UNIVERSIDAD NACIONAL INTERCULTURAL DE LA SELVA CENTRAL JUAN SANTOS ATAHUALPA

FACULTAD DE INGENIERÍA

ESCUELA PROFESIONAL DE INGENIERÍA AMBIENTAL

Imputación de datos faltantes del material particulado $PM_{2.5}$ y su asociación con variables meteorológicas en la ciudad de Chanchamayo, región Junín, 2025

TESIS

Para optar el título profesional de Ingeniero Ambiental

AUTORES

Jeferson Walter Barrera Quispe

Fredy Andres Perez Julian

ASESOR

Dr. Daniel Martin Álvarez Tolentino

Chanchamayo, Perú

2026

AGRADECIMIENTOS

A Dios, por protegernos, guiarnos y ser nuestro amparo constante, especialmente en los momentos de mayor dificultad; por concedernos la perseverancia para confiar en nuestras capacidades y por resguardar a nuestras familias. Asimismo, le agradecemos por las personas que dispuso en nuestro camino, quienes con su apoyo contribuyeron significativamente al éxito de esta investigación.

A nuestro asesor, el Dr. Daniel Martin Álvarez Tolentino, por su invaluable labor como guía, así como por su comprensión, paciencia y sabiduría compartidas en cada una de las etapas del desarrollo de este estudio.

A nuestros padres, por depositar su confianza incondicional en nosotros y por brindarnos enseñanzas de perseverancia, paciencia y tolerancia, las cuales han sido pilares fundamentales en nuestro crecimiento profesional.

A nuestros docentes y amistades, quienes a través de sus conocimientos y experiencias nos motivaron constantemente a seguir esta línea de investigación y a consolidar nuestro camino profesional.

DEDICATORIA

A Dios, por ser mi constante amparo y fortaleza. A mis padres, Andrés y Victoria, pilares fundamentales de mi vida, por su apoyo incondicional y por sembrar en mí la sabiduría a través de sus consejos. A mi hermano, Elvis, por transmitirme su experiencia y demostrarme que el camino se recorre mejor con valentía y perseverancia. Este logro es de ellos, pues sin su guía nada de esto habría sido posible.

Pérez Julián Fredy Andrés

A Dios por brindarme la vida, fortaleza y sabiduría, a mis padres, por su amor incondicional, sacrificio y apoyo constante; a mi hermano, por su compañía y motivación en todo momento y a una persona muy especial que estuvo a mi lado brindándome su apoyo en todo momento. Por lo que, dedico este trabajo de investigación, que representa el esfuerzo, la perseverancia y la culminación de una importante etapa de mi vida, a las personas que han sido mi mayor fortaleza y motivación para nunca rendirme.

Jeferson Walter Barrera Quispe

RESUMEN

El trabajo de investigación tuvo como objetivo principal determinar la mejor técnica de imputación más eficiente para completar datos faltantes de material particulado fino ($PM_{2.5}$) asociado a variables meteorológicas en el distrito de Chanchamayo durante el periodo 2025, se realizó el procesamiento de series temporales provenientes de la estación La Merced que se ubica dentro del entorno urbano presento un 16.6 % de pérdida de información, y en la estación Santa Isabel ubicado en el entorno rural, presentó un 10.1 % de vacíos en sus registros. Se evaluaron comparativamente tres metodologías estadísticas, el algoritmo de aprendizaje MissForest, imputación múltiple por ecuaciones encadenadas – MICE y el método de sustitución por valor simple. El rendimiento y la precisión de cada técnica se cuantificaron mediante un esquema de validación cruzada utilizando como métricas de control el Error Cuadrático Medio (RMSE), el Error Absoluto Medio (MAE) y el coeficiente de determinación (R^2). Los resultados del proceso de validación demostraron la superioridad del algoritmo MissForest, al alcanzar los menores valores de error en la estación urbana (RMSE = 6.945 ug/m^3; MAE = 4.659 ug/m^3; R^2 = 0.319) y en la estación rural (RMSE = 8.854 ug/m^3; MAE = 4.852 ug/m^3; R^2 = 0.320), superando el desempeño de MICE e imputación por media. Se concluye que la integración de algoritmos de Machine Learning no paramétricos asociados a las variables meteorológicas resulta ser más precisa para el control de calidad y la reconstrucción de series temporales ambientales.

Palabras clave: Material particulado ($PM_{2.5}$), imputación de datos faltantes, MissForest, modelo lineal generalizado, variables meteorologicas, Chanchamayo.

ABSTRACT

The main objective of this research was to determine the most efficient imputation technique for completing missing data on fine particulate matter ($PM_{2.5}$) associated with meteorological variables in the Chanchamayo district during the period 2025. Time series data from the La Merced station, located within the urban area, were processed. This station presented a 16.6% data loss. Data from the Santa Isabel station, located in the rural area, presented a 10.1% gap in its records. Three statistical methodologies were comparatively evaluated: the MissForest learning algorithm, multiple imputation by chained equations (MICE), and the single value substitution method. The performance and accuracy of each technique were quantified using a cross-validation scheme with the Mean Squared Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (R^2) as control metrics. The validation process results demonstrated the superiority of the MissForest algorithm, achieving the lowest error values at the urban station (RMSE = 6.945 ug/m³; MAE = 4.659 ug/m³; R^2 = 0.319) and the rural station (RMSE = 8.854 ug/m³; MAE = 4.852 ug/m³; R^2 = 0.320), outperforming MICE and mean imputation. It is concluded that integrating non-parametric machine learning algorithms with meteorological variables proves to be more accurate for quality control and the reconstruction of environmental time series.

Keywords: Particulate matter ($PM_{2.5}$), Missing data imputation, MissForest, Generalized linear models – GLM, Meteorological variables, Chanchamayo

ÍNDICE

AGRADECIMIENTOS ii

DEDICATORIA iii

RESUMEN iv

ABSTRACT v

1. CAPÍTULO I 12

PLANTEAMIENTO DEL PROBLEMA 12

1.1. Identificación y Determinación del Problema 12

1.2. Formulación del Problema 14

1.2.1. Problema General 14

1.2.2. Problemas Específicos 14

1.3. Objetivos 14

1.3.1. Objetivo General 14

1.3.2. Objetivos Específicos 14

1.4. Definición y Operacionalización de Variables 15

1.5. Hipótesis 15

1.5.1. Hipótesis Alternativa 15

1.5.2. Hipótesis Nula 16

1.6. Justificación 16

1.6.1. Social 16

1.6.2. Ambiental 16

1.6.3. Económica 16

2. CAPÍTULO II 17

MARCO TEÓRICO 17

2.1. Antecedentes 17

2.1.1. Internacionales 17

2.1.2. Nacional 18

2.1.3. Regional 20

2.2. Bases Teóricas 21

2.2.1. Calidad del Aire 21

2.2.2. Contaminantes Atmosféricos 21
2.2.3. Componentes Primarios 21
2.2.4. Contaminantes Secundarios 21
2.2.5. Cambio Climático 21
2.2.6. Condiciones Meteorológicas y su Influencia en la Calidad del Aire 22
2.2.7. Temperatura 22
2.2.8. Humedad 22
2.2.9. Viento 22
2.2.10. Presión Atmosférica 22
2.2.11. Monitoreo de Calidad de Aire 22
2.2.12. Material Particulado 22
2.2.12.1. Definición y Características Físicas. 22
2.2.12.2. Material Particulado Fino PM_{10}. 23
2.2.12.3. Material Particulado Fino $PM_{2.5}$. 23
2.2.12.4. Fuente de Emisión de Material Particulado. 23
2.2.12.5. Efectos de Salud. 24
2.2.12.6. Monitoreo y Medición de Datos. 24
2.2.12.6.1. Temperatura y Humedad Relativa. 24
2.2.12.6.2. Velocidad y Dirección del Viento. 25
2.2.12.6.3. Presión Atmosférica 25
2.2.12.6.4. Precipitación. 26
2.2.12.6.5. Nubosidad y Radiación Solar. 26
2.2.12.6.6. Imputación de Datos Faltantes. 26
2.2.13. Métodos de Imputación para Datos Faltantes 26
2.2.13.1. Métodos de Imputación Simple. 26
2.2.13.2. Métodos de Imputación Avanzados. 27
2.2.14. R Core Team 28
3. CAPÍTULO III 29
METODOLOGÍA DE LA INVESTIGACIÓN 29
3.1. Ámbito de Estudio 29

