mejor rendimiento en las pruebas. Una vez que todas las combinaciones han sido evaluadas, GridSearchCV selecciona la combinación que obtuvo el mejor rendimiento promedio en las pruebas de validación cruzada.

Cuando se obtiene una combinación óptima, maximiza el rendimiento del modelo en datos de validación o prueba. La elección correcta de hiperparámetros es primordial para obtener un modelo de aprendizaje automático eficaz y bien ajustado a los datos, con el fin de realizar imputaciones de datos faltantes para la variable precipitación con mayor precisión.

En el caso del modelo Random Forest, para realizar la imputación de datos faltantes de la precipitación, se usa la función llamada MICEforest (Wilson et al.2022). En este caso los hiperparámetros predeterminados se utilizan en función de los valores estándar establecidos por la librería, y no hay un proceso de búsqueda automática de los hiperparámetros más efectivos. En la tabla **5-6** se muestran los valores por defecto de los hiperparametros.

Tabla 5-6.: Parámetros Óptimos Random Forest

HIPERPARÁMETRO	RESULTADO
MIN-SAMPLES-SPLIT	2
MAX DEPTH	0
n-ESTIMATORS	500

En la tabla **5-7** se tienen las métricas de desempeño del modelo XGBoost y Random Forest para los diferentes porcentajes de datos faltantes:

Tabla 5-7.: Comparación de Métricas para XGBoost y Random Forest

Métrica	Modelo	10 %	15 %	20 %
RMSE	**XGBoost**	5.216	6.389	7.38
	Random Forest	7.4825	9.284	10.570
MAE	**XGBoost**	1.013	1.518	2.02
	Random Forest	1.278	1.933	2.533
MAPE	**XGBoost**	26.01	38.373	51.397
	Random Forest	32.708	53.573	65.075
R^2	**XGBoost**	0.914	0.8721	0.829
	Random Forest	0.8245	0.729	0.649

Las métricas demuestran que el desempeño de los modelos XGBoost y Random Forest en la imputación de la precipitación se deteriora a medida que aumenta el porcentaje de datos faltantes. Esto se refleja en el incremento de los valores de RMSE, MAE y MAPE, y la disminución del R^2, indicando una menor precisión y capacidad explicativa del modelo bajo condiciones con más datos faltantes.

De manera general, se observa que los resultados de las métricas de desempeño de los modelos indican que ambos modelos utilizados para la imputación de datos faltantes de precipitación, tienen un buen desempeño para el escenario del 10 % de datos faltantes, ya que muestran un RMSE y MAE relativamente bajos, lo que indica que los modelos están imputando los datos faltantes de manera precisa y confiable, esto significa que las diferencias entre los valores imputados y los valores reales son mínimas. El MAPE sugiere que hay un error porcentual moderado en los valores imputados, esto significa que, aunque el modelo logra una imputación

relativamente precisa, existe una variabilidad notable en los errores relativos. El coeficiente de determinación (R^2) es alto, lo que indica que es capaz de explicar una gran parte de la variabilidad en los datos de precipitación observados.

Ambos modelos, XGBoost y Random Forest, muestran una tendencia similar en su desempeño a medida que aumenta el porcentaje de datos faltantes, lo cual indica que la imputación de datos faltantes es un problema recurrente para ambos modelos. cabe destacar, que el modelo XGBoost tiende a mostrar una mayor robustez o una degradación más lenta en su desempeño a medida que aumenta el porcentaje de datos faltantes.

La Tabla **5-8** muestra los tiempos de ejecución de ambos modelos con diferentes porcentajes de datos faltantes, contrastando la ejecución secuencial con la ejecución paralelizada.

Tabla 5-8.: Comparación de Tiempos de Ejecución para XGBoost y Random Forest

Método	**XGBoost**		**Random Forest**	
% NA / Tipo Ejecución	**Secuencial**	**Paralelizada**	**Secuencial**	**Paralelizada**
10 %	2.7 minutos	8.1 segundos	3.5 minutos	10.4 segundos
15 %	3.8 minutos	15.6 segundos	4.3 minutos	17.8 segundos
20 %	5.6 minutos	27.2 segundos	7.2 minutos	31.6 segundos

La ejecución secuencial y la ejecución paralelizada tienen tiempos de ejecución significativamente diferentes para ambos modelos. Para todos los niveles de porcentaje de datos faltantes, la ejecución paralelizada generalmente muestra tiempos de ejecución más cortos que la ejecución secuencial.

Además, se puede observar que el tiempo de ejecución tanto en la ejecución secuencial como en la paralelizada aumenta con el porcentaje de datos faltantes. Esto indica que la complejidad computacional del proceso de imputación está directamente influenciada por la cantidad de datos faltantes. Estos resultados dejan evidencia de que el XGBoost tiende a ser más eficiente en términos de tiempo de ejecución que el Random Forest en todos los escenarios considerados, y que la ejecución paralelizada proporciona una mejora sustancial en los tiempos de ejecución en comparación con la ejecución secuencial para ambos modelos.

5.5. Resultados Por Estación

Los resultados anteriores mostraron que el XGboost presenta los mejores resultados. Esto concuerda con Rodríguez Leguizamón (2023) en la cual compara métodos de imputación para completar los datos de un generación de energía fotovoltaica en el cual XGboost fue el que obtuvo los mejores resultados. Es importante considerar que el rendimiento de la imputación de datos faltantes, presenta variaciones en función del desempeño específico de cada estación, lo que muestra mayor o menor eficiencia según la posición de la estación meteorológica. Por lo que es necesario ampliar las métricas conforme a la estación, por ende, los resultados de las métricas por estación se presentan como resumen en la Figura **5-9** y el conjunto completo en la tabla del Anexo **A-5**

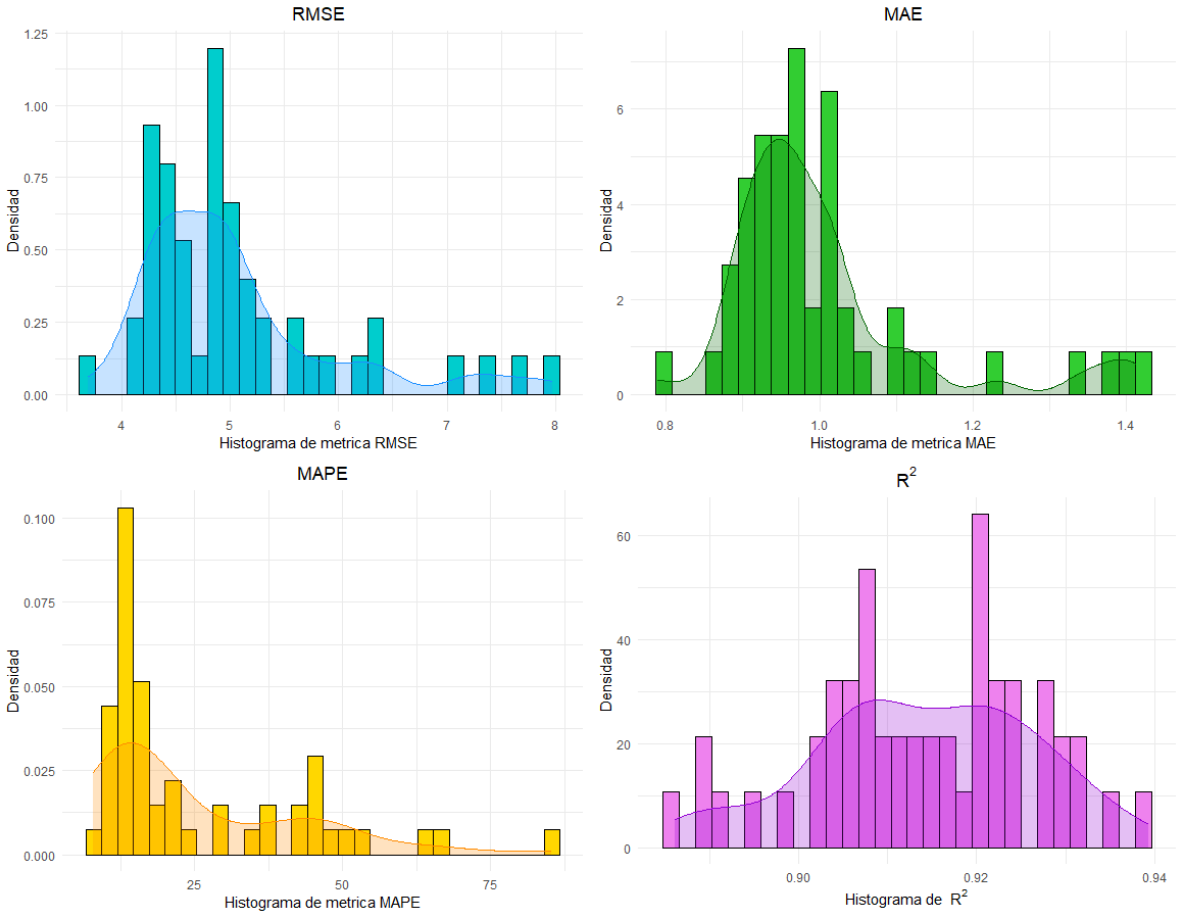

Figura 5-9.: Histograma de métricas de desempeño por estación para XGboost

Las métricas de desempeño varían por estación, esto se ve reflejado en los histogramas, ya que la distribución, muestra una asimetría hacia la izquierda, excepto para el caso de el R^2. La mayoría de las estaciones presentan métricas de desempeño similares, con algunas excepciones en las que el error fue más significativo para un número reducido de estaciones.