3.2. Nivel, Tipo y Diseño de Investigación 30
3.2.1. Nivel 30
3.2.2. Tipo 30
3.2.3. Diseño 31
3.3. Población y Muestra 31
3.3.1. Población 31
3.3.2. Muestra 33
3.4. Procedimientos, Técnicas e Instrumentos de Recolección de Datos 34
3.4.1. Procedimiento 34
3.4.2. Técnicas 36
3.4.3. Instrumentos 37
3.5. Análisis de Datos 37
3.6. Consideraciones Éticas 39
4. CAPÍTULO IV 40
RESULTADOS Y DISCUSIONES 40
4.1. Presentación de Resultados y Discusiones 40
4.1.1. Identificación de Patrones de Datos Faltantes en las Series Temporales de $PM_{2.5}$ 40
4.1.1.1. Estación La Merced. 40
4.1.1.2. Estación Santa Isabel 43
4.1.2. Análisis de Patrones de Datos Faltantes en las Series Temporales de $PM_{2.5}$ 46
4.1.2.1. Estación Santa Isabel 46
4.1.2.2. Estación La Merced 48
5.1.1. Sistema de validación cruzada en los métodos de imputación aplicadas al Material Particulado $PM_{2.5}$ y su asociación a las variables meteorológicas. 51
5.1.1.1. Series de tiempo medidos en la estación de Santa Isabel 51
5.1.1.2. Series de tiempo medidos en la estación de La Merced 53
5.1.1.3. Análisis de correlación entre concentraciones de $PM_{2.5}$ y variables meteorológicas en la estación La Merced 54
5.1.1.4. Evaluación de significancia estadística del Modelo Lineal Generalizado (GLM) en la estación La Merced 57

5.1.1.5. Modelo Lineal Generalizado de material particulado $PM_{2.5}$ observado vs el predicho en la estación La Merced 59

5.1.1.6. Rendimiento de Random Forest meteorológico del material particulado $PM_{2.5}$ observado vs el predicho en la estación La Merced 60

5.1.1.7. Desempeño comparativo de los modelos de imputación de material particulado $PM_{2.5}$ en la estación La Merced 61

5.1.1.8. Análisis de correlación entre concentraciones de $PM_{2.5}$ y variables meteorologicas en la estación Santa Isabel 62

5.1.1.9. Evaluación de significancia estadística del Modelo Lineal Generalizado (GLM) en la estación Santa Isabel 65

5.1.1.10. Modelo Lineal Generalizado de material particulado $PM_{2.5}$ observado vs el predicho en la estación Santa Isabel 67

5.1.1.11. Rendimiento de Random Forest meteorológico del material particulado $PM_{2.5}$ observado vs el predicho en la Estación Santa Isabel 68

5.1.1.12. Desempeño comparativo de los modelos de imputación de material particulado $PM_{2.5}$ en la estación Santa Isabel 70

4.2. Prueba de Hipótesis 74

CONCLUSIONES 76

RECOMENDACIONES 77

REFERENCIAS BIBLIOGRÁFICAS 78

ANEXOS 86

Anexo 1. Matriz de consistencia 86

Anexo 2. Panel fotográfico 87

Anexo 3. Instrumentos de recolección de datos 92

ÍNDICE DE TABLAS

Tabla 1. Operacionalización de variables 15
Tabla 2. Modelos estadísticos aplicados en la imputación de datos faltantes de $PM_{2.5}$ 39
Tabla 3. Resumen estadístico mensual de concentraciones de $PM_{2.5}$ medidos en la estación de La Merced, Chanchamayo, 2025 42
Tabla 4. Resumen mensual de $PM_{2.5}$ en la estación de Santa Isabel, Chanchamayo, 2025 45
Tabla 5. Criterios para permitir la imputación de $PM_{2.5}$ medidos en la estación de Santa Isabel, Chanchamayo, 2025 47
Tabla 6. Evaluación de métodos de imputación de $PM_{2.5}$ medidos en la estación de Santa Isabel, Chanchamayo, 2025 48
Tabla 7. Criterios para permitir la imputación de $PM_{2.5}$ medidos en la estación de La Merced, Chanchamayo, 2025 49
Tabla 8. Evaluación de métodos de imputación de $PM_{2.5}$ medidos en la estación de La Merced, Chanchamayo, 2025 50
Tabla 9. Métricas de ACF y PACF de modelos de series de tiempo medidos en la estación de Santa Isabel, Chanchamayo, 2025 52
Tabla 10. Métricas de ACF y PACF de modelos de series de tiempo medidos en la estación de La Merced, Chanchamayo, 2025 54
Tabla 11. Matriz de correlación $PM_{2.5}$ observada y variables meteorológicas en la estación de La Merced, Chanchamayo, 2025 56
Tabla 12. ANOVA del GLM meteorológico observad - Significancia de variables en la estación de La Merced, Chanchamayo, 2025 58
Tabla 13. Comparación de modelos de $PM_{2.5}$: GLM vs Random Forest vs Imputación en la estación de La Merced, Chanchamayo, 2025 62
Tabla 14. Matriz de correlación $PM_{2.5}$ observada y variables meteorologícas en la estación de Santa Isabel, Chanchamayo, 2025 64
Tabla 15. ANOVA del GLM meteorológico (observado) - Significancia de variables en la estación de Santa Isabel, Chanchamayo, 2025. 66
Tabla 16. Comparación de modelos de $PM_{2.5}$: GLM vs Random Forest vs Imputación en la estación de Santa Isabel, Chanchamayo, 2025 71
Tabla 17. Modelos predictivos resultantes de los métodos de imputación para $PM_{2.5}$ 73
Tabla 18. Comparativa de métricas de desempeño y significancia estadística de los modelos de imputación por estación 74
Tabla 19 Instrumentos de recolección de datos 92

ÍNDICE DE FIGURAS

Figura 1. Material Particulado, comparación de tamaño 23
Figura 2. Ubicación de lugar de estudio 29
Figura 3. Estación de monitoreo La Merced, Chanchamayo 30
Figura 4. Equipos: Estación meteorológica Davis Vantage Pro 32
Figura 5. Equipos: Sensor de monitoreo Purplair PA-II 33
Figura 6. Flujograma de procesos 34
Figura 7. Comportamiento del $PM_{2.5}$ y representatividad de los datos en la estación La Merced, Chanchamayo, 2025 41
Figura 8. Comportamiento del $PM_{2.5}$ y representatividad de los datos en la estación Santa Isabel, Chanchamayo, 2025 44
Figura 9. Gráficos ACF y PACF de series de tiempo medidos en la estación de Santa Isabel, Chanchamayo, 2025 52
Figura 10. Gráficos ACF y PACF de series de tiempo medidos en la estación de La Merced, Chanchamayo, 2025 54
Figura 11. Gráfico de GLM meteorológico con relación del $PM_{2.5}$ observado vs $PM_{2.5}$ predicho 60
Figura 12. Gráfico Random Forest meteorológico del $PM_{2.5}$ observado vs $PM_{2.5}$ predicho 61
Figura 13. Gráfico de GLM meteorológico con relación del $PM_{2.5}$ observado vs $PM_{2.5}$ predicho 68
Figura 14. Gráfico de Random Forest meteorológico del $PM_{2.5}$ observado vs $PM_{2.5}$ predicho 69
Figura 15 Estación de monitoreo Santa Isabel 87
Figura 16 Mantenimiento y limpieza de la estación de monitoreo Santa Isabel 87
Figura 17 Mantenimiento y calibración de la estación de monitoreo Santa Isabel 88
Figura 18 Calibración de estación meteorológica 88
Figura 19 Mantenimiento y calibración de sensor PurPlair 89
Figura 20 Verificación y descarga de data de la estación de monitoreo Santa Isabel 89
Figura 21 Estación de monitoreo La Merced 90
Figura 22 Mantenimiento y limpieza de la estación de monitoreo La Merced 90
Figura 23 *Calibración y mantenimiento de la estación de monitoreo La Merced* 91
Figura 24 Verificación y descarga de datos de la estación de monitoreo La Merced 91

CAPÍTULO I

PLANTEAMIENTO DEL PROBLEMA

1.1. Identificación y Determinación del Problema

Según la Organización Mundial de la Salud (OMS) la contaminación del aire por material particulado fino $PM_{2.5}$ es uno de los principales problemas ambientales y sanitarias que afecta a las ciudades, ecosistemas y personas debido a su pequeño tamaño que es menos de 2.5 micras, así también, aproximadamente el 99% de la población respira aire de mala calidad (OMS, 2022). Es por ello que los efectos del $PM_{2.5}$ en la salud están ampliamente documentados, diversos estudios epidemiológicos han demostrado su relación con enfermedades respiratorias como el asma, la enfermedad pulmonar obstructiva crónica (EPOC) y el cáncer de pulmón (MISAN, 2019).

A nivel internacional, diversos estudios han reportado que una de las dificultades en el monitoreo continuo de $PM_{2.5}$ son los datos faltantes, esto se origina debido a múltiples orígenes como: fallas en el suministro de energía eléctrica o fallas en los sistemas de comunicación o situaciones relacionadas con el clima (Arroyave et al., 2023). Por ende, estos datos sobre la calidad de aire se han convertido en algo muy común, debido a ello se han presentado diversas alternativas sobre cómo resolver estos problemas, siendo uno de los métodos más utilizados actualmente la imputación de datos faltantes. Durán (2005) indica que los datos perdidos afectan la precisión y validez de los resultados obtenidos, además increpa que se debe de realizar el tratamiento de estos datos perdidos con la finalidad de reducirlos.

En el Perú, la contaminación del aire se ha convertido en un desafío de gran envergadura, diversos estudios e informes oficiales señalan que las principales ciudades del país, como Lima, Arequipa y Trujillo, registran concentraciones de material particulado fino $PM_{2.5}$ que superan ampliamente los límites internacionales establecidos por la OMS (Montaño, 2023). En algunas zonas urbanas de Lima como San Martin de Porres y San Juan de Lurigancho se ha registrado 64.3 $\mu g/m^3$, alcanzando estos valores en fiestas navideñas, cifra que contrasta notablemente con el valor guía de 5 $\mu g/m^3$ recomendado para el promedio anual (Arce, 2024). Una de las principales fuentes que contribuyen a la emisión del $PM_{2.5}$ es el transporte urbano, especialmente en Lima Metropolitana, se estima que casi el 47% de la contaminación del aire proviene del parque vehicular (ATU, 2023).

En el país, ciudades como Lima o Arequipa cuentan con estaciones de monitoreo constantes, pero en muchas regiones del país especialmente en la Amazonía y la sierra central los sistemas de vigilancia son limitados, los datos son escasos, y en muchos casos se presentan

registros incompletos o interrumpidos. Esto impide conocer con exactitud la magnitud del problema y limita la capacidad de respuesta de las autoridades de salud y medio ambiente (Montaño, 2023).

En el distrito de Chanchamayo, se aprecia un entorno ambiental complejo caracterizado por una abundante biodiversidad y una actividad turística y agropecuaria en aumento, factores que han fomentado el crecimiento económico de la región. No obstante, en años recientes se ha observado un incremento en los índices de $PM_{2.5}$, vinculado a acciones como la deforestación, los incendios forestales y la incineración de desechos agrícolas, que afectan severamente los ecosistemas locales y dañan la calidad del aire (SERFOR, 2021). Las prácticas de tala y quema para el crecimiento agrícola han tenido un impacto considerable en la emisión de partículas contaminantes en la zona. Adicionalmente el (CEPLAN, 2023) resalta la importancia del seguimiento constante y de investigaciones integradas que toman en cuenta las variables climáticas características de Chanchamayo. Como la elevada humedad, la variabilidad térmica y las fuertes lluvias, que tienen un rol esencial en la dispersión o acumulación del $PM_{2.5}$ frecuentemente, los registros de concentración de $PM_{2.5}$ presentan datos faltantes, por fallas en los equipos de monitoreo, cortes de energía, mantenimiento o incluso por factores climáticos. Estos vacíos dificultan el análisis real de la calidad del aire y pueden llevar a interpretaciones erróneas o incompletas.

En ese sentido una posible solución para enfrentar este problema es la imputación de datos faltantes, una técnica estadística que permite estimar los valores ausentes en una serie de datos, utilizando patrones detectados en la información disponible. En este caso, se considera que las variables meteorológicas podrían ser clave para estimar de forma más precisa los valores de $PM_{2.5}$, ya que están estrechamente relacionadas con la dispersión, acumulación o transporte de contaminantes atmosféricos (Muñoz & Álvarez, 2009).

Por ello, esta investigación buscó no solo ofrecer una solución técnica al problema de los datos faltantes, sino también contribuir al fortalecimiento del monitoreo ambiental en zonas donde los recursos son limitados. Al proponer un enfoque basado en el análisis de variables meteorológicas, se logró generar una herramienta práctica que permita estimar con mayor precisión los niveles de contaminación atmosférica, lo cual es clave para proteger la salud pública y promover un desarrollo sostenible en el distrito de Chanchamayo.

1.2. Formulación del Problema

1.2.1. Problema General

¿Cuál es la mejor técnica de imputación de datos faltantes del material particulado $PM_{2.5}$ y su asociación con variables meteorológicas en la Ciudad de Chanchamayo, región Junín, 2025?

1.2.2. Problemas Específicos

¿Cuáles son los patrones de datos faltantes en las series temporales de $PM_{2.5}$ registradas en la ciudad de Chanchamayo, región Junín, 2025?

¿Cuáles son las técnicas de imputación aplicadas a los datos faltantes de material particulado $PM_{2.5}$ en la ciudad de Chanchamayo, región Junín, 2025?

¿Cuál es la precisión de cada técnica de imputación aplicadas a los datos faltantes de material particulado $PM_{2.5}$ y su asociación con variables meteorológicas en la ciudad de Chanchamayo, región Junín. 2025?

1.3. Objetivos

1.3.1. Objetivo General

Evaluar la mejor técnica de imputación de datos faltantes de material particulado $PM_{2.5}$ y su asociación con variables meteorológicas en la ciudad de Chanchamayo, región Junín, 2025.

1.3.2. Objetivos Específicos

Identificar los patrones de datos faltantes en las series temporales de $PM_{2.5}$ registradas en la ciudad de Chanchamayo, región Junín, 2025.

Implementar las técnicas de imputación aplicadas a los datos faltantes de material particulado $PM_{2.5}$ en la ciudad de Chanchamayo, región Junín, 2025.

Determinar la precisión de cada técnica de imputación aplicadas a los datos faltantes de material particulado $PM_{2.5}$ y su asociado con variables meteorológicas en la ciudad de Chanchamayo, región Junín, 2025.

1.4. Definición y Operacionalización de Variables

Tabla 1.

Operacionalización de variables

Variables	Definición conceptual	Dimensiones	Indicadores	Unidad de medida
a) Variable 1		Temperatura del aire	Temperatura máxima diaria Temperatura mínima diaria Temperatura promedio	°C
Parámetros meteorológicos	Son registros de variables atmosféricas tal como: temperatura, humedad, presión, velocidad y dirección del viento, y precipitación, que describen el estado y la evolución del clima en un lugar (EPA, 2024).	Punto de Rocío	Punto de Rocío	°C
		Velocidad del viento	Promedio horario de velocidad del viento	m/s
		Humedad relativa	% de humedad relativa del aire	%
		Presión atmosférica	Presión atmosférica actual	hPa / mb
		Precipitación	Volumen / totales de lluvia	mm
		Radiación solar	Energía solar incidente	W/m^2
b) Variable 2 **Material Particulado ($PM_{2.5}$)**	Materia liquida o sólida dispersa en el aire, de diámetro inferior o igual a 2.5 micras. Puede proceder de fuentes naturales, dependiendo de su tamaño y pueden permanecer en suspensión en la atmósfera desde unos segundos a varios meses (MINAM, 2020).	Partículas atmosféricas menor a $PM_{2.5}$	Concentración de $PM_{2.5}$ atmosférico ($PM_{2.5}$_atm)	µg/ m³

Fuente: Elaboración Propia

1.5. Hipótesis

1.5.1. Hipótesis Alternativa

H_a: Existe diferencias significativas en la eficiencia de las distintas técnicas de imputación a los datos faltantes de $PM_{2.5}$ y variables meteorológicas en el distrito de Chanchamayo durante el 2025.

1.5.2. Hipótesis Nula

H_o: No Existe diferencias significativas en la eficiencia de las distintas técnicas de imputación a los datos faltantes de $PM_{2.5}$ y variables meteorológicas en el distrito de Chanchamayo durante el 2025.

1.6. Justificación

1.6.1. Social

Este estudio contribuye a mejorar el acceso a información confiable sobre la calidad del aire en el distrito de Chanchamayo, un lugar donde la población está expuesta a fuentes de contaminación como el tránsito vehicular y las quemas agrícolas. La falta de datos completos sobre $PM_{2.5}$ imposibilita evaluar con precisión los riesgos para la salud, lo que influye en particular a los grupos más vulnerables como niños, adultos mayores y personas con enfermedades respiratorias. Al plantear un método para imputar datos faltantes empleando variables meteorológicas, esta investigación buscó fortalecer el monitoreo ambiental local, apoyar la toma de decisiones informadas y, por último, velar el cuidado de la salud pública y promover las condiciones de vida de la población.