En el caso del R^2, la distribución no presenta tanta asimetría, concentrándose la mayor parte de las estaciones en un rango de 0.90 a 0.92, con menor número que presentaron menores valores para el indicador. La eficiencia se debe a la disponibilidad de una mayor cantidad de información en la serie temporal de la misma variable, así como los registros que provienen de las estaciones vecinas. Es importante destacar que las metodologías evaluadas tienden a perder eficiencia con el incremento en la proporción de datos faltantes. Esta observación se aliena con hallazgos de investigaciones como la de Russi Peña & Larrahondo Méndez (2017), demostrando que la escasez de información puede afectar la precisión del modelo y no depender del tipo de metodología implementada.

Para analizar el comportamiento de las estaciones, se seleccionaron de manera aleatoria un total de 10 estaciones. En la Tabla aleatoria, como se muestra en la tabla **5-9** se presentan estas estaciones, las cuales han sido sometidas previamente a un proceso de imputación de datos faltantes utilizando el modelo XGBoost, considerando un 10 % de datos faltantes. Además, en los anexos, específicamente en la Tabla **A-6**, se incluyen las estadísticas descriptivas correspondientes al total de las estaciones evaluadas.

Tabla 5-9.: Estadísticas descriptivas por estaciones

NOMBRE	Count	Mean	Std	Min	Max	Var	CV
URIBE LA [32020020]	7963	9.862	17.328	0	185	300.267	1.757
RAUDAL UNO [32030020]	10851	7.998	14.539	0	190	211.390	1.818
MACARENA LA [32035010]	12023	7.600	13.575	0	187.9	184.270	1.786
MESA DE YAMANES [32060020]	11857	11.558	19.839	0	182	393.570	1.716
CALIME [32060060]	12155	16.144	24.685	0	287	609.346	1.529
CAMPO ALEGRE [32070010]	12406	7.249	14.851	0	200	220.556	2.049
MICOS LOS [32070030]	11997	7.662	14.917	0	325	222.522	1.947
PINALITO [32070040]	12461	7.422	14.909	0	236	222.266	2.009
CARIMAGUA [33035010]	11048	13.676	22.346	0	225	499.337	1.634
APTO VANGUARDIA [35035020]	13326	6.528	14.568	0	160	212.238	2.232

Las estaciones de monitoreo muestran una considerable variabilidad en los niveles de precipitación, reflejada en las diferentes medias y desviaciones estándar. Los factores geográficos y climáticos particulares de cada ubicación pueden afectar esta variabilidad. Por ejemplo, las estaciones como CALIME [32060060] tienen una precipitación media relativamente alta y una variabilidad elevada, lo que indica la posibilidad de episodios de precipitaciones intensas, lo cual puede ser crucial para la prevención de desastres naturales y periodos de alta precipitación.

Las diferencias en la media y el coeficiente de variación (CV) entre estaciones indican diferentes patrones de precipitación. Las estaciones como APTO VANGUARDIA [35035020] tienen un CV alto, lo que indica una alta variabilidad relativa y la posibilidad de eventos

de precipitación extremos. En el ámbito de la agricultura, este conocimiento es fundamental para diseñar sistemas de riego y gestión de cultivos más eficientes que se adapten a las condiciones climáticas locales.

El conocimiento de estas estadísticas de la precipitación en cada estación, es importante para diversas aplicaciones prácticas. Por ejemplo, en las estaciones de LA MACARENA [32035010] y CAMPO ALEGRE [32070010], donde la media de precipitación es baja pero la variabilidad es alta, se podrían necesitar estrategias para el manejo de sequías intermitentes y eventos de lluvia intensa.

Estos resultados resaltan la importancia de comprender la variabilidad espacial y temporal de la precipitación en el Departamento del Meta; esto puede tener implicaciones significativas para la planificación y la gestión de recursos hídricos, así como para la predicción de eventos climáticos extremos.

6. Conclusiones, limitaciones y recomendaciones

6.1. Conclusiones

En conclusión, el presente trabajo ha enfatizado la importancia y eficacia de los modelos de imputación de datos faltantes, centrándose en la imputación de la variable de precipitación para el Departamento del Meta, Colombia. Se utilizaron métricas fundamentales como MAE, RMSE, R^2 y MAPE. Se evaluaron y compararon varios modelos de aprendizaje automático, incluidos KNN en diferentes versiones, XGBoost y Random Forest.

Entre los modelos evaluados, XGBoost emerge como el modelo más eficiente en términos de rendimiento tanto estadístico como computacional. Su capacidad para manejar grandes conjuntos de datos y su rendimiento superior en términos de métricas de evaluación lo convierten en una opción destacada para la imputación de datos faltantes en la precipitación diaria. El método de Random Forest, presentó el menor desempeño en la estimación de los datos faltantes, según las métricas de desempeño utilizadas, por otro lado presentó tiempos bajos de ejecución comparables con el del XGBoost.

En cuanto al método de K-Vecinos mas Cercanos, se consideraron diferentes modificaciones que eran sugeridas por algunos autores como Mayr Ojeda & Patrone Martirena (2016), sin embargo no demostró tener diferencias significativas entre si, en factores como las métricas de desempeño, por otro lado en la medición de los tiempos, se detectó mejoras en los tiempos de espera por los modelos, al ser de un alto rigor computacional, la paralelización aporta poder para realizar mayores comparaciones de versiones de los modelos o trabajar con mayor volumen de datos.

Además, se examinó la importancia de la paralelización de procesos con herramientas como Dask y Joblib de Python, para acelerar el tiempo de ejecución de los modelos. La paralelización es esencial para trabajar con grandes conjuntos de datos, porque demostró ser crucial para mejorar significativamente la eficiencia computacional y reducir los tiempos de procesamiento. Aunque estas herramientas mejoran considerablemente el rendimiento, es importante tener en cuenta que el rendimiento computacional puede variar según el hardware y la implementación específica de los modelos.

La selección del mejor modelo de imputación de datos faltantes en el contexto de la predicción de la precipitación, es crucial para diversos aspectos ambientales, agrícolas y sociales. Este proceso sugiere que la precisión de las predicciones, se pueden acompañar con una comprensión detallada de las características específicas del conjunto de datos y las necesidades prácticas relacionadas con su relevancia.

Los hallazgos de esta investigación demuestran la importancia de mantener la calidad y la integridad de los datos para aplicaciones que dependen de la imputación de datos faltantes. En el contexto de la predicción de la precipitación, garantizar un bajo porcentaje de datos faltantes es esencial para mantener la precisión y la fiabilidad de los modelos predictivos.

6.2. Limitaciones

Una limitación se presenta en la falta de información sobre las causas de los datos faltantes en la red de estaciones de monitoreo del IDEAM. Esto conlleva a la adopción de un enfoque de datos faltantes completamente aleatorios (MCAR), donde se presume que la ausencia de datos es independiente de cualquier variable observada o no observada. Sin embargo, no se cuenta con la capacidad para determinar con certeza las razones detrás de los datos faltantes.

6.3. Recomendaciones

Para mejorar la velocidad y eficiencia en el procesamiento, se sugiere aprovechar la paralelización mediante GPU o una CPU de mayor potencia, como lo indica Sandín Carral (2015). Estas mejoras pueden incrementar directamente las velocidades de computo y la capacidad de análisis. Además, complementar con variables adicionales al estudio, como corrientes de aire, humedad relativa y datos satelitales, entre otras, por lo que proporcionan un contexto más completo, permitiendo una mejor captura de las interdependencias, variaciones espaciales y temporales, como lo mencionan Pielke Sr (2013) y Sánchez Quiroga (2020). Esta integración podría aumentar la capacidad de los modelos para imputar datos faltantes.

El conjunto de datos resultante de este trabajo se presenta como un recurso para futuros estudios sobre la precipitación en el Departamento del Meta. Al contar con registros completos, permiten realizar investigación y profundizar en el entendimiento de este fenómeno hidrometeoro. Dado el volumen de datos resultante, se podría implementar metodologías mas robustas de análisis como datos funcionales.

Realizar estudios comparativos considerando otros mecanismos de datos faltantes, como MAR (Missing At Random) y MNAR (Missing Not At Random), para determinar el impacto que tienen los modelos de imputación (KNN, XGBoost y Random Forest) en la eficiencia y precisión computacional. Esto mejorará la comprensión de cómo varios mecanismos de datos faltantes que pueden afectar la eficacia de los métodos de imputación.

Bibliografía

Agnihotri, A., Sahoo, J. R. P. & Diwakar, M. (2022), *Flood Prediction Using Hybrid ANFIS-ACO Model: A Case Study*, pp. 169–180.

Alperin, M. (2013), 'Introducción al análisis estadístico de datos geológicos', *Series: Libros de Cátedra* .

Aziz, A., Abdulqader, N., Sallow, A. B. & Omer, K. K. (2021), 'Python parallel processing and multiprocessing: A review', *Academic Journal of Nawroz University* **10**(3), 345–354.

Barney, B. (2023), 'Introduction to parallel computing', [Online]. Available: https://computing.llnl.gov/tutorials/parallelcomp/#Whatis.

Bello, A. M., C. J. A. . G. E. K. (2019), 'Técnicas de imputación para datos de precipitación máxima mensual en la zona central de boyacá', *Revista Ingeniería Investigación y Desarrollo* **19(1)**, 64–79.

Breiman, L. (2001), 'Random forests', *Machine learning* **45**, 5–32.

CAM (2023), https://nzxt.com/es-ES/software/cam. Software para monitoreo de CPU.

Campozano, L., Sánchez, E., Avilés, Á. & Samaniego, E. (2014), 'Evaluation of infilling methods for time series of daily precipitation and temperature: The case of the ecuadorian andes', *Maskana* **5**(1), 99–115.

Chen, T. & Guestrin, C. (2016*a*), Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining', pp. 785–794.

Chen, T. & Guestrin, C. (2016*b*), Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '16, ACM.
URL: *http://dx.doi.org/10.1145/2939672.2939785*

Corporation, M. (2024), 'Visual studio code: Code editing. redefined', https://code.visualstudio.com/. Accesado: 29-07-2024.

Cover, T. & Hart, P. (1967), 'Nearest neighbor pattern classification', *IEEE Transactions on Information Theory* **13**, 21–27.

Cutler, A., Cutler, D. R. & Stevens, J. R. (2012), Random forests, *in* 'Ensemble Machine Learning', Springer US, pp. 157–175.
URL: *http://link.springer.com/10.1007/978-1-4419-9326-7*$_5$

DANE (2005), 'Ficha técnica sistema de información del medio ambiente', *Departamento Administrativo Nacional de Estadística (DANE)* .
URL: *https://www.dane.gov.co/index.php/estadisticas-por-tema/ambientales/indicadores-ambientales-iaii/iniciativa-sima*

Dempster, A. P.and Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the royal statistical society* **39**(1), 1–22.
URL: *https://doi.org/10.1111/j.2517-6161.1977.tb01600.x*

Dumedah, G. & Coulibaly, P. (2011), 'Evaluation of statistical methods for infilling missing values in high-resolution soil moisture data', *Journal of Hydrology* **400(1)**, 95–102.

Ferrari, G. T. & Ozaki, V. (2014), 'Missing data imputation of climate datasets: Implications to modeling extreme drought events', *Revista Brasileira de Meteorologia* **29**, 21–28.

Figueras, S. M. & Gargallo, P. (2003), 'Análisis exploratorio de datos', http://www.5campus.com/leccion/aed. Consultado el 28 de julio de 2017.

Fontecha Bello, M., Cuta, J. A. & García, E. K. (2020), 'Técnicas de imputación para datos de precipitación máxima mensual en la zona central de boyacá', *Ingeniería Investigación y Desarrollo* **19(1)**, 64–79.

Foundation, P. S. (2024), 'Python: A powerful programming language', https://www.python.org/. Accesado: 29-07-2024.

García, R. P. L. (2015), 'Imputación de datos en series de precipitación diaria caso de estudio cuenca del río quindío', *Ingeniare* **10**(18), 73–86.

González, I., Rodríguez, A. & Sastoque, J. I. (2020), 'La incidencia de la precipitación y el Área sembrada frente a la producción de arroz en el departamento del meta', *Revista Boletín El Conuco: investigación, economía y sociedad* **3**, 1–15.

Grama, A., Gupta, A., Karypis, G. & Kumar, V. (2003), *Introduction to Parallel Computing, Second Edition*, Addison Wesley.

Gómez, J., Palarea, A. & Matín, F. J. (2006), 'Métodos de inferencia estadística con datos faltantes. estudio de simulacion sobre los efectos en las estimaciones', *Estadística Española* **48**(162), 241–270.

Hastie, T., Tibshirani, R., Narasimhan, B. & Chu, G. (2001), 'impute: Imputation for microarray data', *Bioinformatics* **17**(6), 520–525.

Hurtado, G., González, O. C. & Montaña, J. A. (2001), 'Aspectos departamentales. atlas climatológico de colombia', *Bogotá, Instituto de Hidrología, Meteorología y Estudios Ambientales de Colombia IDEAM* .

Ibarra-Berastegi, G., Saénz, J., Ezcurra, A., Elías, A., Diaz Argandoña, J. & Errasti, I. (2011), 'Downscaling of surface moisture flux and precipitation in the ebro valley (spain) using analogues and analogues followed by random forests and multiple linear regression', *Hydrology and Earth System Science* **15**, 1895–1907.

IDEAM (2023), 'Instituto de hidrología, meteorología y estudios ambientales. atlas climatológico de colombia - IDEAM'.
URL: *http://www.ideam.gov.co/web/tiempo-y-clima/tiempo-clima*

Instituto de Hidrología, M. y. E. A. I. (2023), 'Fenómenos el niño-la niña'.
URL: *http://atlas.ideam.gov.co*

Jacob Junior, A. F. L., do Carmo, F. A., de Santana, A. L., Santana, E. E. C. & Lobato, F. M. F. (2024), 'Evoimp: Multiple imputation of multi-label classification data with a genetic algorithm', *Plos one* **19**(1), e0297147.

Jupyter, P. (2024), 'Project jupyter: Open tools for interactive computing', https://jupyter.org/. Accesado: 29-07-2024.

Kim, S. & Kim, H. (2016), 'A new metric of absolute percentage error for intermittent demand forecasts', *International Journal of Forecasting* **32**(3), 669–679.
URL: *https://www.sciencedirect.com/science/article/pii/S0169207016000121*

learn developers, S. (2014), 'Grid search — scikit-learn 0.15 documentation', https://scikit-learn.org/0.15/modules/grid_search.html#grid-search. Accesed: 2024-08-10.

Little, R. J. & Rubin, D. B. (2019), *Statistical analysis with missing data*, Vol. 793, John Wiley & Sons.

Little, R. & Rubin, D. (1987), *Statistical Analysis With Missing Data*, Wiley Series in Probability and Statistics, Wiley.
URL: *https://books.google.com.co/books?id=w40QAQAAIAAJ*

Lorena, U. H. A. (2022), 'Aceleración de algoritmos de clasificación basados en disimilitudes, utilizando arquitecturas computacionales con múltiples núcleos.'.
URL: *https://repositorio.unal.edu.co/handle/unal/81347*

Mayr Ojeda, F. & Patrone Martirena, F. (2016), 'Mejora de la eficiencia de knn utilizando programación paralela en f'.

McKnight, P., McKnight, K., Sidani, S. & Figueredo, A. (2007), *Missing Data: A Gentle Introduction*, Methodology in the Social Sciences Series, Guilford Publications.
URL: *https://books.google.com.co/books?id=Oel21pwDWXQC*

Medina Rivera, R. D. (2008), Estimación estadística de valores faltantes en series históricas de lluvia, Master's thesis.
URL: *https://hdl.handle.net/11059/1126*

Membreño, A. M. & Zavala, A. d. J. R. (2015), 'Aplicación de técnicas estadísticas y geoestadísticas para elaborar cartografía de precipitaciones. departamentos del occidente de nicaragua', *Ciencias Espaciales* **8**(1), 143–159.

Montealegre, J. (2009), 'Estudio de la variabilidad climática de la precipitación en colombia asociada a procesos oceánicos y atmosféricos de meso y gran escala'.
URL: *http://institucional. ideam. gov. co/jsp/812*

Morales España, G., Mora Flórez, J. & Vargas Torres, H. (2008), 'Estrategia de regresión basada en el método de los k vecinos más cercanos para la estimación de la distancia de falla en sistemas radiales', *Revista Facultad de Ingeniería Universidad de Antioquia* (45), 100–108.

Moreno, F. (2004), Clasificadores eficaces basados en algoritmos rápidos de búsqueda del vecino más cercano, PhD thesis, Universidad de Alicante. Departamento de lenguajes y sistemas informáticos.