1.6.2. Ambiental

Esta investigación es fundamental en el ámbito ambiental porque ayuda a tener un mejor control y seguimiento de la calidad del aire en Chanchamayo, una zona afectada por actividades como las quemas agrícolas y el tráfico. Al imputar los datos faltantes de $PM_{2.5}$ con apoyo de variables meteorológicas, se pudo contar con información más completa para tomar decisiones que ayuden a proteger el medio ambiente y minimizar los riesgos de contaminación.

1.6.3. Económica

Esta investigación es útil porque contar con datos más completos sobre la calidad del aire permite optimizar recursos en salud y gestión ambiental. Al aminorar la incertidumbre sobre los niveles reales de contaminación, se pueden tomar decisiones más eficientes efectuando gastos innecesarios en medidas correctivas y primando acciones preventivas que protejan tanto a la población como a las actividades productivas locales.

CAPÍTULO II
MARCO TEÓRICO

2.1. Antecedentes

2.1.1. Internacionales

Soler (2025), en su investigación titulada *"Evaluación de desempeño predictivo de modelos* basados *en análisis funcional y algoritmos de machine learning supervisado para concentraciones de $PM_{2.5}$ y PM_{10} en estaciones de monitoreo ambiental de Bogotá D.C. durante el periodo 2023 – 2024"*, tuvo como objetivo analizar las variaciones horarias que experimentan el $PM_{2.5}$ y el PM_{10} en zonas urbanas densamente pobladas, en las estaciones de Fontibón, Jazmín y Puente Aranda, con el propósito de mejorar su comprensión y fortalecer su actividad predictiva disponible. Empleó un enfoque basado en análisis funcional de datos (AFD), luego ajustaron tres modelos predictivos: ARIMA, FAR Y Random Forest. Sus resultados evidenciaron que el modelo FAR obtuvo el mejor desempeño al presentar valores más bajos de RMSE, MAE y MAPE. El modelo ARIMA mostró un rendimiento aceptable solo en escenarios con menor variabilidad horaria y Random Forest obtuvo una precisión moderada.

Naranjo & Escudero (2025) en su investigación titulada *"Random Forest como herramienta para mejorar la precisión en la imputación de datos meteorológicos en Chimborazo, Ecuador",* tuvieron como objetivo gestionar la información faltante en los registros meteorológicos monitoreados por el Grupo de Energías Alternativas y Ambientales (GEAA), mediante el análisis climático y el uso de técnicas de aprendizaje automático. Emplearon un enfoque cuantitativo con alcance descriptivo y comparativo para evaluar distintos algoritmos. Entre los métodos analizados se consideraron Random Forest, K-vecinos más cercanos (KNN) y el método de imputación por promedio. Realizaron la validación de los modelos mediante parámetros como el Error Cuadrático Medio (RMSE), el Error Medio Absoluto (MAE) y el coeficiente de determinación (R^2). En sus resultados se evidencio que el algoritmo Random Forest presenta mayor precisión en comparación con los otros métodos evaluados. Asimismo, observaron que, aunque este modelo requiere mayor capacidad computacional, ofrece mejores resultados en el procesamiento de datos complejos.

Moreno (2024), en su trabajo de investigación titulado "*Modelo predictivo de difusión de* contaminante *$PM_{2.5}$ en el área urbana de Bogotá entre 2017 – 2020 en presencia del Covid – 19*" tuvo como objetivo establecer un modelo predictivo para variables asociadas a problemáticas con comportamientos de difusión y generar un ejemplo para la ventana de 2020

- 2021 correspondiente a la presencia del confinamiento COVID-19, lo cual discute la difusión de material contaminante proveniente de fuentes estáticas y móviles registrado en la red de estaciones de calidad del aire de Bogotá, proponiendo un modelo predictivo espacio-temporal a partir de las metodologías de imputación, "MissForest" y de predicción "Spatial Autorregressive Hilbert (SARH,1)". Sus procedimientos le permitieron configurar espacialmente el dominio de las variables de difusión, así como, generar las superficies de concentración de material particulado $PM_{2.5}$, prediciendo los valores respectivos de las estaciones faltantes en la ciudad. Con ello concluyó que la combinación de técnicas de Machine Learning y modelos estadísticos espacio temporales potencia en gran medida los resultados en el campo de la predicción.

Cáceres (2018), en su tesis titulada "*Análisis estadístico de la razón entre el $PM_{2.5}$ y el PM_{10} de los datos continuos de las estaciones de la red de monitoreo de calidad del aire en Bogotá"*, tuvo como objetivo obtener la razón $PM_{10}/PM_{2.5}$ junto con variables meteorológicas como la precipitación, la velocidad y dirección del viento nos da información de los procesos y la procedencia del material particulado. Tomaron registros horarios de 6 estaciones de monitoreo entre los años 2014 - 2016, determinaron que el promedio fue de 0.44W de toda la localidad, el $PM_{2.5}$ y el PM_{10} explica casi totalmente la constancia en las velocidades del viento ya que los mecanismos de suspensión se reducen. Así mismo, por la tarde dicha correlación disminuye en la mayoría de las estaciones a causa de la presencia de emisiones regionales influenciadas por la velocidad del viento, las cuales se determinó su procedencia y su proceso de emisión principal a través de gráficos polares. Finalmente, comprobaron el fenómeno de barrido causado por la precipitación y su efecto tanto en el PM_{10} como en el $PM_{2.5}$ encontrando que este fenómeno no es constante en todas las estaciones de la red.

2.1.2. Nacional

Choquetico & Silva (2024), en su investigación titulada "*Modelo de regresión funcional para la predicción de $PM_{2.5}$ en función de PM_{10} en Lima-Perú"*, tuvo el objetivo desarrollar un modelo predictivo avanzado que estime los niveles de $PM_{2.5}$ en función de las concentraciones de PM_{10} utilizando registros diarios de calidad del aire desde 2023. Utilizo registros procedentes de la Plataforma de información históricos de la calidad del aire, el grupo de trabajo encargado del desarrollo del Índice Mundial de la Calidad del Aire (WAQI) y del Sistema de Gestión de Información de la Calidad del Aire en Perú (SENAMHI), inicio con 348 variables buscando determinar su idoneidad para su análisis propuesto, concluyó que la implementación del modelo de regresión funcional concurrente en esta investigación resulto clave para esclarecer el vínculo

complejo y evolutivo entre los contaminantes $PM_{2.5}$ y PM_{10}. El análisis evidencio oscilaciones en las concentraciones de $PM_{2.5}$ que no dependen directamente a variaciones en PM_{10}, lo cual propone que se presentan variable independientes y potencialmente invisibles que causan los niveles de $PM_{2.5}$.

Tello (2023), en su tesis titulada "R*elación entre las variables meteorológicas con la concentración de material particulado $PM_{2.5}$ en la ciudad del Cusco, periodo 2016 – 2021"*, su objetivo principal fue encontrar la relación entre variables meteorológicas de temperatura, rapidez del viento, precipitación, humedad relativa y presión atmosférica con la concentración de material particulado $PM_{2.5}$ entre los años 2016 al 2021. Empleó un diseño no experimental, contó con 5 puntos de muestreo. Entre sus resultados obtuvieron que las 5 áreas de investigación presentaron valores de $PM_{2.5}$ superiores al nivel admisible (ECA anual de 25 µg/m^3 D.S. N° 003.2017-MINAM y demostró que establece una correlación inversa entre $PM_{2.5}$, a menor temperatura mayor concentración de $PM_{2.5}$, a mayor rapidez del viento menor concentración de $PM_{2.5}$, a mayor humedad relativa menor concentración de $PM_{2.5}$, a mayor precipitación menor concentración de $PM_{2.5}$, a mayor presión atmosférica mayor concentración de $PM_{2.5}$. Concluyendo, que las zonas con mayor concentración de $PM_{2.5}$ son los distritos de San Jerónimo (103,92 µg/m$^{3)}$, Cusco (61,51 µg/m^3), San Sebastián (60,82 µg/m^3) y Wánchaq (55,08 µg/m^3).

Motocanche & Lazo (2020), en su estudio titulado "*Condiciones meteorológicas y material particulado $PM_{2.5}$ y PM_{10} en la construcción del hospital Hipólito Unanue, Tacna 2019"*, tuvo como propósito evaluar si las condiciones meteorológicas influyen en la dispersión de las concentraciones de Material Particulado en la fase de construcción del Hospital Hipólito Unanue de Tacna. Contó con un tipo de investigación descriptiva y 02 estaciones de muestreo de aire. Obteniendo como resultado que las dos estaciones de monitoreo EMHU-01 y EMHU-02 consignaron en el mes de junio para $PM_{2.5}$ cifras de 69.53 y 63.04 ug/m^3 y en el mes de setiembre en la estación EMHU-01 el valor de 64.40 ug/m^3 los cuales excedieron el estándar de calidad ambiental para aire. Estas cifras se debieron directamente a los elevados valores de Temperatura de 18 y 19.6°C y la velocidad de viento de 2.5 m/s y una relación inversa con la humedad relativa 76.3 y 75%. Además, el análisis de la varianza preciso que no existe diferencia significativa entre los valores obtenidos y ambas variables fueron comprobadas con la prueba estadística de Pearson.