Niako, N. (2023), Effects of Missing Data Imputation Methods on Univariate Time Series Forecasting with Arima and LSTM, PhD thesis, The University of Texas Rio Grande Valley.

Ortega, J. H. G. & Rodriguez, L. M. S. (2011), 'Aplicación de tecnologías de información geográfica para el estudio de la variabilidad climática en la cuenca alta del río cauca.'.
URL: *http://www.observatoriogeograficoamericalatina.org.mx/egal12/Nuevastecnologias/Cartografiaautom*

Pang-Ning Tan, M. S. et al. (2006*a*), 'Introduction to data mining'. Page 69.

Pang-Ning Tan, M. S. et al. (2006*b*), 'Introduction to data mining'. Page 81.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.

Pielke Sr, R. A. (2013), *Mesoscale meteorological modeling*, Academic press.

Prácticas Hidrológicas, G. (1994), 'Organización meteorológica mundial-n 168, 5ta ed., 1994. 781 p'.
URL: *https://www.academia.edu/14015021/Guia$_d$e$_P$racticas$_H$idrogeologicas$_O$MM*

Rachdi, M., Laksaci, A., Kaid, Z., Benchiha, A. & Al-Awadhi, F. A. (2021), 'k-nearest neighbors local linear regression for functional and missing data at random', *Statistica Neerlandica* **75**(1), 42–65.

Reinoso, P. L. G. (2015), 'Imputación de datos en series de precipitación diaria caso de estudio cuenca del río quindío', *Ingeniare* **18**, 73–86.
URL: *https://dialnet.unirioja.es/servlet/articulo?codigo=5478781*

Rodríguez Leguizamón, C. (2023), Pronóstico de la generación de energía fotovoltaica del campus de la Universidad Autónoma de Occidente mediante modelo basado en inteligencia artificial, PhD thesis, Universidad Autónoma de Occidente.

RStudio Team (2012), *RStudio: Integrated Development Environment for R*, RStudio, PBC, Boston, MA.
URL: *http://www.rstudio.com/*

Rubin, D. B. (1976), 'Inference and missing data', *Biometrika* **63**(3), 581–592.
URL: *https://doi.org/10.1093/biomet/63.3.581*

Russi Peña, J. & Larrahondo Méndez, E. A. (2017), 'Comparación de métodos de estimación de datos faltantes en series de precipitación diaria en el valle del cauca.'.

Sabeena, J. & Reddy, P. V. S. (2017), 'A review of weather forecasting schemes', *i-manager's Journal on Pattern Recognition* **4**(2), 27.

Sahoo, A., Samantaray, S. & Ghose, D. (2021*a*), 'Prediction of flood in barak river using hybrid machine learning approaches: A case study', *Journal of the Geological Society of India* **97**, 186–198.
URL: *https://doi.org/10.1007/s12594-021-1650-1*

Sahoo, A., Samantaray, S. & Ghose, D. (2021*b*), 'Prediction of flood in barak river using hybrid machine learning approaches: a case study', *Journal of the Geological Society of India* **97**(2), 186–198.

Sahoo, A., Samantaray, S. & Paul, S. (2021), 'Efficacy of anfis-goa technique in flood prediction: a case study of mahanadi river basin in india', *H2Open Journal* **4**(1), 137–156.

Sahoo & Dillip Kumar, G. (2022), 'Imputation of missing precipitation data using knn, som, rf, and fnn'.

Samantaray, S., Sahoo, A. & Agnihotri, A. (2021), 'Assessment of flood frequency using statistical and hybrid neural network method: Mahanadi river basin, india', *Journal of the Geological Society of India* **97**, 867–880.
URL: *https://doi.org/10.1007/s12594-021-1785-0*

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, PhD thesis, Universidad Santo Tomás.

Sandín Carral, G. A. (2015), 'Ocr: Análisis e implementación de algoritmos en nuevas tecnologías de paralelización'.

Sandín Carral, G. A. (2015), 'Análisis e implementación de algoritmos en nuevas tecnologías de paralelización'.
URL: *http://hdl.handle.net/2445/67785*

Shao, L. & Chen, W. (2023), 'Coal and gas outburst prediction model based on miceforest filling and phho–kelm', *Processes* **11**, 2722.

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'.
URL: *http://unidadinvestigacion.usta.edu.co*

Team, D. D. (2024*a*), 'Dask: Parallel computing with task scheduling', https://www.dask.org/. Accesado: 29-07-2024.

Team, R. C. (2024*b*), 'R: A language and environment for statistical computing', https://www.r-project.org/. Accesado: 29-07-2024.

The Weather Channel (2017), 'Cómo se mide la lluvia', https://weather.com/es-ES/espana/tiempo/news/como-se-mide-la-lluvia-14102017.

TodaColombia.com (2023*a*), 'Departamento del meta - todacolombia.com', https://www.todacolombia.com/departamentos-de-colombia/meta/index.html.

TodaColombia.com (2023*b*), 'Departamento del meta - todacolombia.com', https://www.todacolombia.com/departamentos-de-colombia/meta/clima.html.

Varoquaux, G. (2024), 'Joblib: Efficient compression and caching for python', https://pypi.org/project/joblib/#description. Accesado: 29-07-2024.

Verbeke, G. & Molenberghs, G. (2009), *Linear Mixed Models for Longitudinal Data*, Springer Series in Statistics, Springer New York.
URL: *https://books.google.com.co/books?id=jmPkX4VU7h0C*

Vicente-Serrano, S. M., Beguería, S., López-Moreno, J. I., García-Vera, M. A. & Stepanek, P. (2010), 'A complete daily precipitation database for northeast spain: Reconstruction, quality control, and homogeneity', *International Journal of Climatology* **30**(8), 1146–1163.

White, M. J. (2024), 'miceforest: Fast, memory efficient imputation with lightgbm', https://pypi.org/project/miceforest/. Accesado: 29-07-2024.

Willmott, C., Matsuura, K. & Robeson, S. (2009), 'Ambiguities inherent in sums-of-squares-based error statistics', *Atmospheric Environment* **43**, 749–752.

Wilson, S. V., Cebere, B., Myatt, J. & Wilson, S. (2022), 'AnotherSamWilson/miceforest: Release for Zenodo DOI', https://doi.org/10.5281/zenodo.7428632.

Yang, J.-H., Cheng, C.-H. & Chan, C.-P. (2017), 'A time-series water level forecasting model based on imputation and variable selection method', *Computational Intelligence and Neuroscience* **2017**, Article ID 8734214.
URL: *https://doi.org/10.1155/2017/8734214*

Yozgatligil, C., Aslan, S., Iyigun, C. & Batmaz, I. (2013), 'Comparison of missing value imputation methods in time series: the case of turkish meteorological data', *Theoretical and applied climatology* **112**, 143–167.

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M. J. et al. (2016), 'Apache spark: A unified engine for big data processing', *Communications of the ACM* **59**(11), 56–65.

A. Anexos

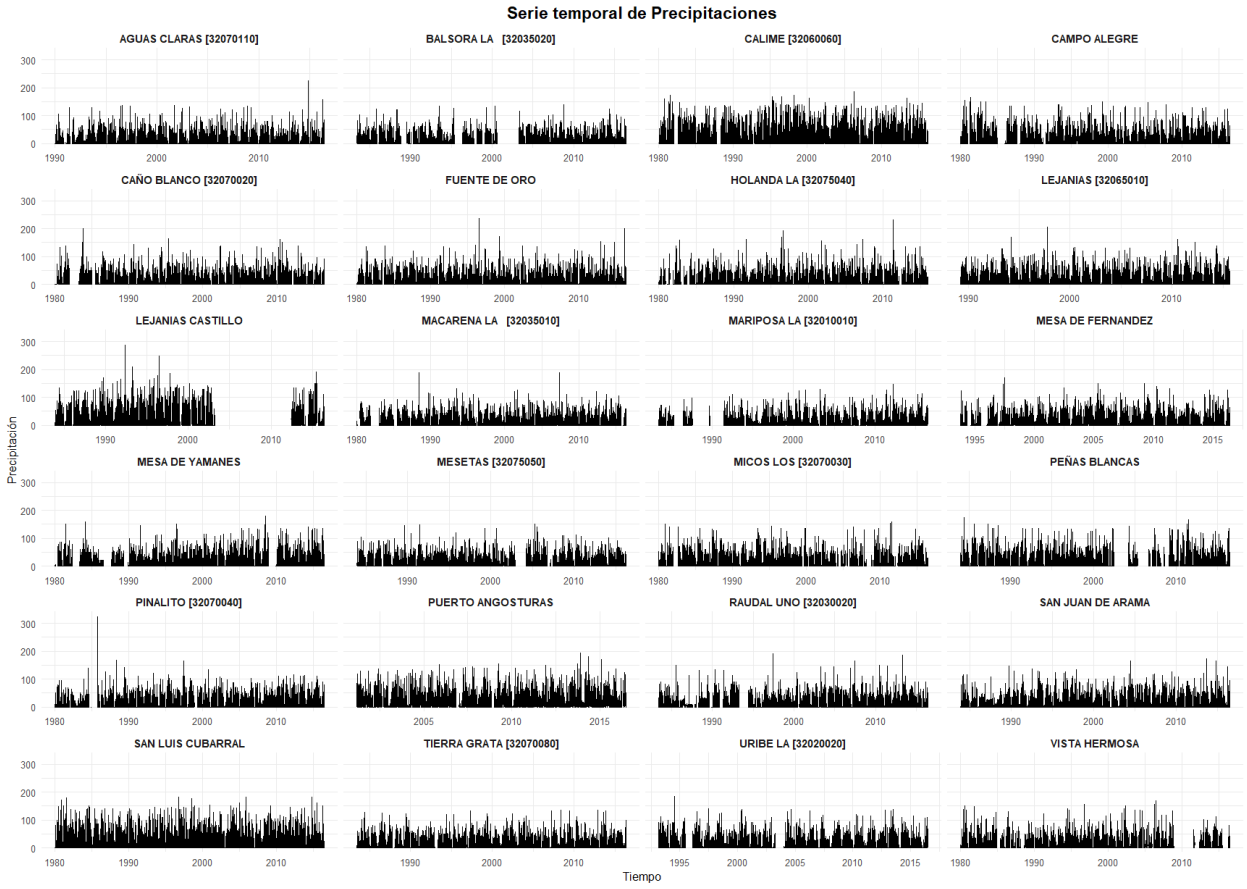

Figura A-1.: Dispersión de la Precipitación por estaciones 18

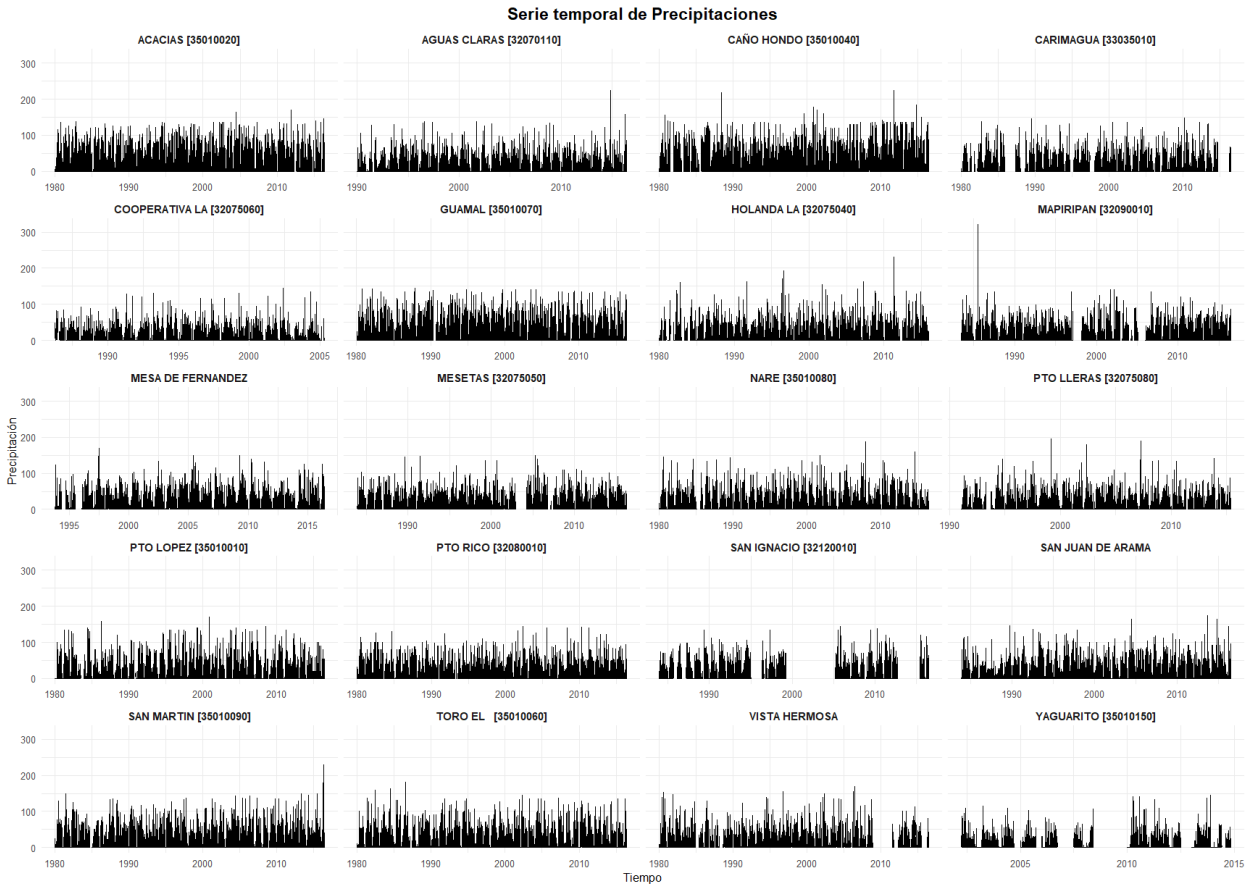

Figura A-2.: Dispersión de la Precipitación por estaciones 38

Tabla A-1.: Tabla de estaciones con ID

ID	Nombre	Municipio	lat	lon
32010010	APTO VANGUARDIA	VILLAVICENCIO	4,161919444	-73,61757778
32020010	LEJANIAS [32065010]	LEJANÍAS	3,533333333	-74,01666667
32020020	SAN JUAN DE ARAMA	SAN JUAN DE ARAMA	3,375083333	-73,89194444
32030010	SAN LUIS CUBARRAL	CUBARRAL	3,790722222	-73,84277778
32030020	POMPEYA [35020060]	VILLAVICENCIO	4,03975	-73,368
32035010	FUENTE DE ORO	FUENTE DE ORO	3,466666667	-73,63333333
32035020	CABUYARO [35100020]	CABUYARO	4,283722222	-72,79283333
32060020	TORO EL [35010060]	CASTILLA LA NUEVA	3,786666667	-73,40055556
32060030	SAN MARTIN [35010090]	SAN MARTÍN	3,750277778	-73,70083333
32060050	ACACIAS [35010020]	ACACÍAS	3,994638889	-73,76558333
32060060	NARE [35010080]	PUERTO LÓPEZ	3,793888889	-73,14972222
32060090	AGUAS CLARAS [32070110]	GRANADA	3,472027778	-73,85783333
32060100	COOPERATIVA LA [32075060]	FUENTE DE ORO	3,366666667	-73,7
32065010	PTO RICO [32080010]	PUERTO RICO	2,943333333	-73,20972222
32070010	CABAÑA LA HDA	CUMARAL	4,300444444	-73,3575
32070020	TIERRA GRATA [32070080]	PUERTO LLERAS	3,0875	-73,22638889
32070030	GUAMAL [35010070]	GUAMAL	3,878333333	-73,75861111
32070040	PTO LOPEZ [35010010]	PUERTO LÓPEZ	4,105027778	-72,9365
32070050	PTO LLERAS [32075080]	PUERTO LLERAS	3,267694444	-73,37319444
32070060	PUERTO ANGOSTURAS	CUBARRAL	3,792666667	-73,92247222
32070080	CAÑO HONDO [35010040]	GUAMAL	3,924222222	-73,81458333
32070090	MESA DE FERNANDEZ	SAN JUAN DE ARAMA	3,456555556	-74,02994444
32070100	PTO GAITAN [35120010]	PUERTO GAITÁN	4,311388889	-72,07663889
32070110	GUAICARAMO [35105050]	BARRANCA DE UPÍA	4,468666667	-72,95358333
32070120	FUNDO NUEVO HUMAPO	PUERTO LÓPEZ	4,327722222	-72,39144444
32075020	CAÑO BLANCO [32070020]	FUENTE DE ORO	3,25	-73,51666667
32075030	PINALITO [32070040]	VISTAHERMOSA	2,983333333	-73,63333333
32075040	URIBE LA [32020020]	URIBE	3,243222222	-74,35372222
32075050	PLATA LA [35130010]	PUERTO LÓPEZ	3,956083333	-72,76572222
32075060	CAMPO ALEGRE	VISTAHERMOSA	3,2	-73,75
32075070	LIBERTAD LA [35025020]	VILLAVICENCIO	4,057361111	-73,46791667
32075080	BARBASCAL [35015050]	SAN MARTÍN	3,630777778	-73,34927778
32080010	OJO DE AGUA [35030050]	VILLAVICENCIO	4,091138889	-73,44877778
32080020	HOLANDA LA [32075040]	GRANADA	3,516333333	-73,71602778
32080030	CALIME [32060060]	EL DORADO	3,742333333	-73,83469444
32080040	MESETAS [32075050]	MESETAS	3,380055556	-74,04297222
32090010	MAPIRIPAN [32090010]	MAPIRIPÁN	2,889722222	-72,13011111
32095010	MACARENA LA [32035010]	LA MACARENA	2,176111111	-73,79333333
32120010	RAUDAL UNO [32030020]	LA MACARENA	2,3225	-73,94388889
33035010	MICOS LOS [32070030]	SAN JUAN DE ARAMA	3,316388889	-73,89216667
35010010	MONFORT [35030020]	EL CALVARIO	4,310083333	-73,64802778
35010020	MESA DE YAMANES	EL CASTILLO	3,530944444	-73,90108333
35010040	PEÑAS BLANCAS	SAN JUAN DE ARAMA	3,316666667	-73,91666667
35010050	CARIMAGUA [33035010]	PUERTO GAITÁN	4,574416667	-71,34075
35010060	IDEAM V/CIO [35030380]	VILLAVICENCIO	4,144083333	-73,64183333
35010070	MARIPOSA LA [32010010]	URIBE	2,562833333	-74,10308333
35010080	BALSORA LA [32035020]	LA MACARENA	2,438388889	-73,93227778