Valeriano (2021), en su tesis titulada "*Relación de las variables meteorológicas en la dispersión* espacio *temporal de los contaminantes atmosféricos ($PM_{2.5}$, PM_{10}, SO_2, NO_2, O_3) del distrito de Pacocha, Moquegua, 2019 – 2020"*, su objetivo principal fue determinar la

relación de la temperatura, humedad relativa, velocidad del viento y la dirección del viento en la dispersión espacio temporal de los contaminantes atmosféricos (NO_2, O_3, SO_2, PM_{10} y $PM_{2.5}$). Empleó un esquema no experimental, longitudinal y una muestra de 01 estación de monitoreo de calidad de aire. Obtuvo que los sucesos meteorológicos sobre la variabilidad estacional de la concentración de contaminantes atmosféricos están vinculados con el proceso productivo. Para el 2019 se observa una fuerte correlación del PM_{10} y WS, O_3 y WS, Humedad y SO_2, NO_2 y Humedad, O_3 y temperatura. Para el 2020 se observa una fuerte correlación del PM_{10} y WS, O_3 y WS, O_3 y WD, PM_{10} y $PM_{2.5}$ con WD, SO_2 y temperatura, SO_2 humedad, NO_2 temperatura, NO_2 y humedad. Concluyendo que el estudio de conglomerados de las trayectorias de retorno de HYSPLIT muestra que los vientos proceden de Chile, Bolivia y el ámbito nacional.

2.1.3. Regional

Pardo & Soto (2026), en su trabajo titulado *"Estimación de concentraciones de Material Particulado fino $PM_{2.5}$ mediante modelos de Machine Learning en la ciudad de Concepción durante el periodo 2023-2025"* tuvo como objetivo evaluar las concentraciones de Material Particulado fino $PM_{2.5}$ por medio de modelos de Machine Learning, siendo una investigación cuantitativa de carácter explicativo - predictivo y contando con un diseño no experimental. Aplicaron tres modelos predictivos: Regresión Lineal, Random Forest y Redes Neuronales Artificiales (ANN). Obtuvieron resultados que demuestran que las concentraciones de $PM_{2.5}$ muestran elevada variabilidad temporal, con aumentos críticos en los meses secos de junio a septiembre. La ANN ofreció la mejor capacidad predictiva ($R^2 = 0.611$; RMSE = 5.417; MAE = 2.566), superando significativamente a los otros modelos ($p < 0.001$). Asimismo, se determinó que la variable más predominante es el $PM_{2.5}$ rezagado ($p = 0.917$), corroborando la fuerte sujeción temporal del contaminante. Se Deduce que las redes neuronales artificiales representan una herramienta efectiva para la predicción de la calidad del aire en zonas andinas, con ampliación viable en sistemas de alerta temprana y gestión ambiental.

Olarte et al., (2025) en su trabajo titulado "Modelo predictivo con aprendizaje automático para variables ambientales y $PM_{2.5}$ en Huachac, Junín, Perú" tuvo como objetivo modelar el comportamiento del ICA (índice de calidad del aire) de $PM_{2.5}$ empleando modelos predictivos de aprendizaje automático (ML) de regresión lineal, lasso, ridge y elastic net. Utilizaron un total de 16,543 registros de la zona de Huachac con regresores de humedad en % y temperatura en °C. Su enfoque de su estudio fue el ICA de $PM_{2.5}$ y las variables ambientales. Métodos: Aplicaron un análisis exploratorio de datos (AED) y modelos predictivos de

aprendizaje automático. Sus resultados: El ICA de $PM_{2.5}$ cuenta con valores altos en invierno y primavera, con promedios de 52.6 y 36.9, respectivamente, y valores bajos en verano. El uso de modelos de regresión produjo métricas precisas para elegir el mejor modelo para la predicción del ICA de $PM_{2.5}$. La comparación que realizó con otras investigaciones resalta la robustez de los modelos ML elegidos, resaltando el potencial del ML en el ICA de $PM_{2.5}$. Llegaron a las siguientes conclusiones: el modelo predictivo que encontraron fue $\alpha = 0{,}1111111$ y un valor Lambda $\lambda = 0{,}150025$, representado por ICA de $PM_{2.5} = 83{,}0846522 - 10{,}302222000$ (Humedad) $- 0{,}1268124$ (Temperatura).

2.2. Bases Teóricas

2.2.1. Calidad del Aire

La calidad del aire es una señal de cuando el aire está eximido de contaminantes atmosféricos y por lo tanto adecuado para ser respirado. No gozar de un ambiente con aire de calidad es un problema que implica riesgo o daño para la seguridad y la salud de las personas, el medio ambiente y bienes de cualquier naturaleza (INEI, 2018).

2.2.2. Contaminantes Atmosféricos

Según Reyes (2025), menciona que son sustancias, formas de energía o partículas físicas presentes en la atmósfera que, en concentraciones elevadas o de forma no natural, pueden causar daño o riesgo a la salud humana, el medio ambiente.

2.2.3. Componentes Primarios

Son los que se expulsa directamente a la atmósfera de manera nociva, como: el dióxido de azufre SO_2, monóxido de carbono (CO), óxido de nitrógeno (NO_x), entre otros. Las fuentes más comunes son: vehículos, incendios forestales, fábrica y agricultura (Davda, 2025).

2.2.4. Contaminantes Secundarios

Son aquellos que se componen en la atmósfera por medio de reacciones químicos atmosféricos que interviene sobre los contaminantes primarios. Son importantes contaminantes secundarios: ozono troposférico (O_3), lluvia ácida (SO_2 y NO_x) y ácido sulfúrico (H_2SO_4) (Davda, 2025).

2.2.5. Cambio Climático

El Cambio Climático (CC) constituye la alteración del clima a escala global como consecuencia de las altas emisiones y concentraciones de Gases de efecto invernadero (GEI) en la atmósfera (Salaverry & Botana, 2021).

2.2.6. Condiciones Meteorológicas y su Influencia en la Calidad del Aire

La concentración de contaminantes atmosféricos nocivos está aumentando, y una de las razones son los factores meteorológicos (Parth, 2024).

2.2.7. Temperatura

Las altas temperaturas pueden favorecer la formación de contaminantes secundarios, como el ozono, lo que incrementa las concentraciones de partículas finas.

2.2.8. Humedad

Un aumento en la humedad relativa puede causar que las partículas finas se agrupen, lo que modifica su tamaño y distribución.

2.2.9. Viento

La velocidad del viento es un elemento clave en la dispersión de los contaminantes. Vientos fuertes tienden a diluir las concentraciones de $PM_{2.5}$, mientras que vientos calmados favorecen su acumulación en áreas urbanas y rurales.

2.2.10. Presión Atmosférica

En condiciones de alta presión, los contaminantes tienden a concentrarse en la atmósfera, creando eventos de inversión térmica, donde el aire frío queda atrapado debajo de una capa de aire caliente, impidiendo la dispersión de los contaminantes.

Estudios previos han demostrado que las condiciones meteorológicas influyen significativamente en la dinámica de la calidad del aire, especialmente en áreas urbanas densamente pobladas, como el caso del distrito de Chanchamayo (Zhao et al., 2019; Seinfeld & Pandis, 2016).

2.2.11. Monitoreo de Calidad de Aire

El monitoreo de calidad del aire es fundamental para evaluar la presencia de contaminantes como partículas en suspensión, dióxido de azufre (SO_2), óxidos de nitrógeno (NO_x) y compuestos orgánicos volátiles (COV) (MEIN INGENIEROS S.A.C., 2025).

2.2.12. Material Particulado

2.2.12.1. **Definición y Características Físicas.**

Se define como una mezcla de partículas sólidas y gotas líquidas que se ubican en el aire. Algunas partículas, como la suciedad, el polvo, el hollín, o el humo, son lo debidamente grandes y oscuras como para observarlas a simple vista. Otras son tan

pequeñas que solo pueden detectarse mediante el uso de un microscopio electrónico (EPA, 2025).

2.2.12.2. Material Particulado Fino PM_{10}.

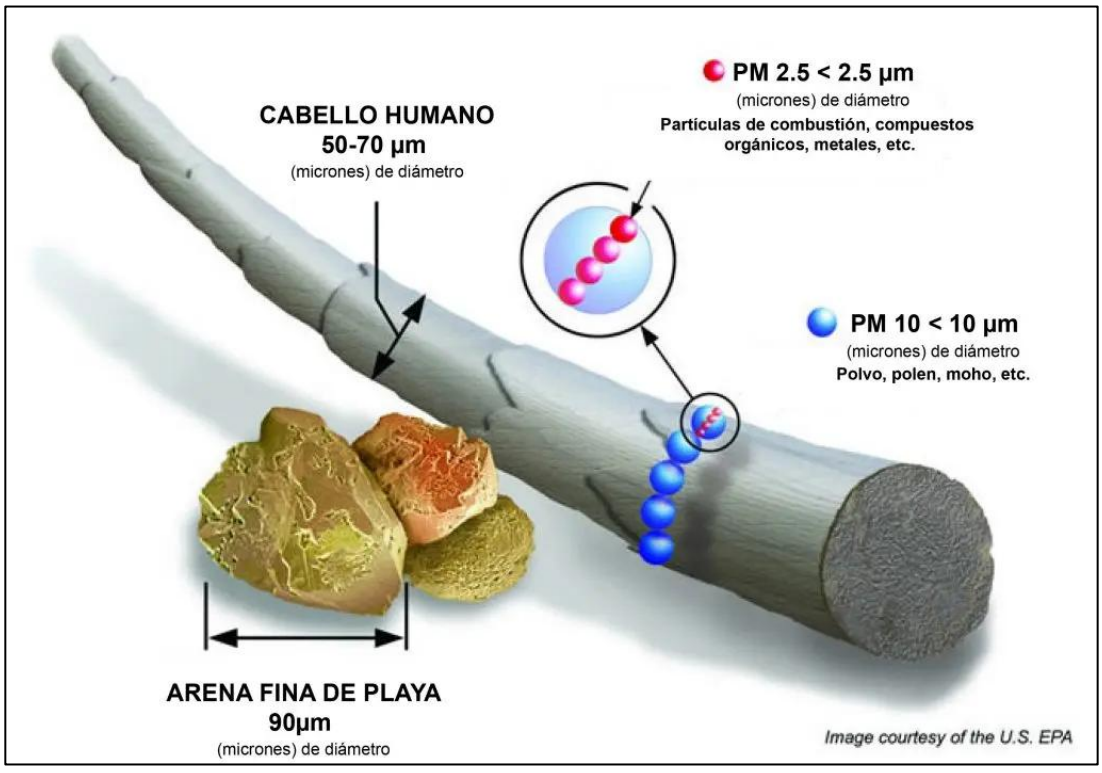
Se define como una mezcla de partículas sólidas y gotas líquidas que se ubican en el aire. Algunas partículas, como el polvo, la suciedad, el hollín, o el humo, son lo suficientemente grandes y oscuras como para observarlas a simple vista. Otras son tan pequeñas que solo pueden detectarse mediante el uso de un microscopio electrónico (EPA, 2025).

2.2.12.3. Material Particulado Fino $PM_{2.5}$.

Estas partículas miden 2.5 micras o menos, estas solo pueden vistos mediante el uso de un microscopio eléctrico, la contaminación por $PM_{2.5}$ es una de las mayores amenazas para la salud, debido a su tamaño y porque puede permanecer suspendido en el aire durante largos periodos de tiempo (Iqair, 2022).

Figura 1.

Material Particulado, comparación de tamaño



Nota. (Purioperú, 2021)

2.2.12.4. Fuente de Emisión de Material Particulado.

Algunas se emiten directamente desde una fuente, como obras en construcción, caminos sin asfaltar, campos, chimeneas o incendios (EPA, 2025).

La mayoría de las partículas se forman en la atmósfera como resultado de reacciones complejas de químicos, como el dióxido de azufre y los óxidos de nitrógeno, que son contaminantes emitidos por centrales eléctricas, industrias y automóviles (EPA, 2025).

2.2.12.5. Efectos de Salud.

La exposición prolongada a altos niveles de $PM_{2.5}$ está asociada con enfermedades respiratorias crónicas, como el asma, la enfermedad pulmonar obstructiva crónica (EPOC), y otras enfermedades cardiovasculares, además de aumentar el riesgo de cáncer de pulmón (Dockery et al., 1993).

2.2.12.6. Monitoreo y Medición de Datos.

2.2.12.6.1. Temperatura y Humedad Relativa.

Para la obtención de estos datos se emplea un termo-higrómetro protegido contra radiación solar. Un sensor Termohigrómetro es un dispositivo de medición que combina dos funciones: la temperatura y la humedad relativa, su utilidad principal es monitorear las condiciones climáticas en diversos espacios para asegurar un ambiente confortable y saludable (Paz, 2025).

a. Tipo de Termohigrómetros

Salazar (2023) Menciona que en el mercado existen diversos tipos de termohigrómetros, cada uno cuenta con características distintas y aplicaciones específicas, lo cual le permite adaptarse y contar con diversas funcionalidades:

- **Analógicos:** Son dispositivos mecánicos que emplean elementos sensibles a la temperatura y la humedad relativa, cuentan con un bimetálico en espiral para medir la temperatura y un higrómetro de cabello o de espiral para medir la humedad, sin embargo, los termohigrómetros analógicos suelen ser menos precisos y cuentan con una menor cantidad de funciones en comparación a los termohigrómetros digitales.
- **Digitales:** Emplean sensores eléctricos para medir la temperatura y la humedad relativa, cuenta con una pantalla digital donde muestra las mediciones de forma numérica, lo cual simplifica la lectura y la interpretación de datos, suelen ser precisos y cuenta con más funciones que los analógicos.

- **Inalámbricos:** Emplean una tecnología de comunicación sin cables para informar las mediciones de la temperatura y la humedad relativa a un dispositivo receptor, lo cual posibilita supervisar y controlar los parámetros ambientales a distancia, son útiles donde se requiere monitorear diversas áreas o lugares con difícil acceso.
- **Con sonda:** Emplea sensores de temperatura y humedad montados en una sonda, lo cual facilita medir las condiciones ambientales en lugares de difícil acceso, y pueden ser analógicos o digitales.

2.2.12.6.2. Velocidad y Dirección del Viento.

Para contar con estos datos es a través de un anemómetro y veleta de alta sensibilidad.

- **Anemómetro:** Es un equipo detector de viento, que reconoce la velocidad, la dirección y la precisión, el anemómetro analiza los cambios en las propiedades del fluido (Caneppele, 2022).
- **Tipos de Anemómetros**: Niubol (2024), menciona cuales son los tipos de anemómetros:
  - ✓ **Anemómetro de cazoletas:** Cuenta con tres o cuatro cazoletas que rotan sobre su eje, su velocidad de rotación es proporcional a la velocidad del viento, son empleados en estaciones meteorológica por su precisión, fiabilidad y sencillez.
  - ✓ **Anemómetro de veleta:** Esta emplea una veleta para medir la dirección y la velocidad del viento, la veleta se conduce hacia el viento, a la vez que la hélice gira a una velocidad proporcional a la velocidad del viento y es ampliamente ocupada en aviación meteorológica.
  - ✓ **Anemómetro de hilo caliente:** Mide la velocidad del viento a través de un alambre delgado calentado eléctricamente, cuando el viento circula a través del alambre este se enfría y la resistencia eléctrica varia, esto permite calcular la velocidad del viento. Este anemómetro es sensible y miden vientos de baja velocidad con precisión.

2.2.12.6.3. Presión Atmosférica.

Son captados mediante un barómetro digital integrado. Los barómetros digitales miden y muestran datos atmosféricos complejos con precisión y rapidez, la

mayoría de barómetros digitales muestran lecturas barométricas actuales como las lecturas previas de hace doce horas en un formato grafico de barras, asimismo tienen en cuenta lecturas atmosféricas como la humedad y el viento (Turgeon, 2023).

2.2.12.6.4. Precipitación.

Su medida es mediante un pluviómetro de balancín, el cual mide de manera precisa la cantidad de lluvia cada durante un balancín de vaciado automático (Darrera, 2025).

2.2.12.6.5. Nubosidad y Radiación Solar.

Son obtenidas mediante sensores ópticos adicionales, cuando estén disponibles. Estos sensores basan su funcionamiento en la emisión de un haz de luz que es reflejado por el objeto a detectar (Leuze, 2025).

2.2.12.6.6. Imputación de Datos Faltantes.

Los datos faltantes son observaciones sobre las cuales se posee la ausencia de determinada información en las variables de una base de datos. La existencia de los valores perdidos suele generar dudas acerca de la calidad del conjunto de datos, aumenta la probabilidad de cometer errores tipo I y II, reduce la potencia estadística y limita la fiabilidad de los intervalos de confianza (Statics, 2015).