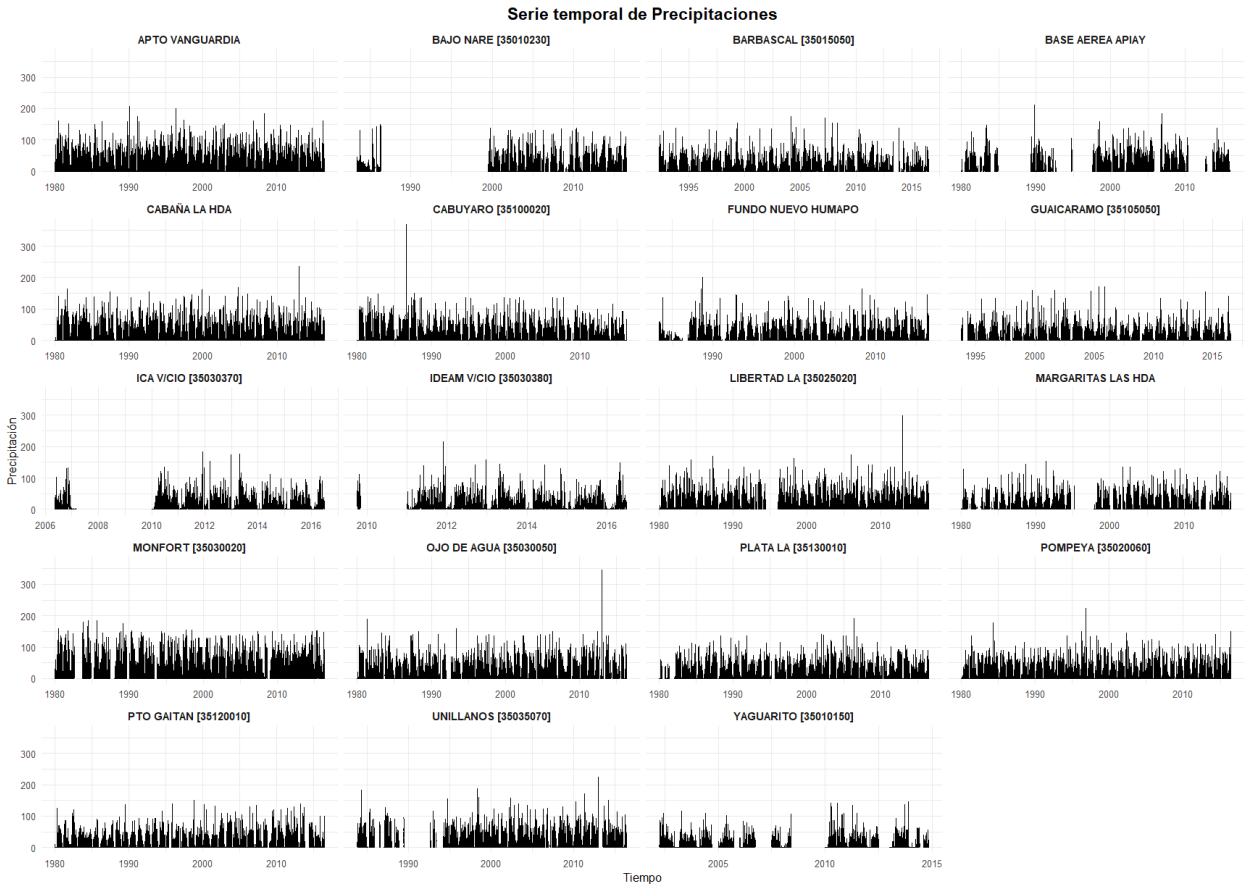

Figura A-3.: Dispersión de la Precipitación por estaciones 56

Tabla A-2.: Tabla con total datos faltantes

Nombre	lat	lon	na_percent
APTO VANGUARDIA	4,161919444	-73,61757778	0,04 %
LEJANIAS [32065010]	3,533333333	-74,01666667	0,27 %
SAN JUAN DE ARAMA	3,375083333	-73,89194444	0,52 %
SAN LUIS CUBARRAL	3,790722222	-73,84277778	0,56 %
POMPEYA [35020060]	4,03975	-73,368	0,56 %
FUENTE DE ORO	3,466666667	-73,63333333	1,03 %
CABUYARO [35100020]	4,283722222	-72,79283333	1,42 %
TORO EL [35010060]	3,786666667	-73,40055556	1,53 %
SAN MARTIN [35010090]	3,750277778	-73,70083333	1,57 %
ACACIAS [35010020]	3,994638889	-73,76558333	1,65 %
NARE [35010080]	3,793888889	-73,14972222	1,71 %
AGUAS CLARAS [32070110]	3,472027778	-73,85783333	2,02 %
COOPERATIVA LA [32075060]	3,366666667	-73,7	2,10 %
PTO RICO [32080010]	2,943333333	-73,20972222	2,14 %
CABAÑA LA HDA	4,300444444	-73,3575	2,21 %
TIERRA GRATA [32070080]	3,0875	-73,22638889	2,71 %
GUAMAL [35010070]	3,878333333	-73,75861111	2,75 %
PTO LOPEZ [35010010]	4,105027778	-72,9365	3,23 %
PTO LLERAS [32075080]	3,267694444	-73,37319444	3,32 %
PUERTO ANGOSTURAS	3,792666667	-73,92247222	3,63 %
CAÑO HONDO [35010040]	3,924222222	-73,81458333	3,77 %
MESA DE FERNANDEZ	3,456555556	-74,02994444	4,27 %
PTO GAITAN [35120010]	4,311388889	-72,07663889	4,46 %
GUAICARAMO [35105050]	4,468666667	-72,95358333	4,72 %
FUNDO NUEVO HUMAPO	4,327722222	-72,39144444	6,12 %
CAÑO BLANCO [32070020]	3,25	-73,51666667	6,38 %
PINALITO [32070040]	2,983333333	-73,63333333	6,53 %
URIBE LA [32020020]	3,243222222	-74,35372222	6,57 %
PLATA LA [35130010]	3,956083333	-72,76572222	6,88 %
CAMPO ALEGRE	3,2	-73,75	6,94 %
LIBERTAD LA [35025020]	4,057361111	-73,46791667	7,17 %
BARBASCAL [35015050]	3,630777778	-73,34927778	7,23 %
OJO DE AGUA [35030050]	4,091138889	-73,44877778	7,35 %
HOLANDA LA [32075040]	3,516333333	-73,71602778	7,94 %
CALIME [32060060]	3,742333333	-73,83469444	8,19 %
MESETAS [32075050]	3,380055556	-74,04297222	9,13 %
MAPIRIPAN [32090010]	2,889722222	-72,13011111	9,36 %
MACARENA LA [32035010]	2,176111111	-73,79333333	9,60 %
RAUDAL UNO [32030020]	2,3225	-73,94388889	9,92 %
MICOS LOS [32070030]	3,316388889	-73,89216667	10,01 %
MONFORT [35030020]	4,310083333	-73,64802778	10,11 %
MESA DE YAMANES	3,530944444	-73,90108333	11,06 %
PEÑAS BLANCAS	3,316666667	-73,91666667	14,72 %
CARIMAGUA [33035010]	4,574416667	-71,34075	17,13 %
IDEAM V/CIO [35030380]	4,144083333	-73,64183333	17,40 %
MARIPOSA LA [32010010]	2,562833333	-74,10308333	18,09 %
BALSORA LA [32035020]	2,438388889	-73,93227778	19,63 %

Tabla A-3.: Suma de precipitación por Mes y año

Año	Mes	Suma precipitación
2002	5	30314
2010	4	29187
2011	5	27759
2000	5	27644
2007	6	27329
2013	5	27320
2004	5	27251
1999	4	27195
1996	5	27072
2011	4	27007
2004	6	26357
2005	5	26049
2014	4	25947
2010	6	25858
1994	5	25247
2005	4	24839
2009	6	24431
1997	5	24013
2014	6	23890
2016	4	23792
2002	6	23703
2008	5	23238
1998	7	22793
2007	5	22734
1998	4	22702
2003	6	22400
2001	5	22298
1998	5	22223
2015	6	21950
2007	4	21799
2013	4	21769
1999	5	21760
1984	6	21730
1995	6	21627
1997	6	21437
2008	6	21388
1990	5	21286
1998	6	21129
2002	4	21027

Tabla A-4.: Meses con mayor Precipitación

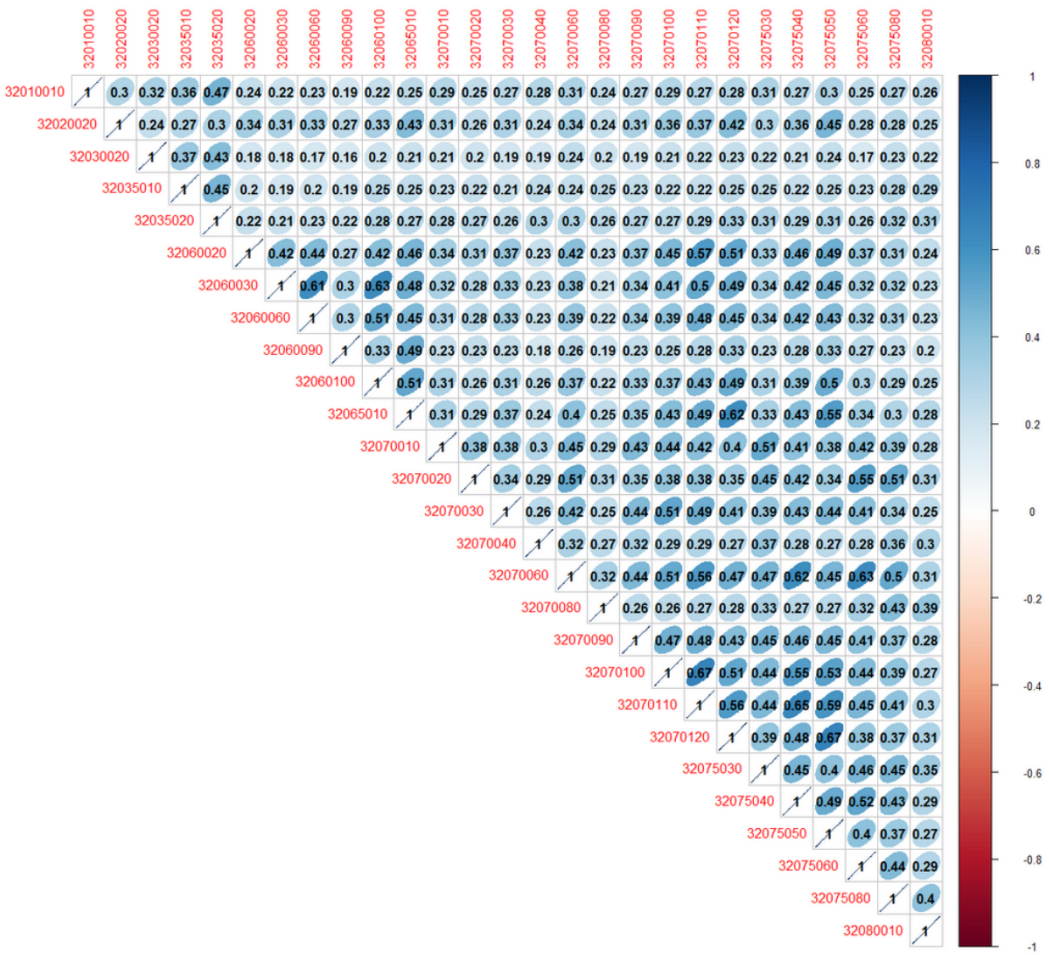

Figura A-4.: Correlación entre estaciones 1

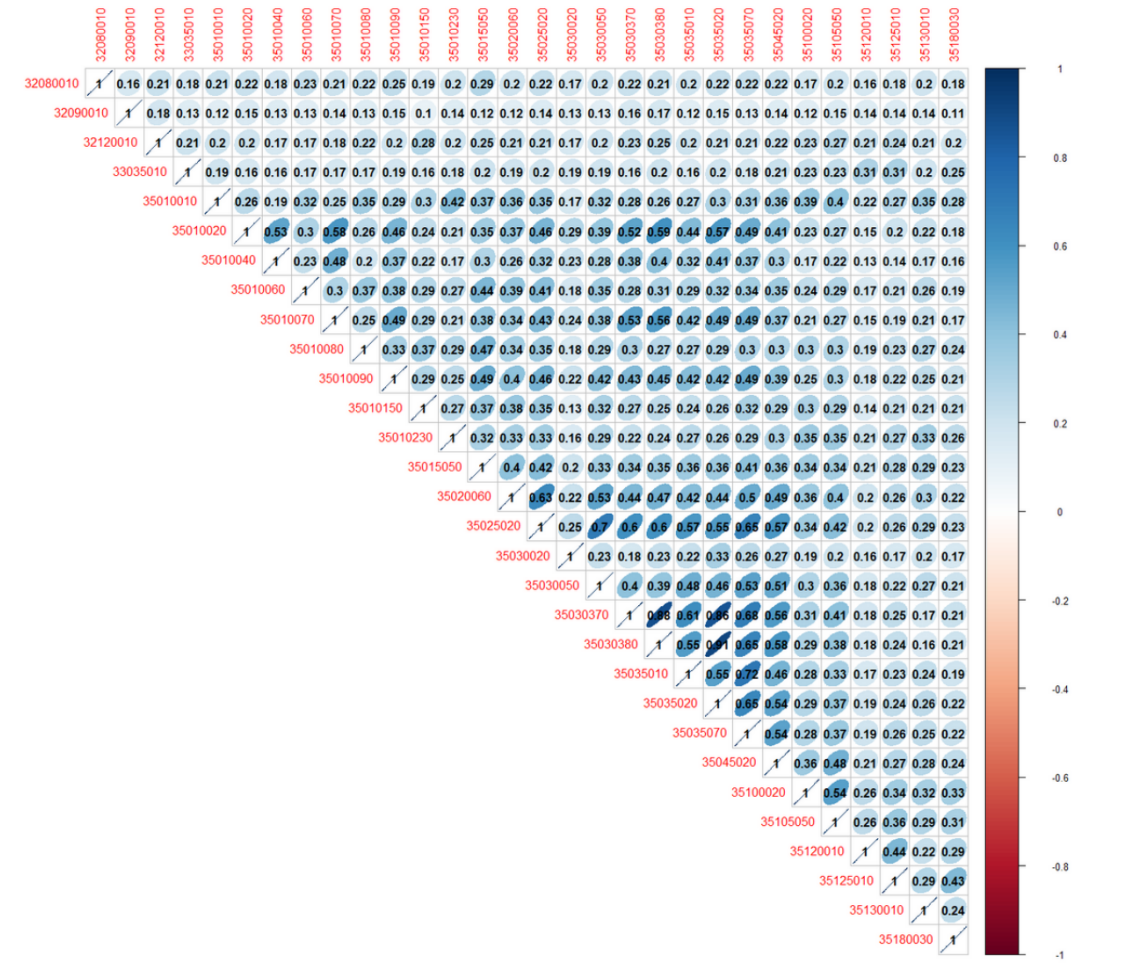

Figura A-5.: Correlación entre estaciones 2

Tabla A-5.: Métricas por estación

Municipio	Nombre de estacion	RMSE	MAE	MAPE	R2
URIBE	MARIPOSA LA [32010010]	4,258066	0,905027	17,578631	0,920582
URIBE	URIBE LA [32020020]	4,851	0,923	15,097	0,915
LA MACARENA	RAUDAL UNO [32030020]	4,798855	0,979426	21,251261	0,926992
LA MACARENA	MACARENA LA [32035010]	7,313	1,372	9,421	0,916
LA MACARENA	BALSORA LA [32035020]	5,075202	0,963752	28,15447	0,931105
EL CASTILLO	MESA DE YAMANES	4,862708	0,942862	14,366346	0,922346
CUBARRAL	SAN LUIS CUBARRAL	4,54904	0,9468	85,226893	0,914113
EL DORADO	CALIME [32060060]	4,862443	1,009352	17,31301	0,921899
LEJANÍAS	LEJANIAS CASTILLO	4,848751	0,966758	45,524525	0,91132
CUBARRAL	PUERTO ANGOSTURAS	5,789162	1,096191	14,397357	0,907167
LEJANÍAS	LEJANIAS [32065010]	4,344699	0,888901	37,744175	0,927714
VISTAHERMOSA	CAMPO ALEGRE	4,984968	0,959127	43,335676	0,894724
FUENTE DE ORO	CAÑO BLANCO [32070020]	4,898594	1,003295	11,600904	0,90585
SAN JUAN DE ARAMA	MICOS LOS [32070030]	4,97123	1,036014	44,928592	0,901648
VISTAHERMOSA	PINALITO [32070040]	4,41959	0,909376	13,980187	0,920208
FUENTE DE ORO	FUENTE DE ORO	4,56939	0,943715	13,772535	0,924298
PUERTO LLERAS	TIERRA GRATA [32070080]	4,306473	0,903624	12,477548	0,932054
SAN JUAN DE ARAMA	PEÑAS BLANCAS	5,281487	1,01156	9,5386486	0,920024
SAN JUAN DE ARAMA	SAN JUAN DE ARAMA	4,569705	0,999643	13,699877	0,920355
GRANADA	AGUAS CLARAS [32070110]	4,957	0,927	14,622	0,913
SAN JUAN DE ARAMA	MESA DE FERNANDEZ	5,11245	1,036294	13,562975	0,91325
VISTAHERMOSA	VISTA HERMOSA	4,515767	0,930445	42,46199	0,924199
GRANADA	HOLANDA LA [32075040]	5,059	1,000	10,158	0,910
MESETAS	MESETAS [32075050]	4,112	0,891	13,997	0,929
FUENTE DE ORO	COOPERATIVA LA [32075060]	5,559	1,051	13,024	0,907
PUERTO LLERAS	PTO LLERAS [32075080]	4,429	0,928	14,736	0,918
PUERTO RICO	PTO RICO [32080010]	5,109329	0,964147	65,723943	0,930439
MAPIRIPÁN	MAPIRIPAN [32090010]	6,295495	1,120558	10,265354	0,911027
MAPIRIPÁN	SAN IGNACIO [32120010]	5,544562	1,021148	52,672095	0,909425
PUERTO GAITÁN	CARIMAGUA [33035010]	4,647358	0,952585	36,846726	0,904791
PUERTO LÓPEZ	PTO LOPEZ [35010010]	7,974	1,402	11,566	0,905
ACACÍAS	ACACIAS [35010020]	4,440801	0,954096	12,12018	0,922684
GUAMAL	CAÑO HONDO [35010040]	5,33866	1,005773	24,694665	0,886141
CASTILLA LA NUEVA	TORO EL [35010060]	4,439	0,959	47,789	0,903
GUAMAL	GUAMAL [35010070]	4,273	0,918	22,485	0,920
PUERTO LÓPEZ	NARE [35010080]	4,873	1,015	12,489	0,904
SAN MARTÍN	SAN MARTIN [35010090]	5,958184	1,097831	45,490802	0,9083
SAN CARLOS DE GUAROA	YAGUARITO [35010150]	6,411	1,139	12,110	0,906
PUERTO LÓPEZ	BAJO NARE [35010230]	4,840595	0,978341	15,551202	0,907513