2.2.13. Métodos de Imputación para Datos Faltantes

2.2.13.1. Métodos de Imputación Simple.

- **Imputación por media mediana y moda:**

Vaj (2024) menciona que la imputación por media en generalmente adecuada para datos numéricos que cuenten con una distribución normal, que tenga variables continuas y datos sin valores atípicos. La imputación mediana es ideal para datos numéricos con imputación que este distribuida y sesgada, variables continuas y datos con valores atípicos, la imputación por moda debe contar con datos categóricos y datos ordinales. Se tienen que tener en cuenta que estos métodos cuentan con limitaciones, la imputación por media y mediana puede introducir sesgo si los valores faltantes no faltan al azar y la imputación por moda puede no ser la adecuada si no existe una moda representativa de los valores faltantes.

- **Imputación con un valor constante:**

Baladram (2024) Indica que con este método se sustituyen todos los valores faltantes por una variable con un valor constante específico, este constante se puede escoger teniendo en cuenta el conocimiento del dominio o un valor predeterminado seguro, este enfoque solo se aplica cuando la característica es numérica.

- **Relleno hacia adelante y relleno hacia atrás:**

Baladram (2024) refiere que este método propaga el último valor conocido hacia adelante para completar los espacios vacíos, mientras que el método de relleno hacia atrás hace lo contrario, este método asume que el valor faltante probablemente sea similar al valor conocido más cercano.

2.2.13.2. Métodos de Imputación Avanzados.

- **Imputación por regresión:**

Mejía et al. (2024) señala que la imputación por regresión emplea un modelo de regresión para predecir los valores faltantes a partir de los valores observados de otras variables, es una técnica sencilla y rápida de implementar, sin embargo, puede introducir sesgos si los supuestos del modelo de regresión no se cumplen. Este método es ideal para variables numéricas que para variables categóricas.

- **Imputación múltiple:**

Mejía et al. (2024) indica que la imputación múltiple genera múltiples conjuntos de datos imputados, ajusta un modelo a cada uno y combina los resultados para lograr obtener estimaciones finales, es un método complejo que genera estimaciones con una menor cantidad de sesgos, principalmente cuando los datos no son datos faltantes completamente aleatorios (MCAR), se puede emplear tanto para variables numéricas o categóricas.

- **Miss forest imputación:**

Morgan (2020) menciona que es un algoritmo de aprendizaje automático basado en Random Forest que sirve para rellenar (imputar) datos faltantes o vacíos en una tabla de datos mixtos (numéricos y categóricos), repitiendo

un proceso de predicción por rondas hasta adivinar los valores perdidos con gran precisión.

- **Imputación Multivariada por ecuaciones encadenadas (MICE)**

  Azur et al. (2011) indica que es un método estadístico para rellenar datos faltantes en una base de datos. Creando múltiples valores de reemplazo mediante una serie de modelos de regresión secuenciales e iterativos, adaptándose de forma flexible a diferentes tipos de variables.

2.2.14. R Core Team

R es un entorno de software utilizado para el análisis estadístico y la representación gráfica de datos. (R Core Team, 2024)

CAPÍTULO III

METODOLOGÍA DE LA INVESTIGACIÓN

3.1. Ámbito de Estudio

Esta investigación se llevó a cabo en el distrito de Chanchamayo, ubicado en la provincia del mismo nombre, en la región Junín, Perú. Este distrito cuenta con una extensión de 919,72 km², con una población estimada de 30 117 habitantes para el año 2023 (INEI, 2023). Para el presente estudios se empleó dos estaciones las cuales se encuentran ubicados en las siguientes coordenadas: Estación la merced 463738.08 E y 8777746.87 N y la estación santa Isabel 461166.42 E y 8778900.17 N.

En relación con las condiciones geográficas y climáticas, Chanchamayo se sitúa a una altitud promedio de 751 m s. n. m., en la zona de selva alta peruana. Presenta un clima tropical húmedo, con temperaturas medias anuales entre 18 °C y 29 °C, y precipitaciones intensas durante gran parte del año, especialmente entre los meses de noviembre y marzo (SENAMHI, 2022). Estas condiciones hacen del distrito un territorio de alta productividad agrícola y con una biodiversidad relevante.

Figura 2.

Ubicación de lugar de estudio

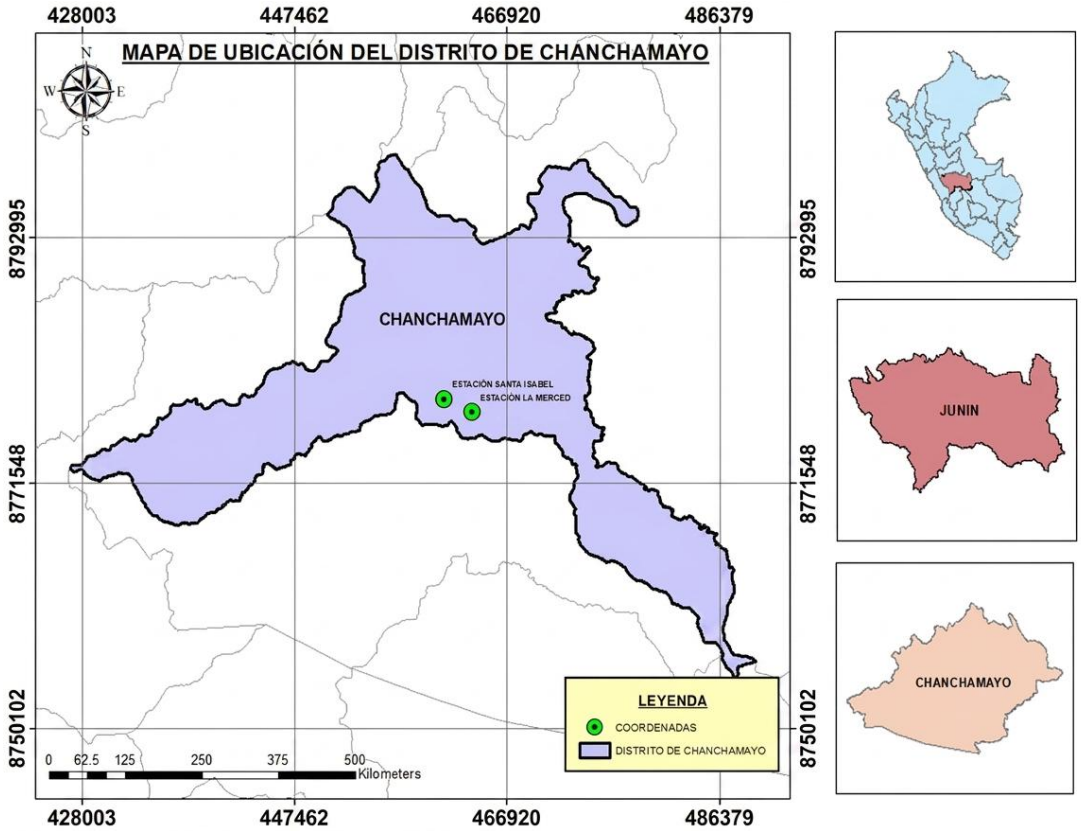


Figura 3.

Estación de monitoreo La Merced, Chanchamayo



3.2. Nivel, Tipo y Diseño de Investigación

3.2.1. Nivel

De acuerdo a Samaniego (2024), el nivel de la investigación es descriptivo - correlacional, descriptivo debido a que busca caracterizar el comportamiento del material particulado $PM_{2.5}$ y correlacional porque busca determinar el grado de asociación existente entra las variables meteorológicas (Temperatura del aire, Punto de Rocío, Velocidad del viento, Humedad relativa, Presión atmosférica, Precipitación, Radiación solar).

3.2.2. Tipo

El tipo de investigación es aplicada. Ñaupas et al. (2018) menciona que las investigaciones aplicadas buscan resolver problemas prácticos utilizando conocimientos ya existentes obtenidas por la investigación básica.

El presente estudio pretende aplicar métodos de imputación para solucionar los datos faltantes de material particulado de $PM_{2.5}$. Para ello, se emplean y evalúan algoritmos de imputación preexistentes que permitan reconstruir y predecir las series temporales de material particulado $PM_{2.5}$ y luego determinar su asociación con variables meteorológicas registradas durante el periodo 2025 en la ciudad de Chanchamayo.

3.2.3. Diseño

El diseño fue no experimental, porque los datos de material particulado $PM_{2.5}$ y las variables meteorológicas provienen de estaciones de monitoreo reales (La Merced y Santa Isabel) y responden a dinámicas climáticas y urbanas que escapan al control de un laboratorio. A ello, Haro et al. (2024), menciona que con este diseño se observa y analiza fenómenos en su contexto natural y sin manipulación de variables a lo largo del tiempo, permitiendo estudiar cambios y tendencias.