SAN MARTÍN	BARBASCAL [35015050]	6,263	1,232	28,560	0,890
VILLAVICENCIO	POMPEYA [35020060]	7,128576	1,339864	14,76607	0,907622
VILLAVICENCIO	LIBERTAD LA [35025020]	4,119381	0,974074	7,9004885	0,916605
EL CALVARIO	MONFORT [35030020]	4,479	0,931	15,050	0,889
VILLAVICENCIO	OJO DE AGUA [35030050]	4,853	0,957	20,913	0,891
VILLAVICENCIO	ICA V/CIO [35030370]	4,256875	0,902277	33,941442	0,91997
VILLAVICENCIO	IDEAM V/CIO [35030380]	4,277574	0,901786	45,919525	0,907293
VILLAVICENCIO	BASE AEREA APIAY	4,225	0,861663	64,351978	0,935281

Tabla A-6.: Estadísticas descriptivas por estación

NOMBRE	count	mean	std	min	max	var	cv
MARIPOSA LA [32010010]	9891	7.92	14.21	0	147	201.81	1.79
URIBE LA [32020020]	7963	9.86	17.33	0	185	300.27	1.76
RAUDAL UNO [32030020]	10851	7.99	14.54	0	190	211.39	1.82
MACARENA LA [32035010]	12023	7.60	13.57	0	187.9	184.27	1.79
BALSORA LA [32035020]	9681	8.39	15.58	0	148	242.76	1.86
MESA DE YAMANES[32060020]	11857	11.56	19.84	0	182	393.57	1.72
SAN LUIS CUBARRAL[32060030]	13257	13.27	22.72	0	250	516.20	1.71
CALIME [32060060]	12155	16.14	24.68	0	287	609.35	1.53
LEJANIAS CASTILLO [32060090]	8087	13.02	20.92	0	193	437.60	1.61
PUERTO ANGOSTURAS[32060100]	5385	8.38	15.98	0	205.2	255.26	1.91
LEJANIAS [32065010]	9623	7.69	15.84	0	165	251.06	2.06
CAMPO ALEGRE[32070010]	12406	7.25	14.85	0	200	220.56	2.05
CAÑO BLANCO [32070020]	12453	8.48	15.99	0	153	255.88	1.89
MICOS LOS [32070030]	11997	7.66	14.92	0	325	222.52	1.95
PINALITO [32070040]	12461	7.42	14.91	0	236	222.27	2.01
FUENTE DE ORO [32070060]	13164	8.64	16.72	0	174.1	279.48	1.93
TIERRA GRATA [32070080]	11684	8.32	15.93	0	174	253.81	1.91
PEÑAS BLANCAS[32070090]	10141	7.93	15.05	0	157	226.41	1.89
SAN JUAN DE ARAMA[32070100]	11839	9.84	17.01	0	170.1	289.22	1.73
AGUAS CLARAS [32070110]	9453	7.61	14.50	0	156	210.18	1.91
MESA DE FERNANDEZ[32070120]	7910	7.84	15.72	0	231	247.23	2.01
VISTA HERMOSA[32075030]	10612	8.12	14.41	0	162.1	207.66	1.77
HOLANDA LA [32075040]	12090	7.25	14.30	0	190.6	204.63	1.97
MESETAS [32075050]	10732	7.86	15.31	0	195	234.26	1.95
COOPERATIVA LA [32075060]	6815	8.20	15.26	0	140	232.98	1.86
PTO LLERAS [32075080]	8572	7.83	15.60	0	322	243.43	1.99
PTO RICO [32080010]	12986	7.62	15.09	0	147	227.84	1.98
MAPIRIPAN [32090010]	10899	6.88	14.86	0	170	220.93	2.16
SAN IGNACIO [32120010]	7740	9.78	18.49	0	170	341.85	1.89
CARIMAGUA [33035010]	11048	13.68	22.35	0	225	499.34	1.63
PTO LOPEZ [35010010]	12900	8.99	17.89	0	182.2	320.10	1.99
ACACIAS [35010020]	13081	9.60	18.02	0	163.8	324.75	1.88
CAÑO HONDO [35010040]	12828	8.29	16.62	0	187	276.28	2.01
TORO EL [35010060]	13098	8.60	16.61	0	230	275.92	1.93
GUAMAL [35010070]	12905	7.83	15.95	0	142	254.58	2.04
NARE [35010080]	13074	7.48	15.50	0	175	240.37	2.07
SAN MARTIN [35010090]	13122	7.55	15.47	0	298	239.30	2.05
YAGUARITO [35010150]	3302	8.07	15.95	0	170	254.26	1.98
BAJO NARE [35010230]	6710	15.11	22.93	0	185	526.00	1.52
BARBASCAL [35015050]	8185	10.86	20.12	0	175	404.81	1.85
POMPEYA [35020060]	13226	9.87	18.60	0	215	345.82	1.88
LIBERTAD LA [35025020]	12375	10.61	18.61	0	212.5	346.23	1.75
MONFORT [35030020]	11949	10.75	18.65	0	223.4	347.80	1.73
OJO DE AGUA [35030050]	12295	9.32	17.45	0	168	304.59	1.87
ICA V/CIO [35030370]	2636	6.49	14.63	0	136	214.07	2.25
IDEAM V/CIO [35030380]	2036	6.54	13.77	0	137	189.54	2.11
BASE AEREA APIAY[35035010]	7049	7.51	15.50	0	172	240.18	2.06
APTO VANGUARDIA[35035020]	13326	6.53	14.57	0	160	212.24	2.23
UNILLANOS [35035070]	9101	6.39	13.51	0	153.5	182.41	2.11