3.3. Población y Muestra

3.3.1. Población

La población comprende al conjunto de registros horarios del material particulado $PM_{2.5}$ y las variables meteorológicas asociados durante el periodo 2025. Estos datos fueron recolectados a través de la plataforma de la nube PurpleAir, mediante el uso de sensores de monitoreo de calidad del aire de bajo costo, específicamente el modelo Purpleair PA-II (que emplea el sensor PMS5003 para la medición de partículas), integrados con estaciones meteorológicas integradas Davis Vantage Pro, instalados dentro del distrito de Chanchamayo, provincia de Chanchamayo, región Junín.

Figura 4.

Equipos: Estación meteorológica Davis Vantage Pro



Figura 5.

Equipos: Sensor de monitoreo Purplair PA-II



3.3.2. Muestra

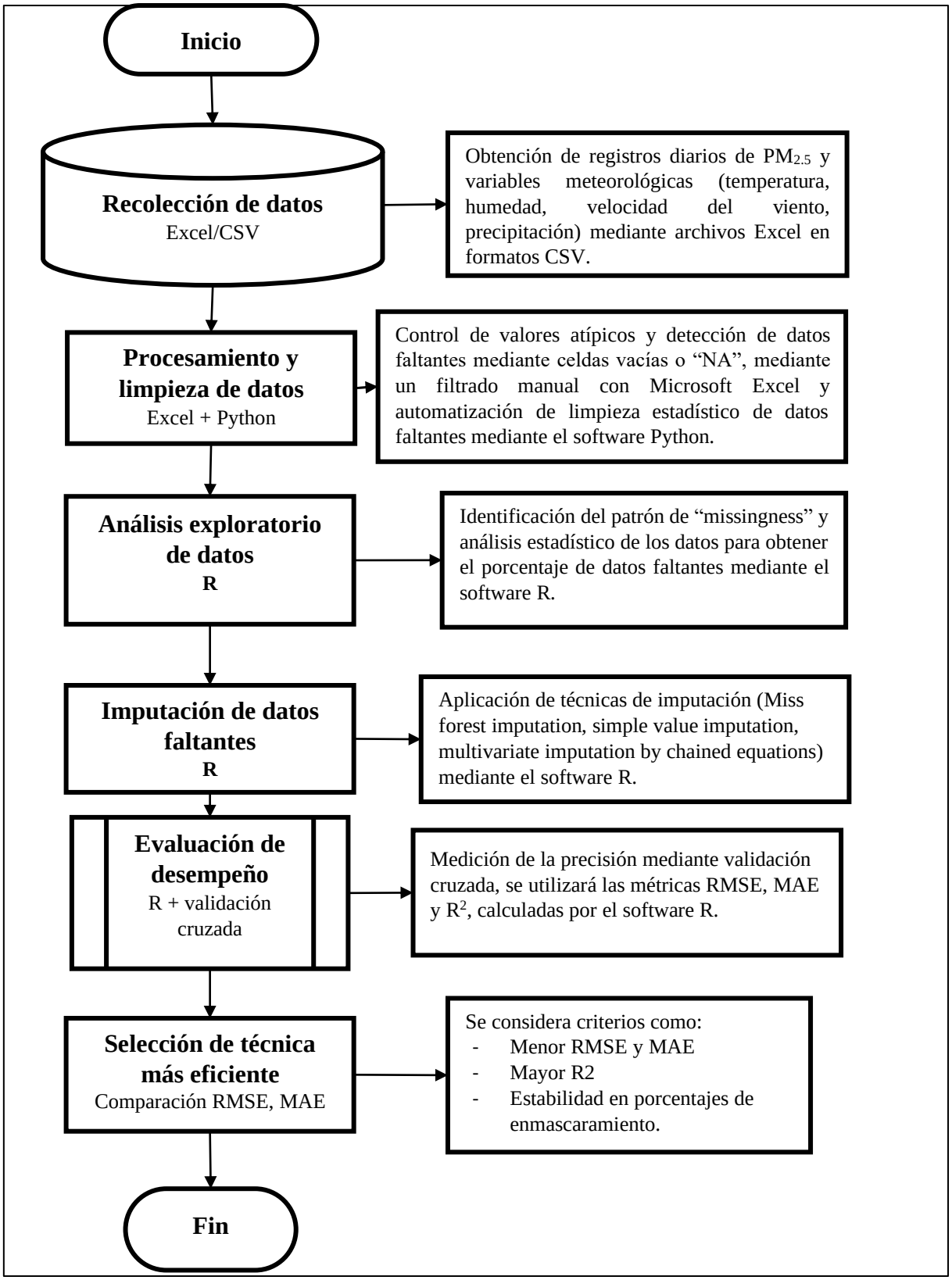
La muestra se refiere al sub conjunto de la población, consistente en registros diarios de $PM_{2.5}$ y variables meteorológicas asociadas (Temperatura del aire, Punto de Rocío, Velocidad del viento, Humedad relativa, Presión atmosférica, Precipitación, Radiación solar). La información fue recopilada mediante descarga y procesamiento de datos alojados en la plataforma PurpleAir, correspondientes a la estación ubicada en la zona urbana de la ciudad de Merced durante los meses desde enero hasta agosto y en la zona rural en el Fundo Santa Isabel durante los meses desde enero hasta setiembre, ambas del distrito de Chanchamayo, periodo 2025.

3.4. Procedimientos, Técnicas e Instrumentos de Recolección de Datos

3.4.1. Procedimiento

Figura 6.

Flujograma de procesos



Se desarrolló una estrategia metodológica compuesta por una serie de secuencias interrelacionadas lo cual conformaron el flujo integral del estudio, desde la adquisición de datos hasta la selección de la técnica de imputación más eficiente.

En esta **primera etapa** consistió en la recolección y el análisis de los registros diarios de la concentración de material particulado $PM_{2.5}$ y los parámetros meteorológicos (Temperatura del aire, Punto de Rocío, Velocidad del viento, Humedad relativa, Presión atmosférica, Precipitación, Radiación solar) registradas en el distrito de Chanchamayo durante el año 2025, a partir de la información recopilada por el equipo de estación meteorológica Davis Vantage Pro 2 y el sensor de material particulado PurpleAir modelo del sensor (PA-II y PMS5003), ubicadas en la ciudad de La Merced y en la zona rural de Chanchamayo. Los archivos fueron descargados en formatos CSV.

En la **segunda etapa** se realizó el procesamiento y limpieza de datos, para ello se utilizó el software Microsoft Excel donde al revisar los datos obtenidos se eliminaron los valores vacíos o inconsistentes. Asimismo, se procesó mediante el software R para unificar unidades de medida y cuantificar el porcentaje de datos faltantes por columna (Harris et al.,2020).

En la **tercera etapa** consistió en el análisis exploratorio de datos utilizando el Software R para caracterizar el patrón de la ausencia de información "missingness". Esto permitió visualizar la distribución temporal de los vacíos y la relación con las variables meteorológicas.

En la **cuarta etapa** se aplicó y evaluó las técnicas de imputación, mediante el software R para su posterior comparación de los cuáles serán aplicadas en las siguientes técnicas.

- **Miss Forest imputación**

  Se implementó este método basado en el algoritmo de Random Forest (Stekhoven & Bühlmann, 2012) para realizar imputaciones iterativas mediante árboles de decisión. Se seleccionó esta técnica debido a su capacidad de modelar relaciones no lineales y complejas entre las variables meteorológicas y el $PM_{2.5}$, sin requerir transformaciones previas de los datos. El procedimiento consistió en utilizar las variables completas como predictoras de las variables con valores faltantes, generando ensambles de árboles que permiten obtener imputaciones resistentes al overfitting y aptas para datos numéricos.

- **Imputación por valor simple (Simple Value Imputation)**

Esta técnica se utilizó como procedimiento complementario para variables que presentaron una incidencia mínima de datos faltantes, donde la complejidad de un modelo iterativo no era necesaria. El método consistió en la sustitución de los vacíos mediante el cálculo de la mediana de la serie temporal, garantizando una corrección computacionalmente eficiente y manteniendo la tendencia central de la variable tratada.

- **Imputación multivariada mediante ecuaciones encadenadas**

Se aplicó el método (Van Buuren & Groothuis-Oudshoorn, 2011) para los casos donde se requería preservar la estructura de correlación multivariada entre los sensores de calidad del aire y las variables climáticas. El proceso se ejecutó modelando cada variable con datos faltantes como una función de las demás, mediante un ciclo de imputaciones múltiples. Esta metodología permitió capturar la incertidumbre estadística y asegurar que las relaciones estructurales entre el material particulado y las variables asociadas se mantuvieran coherentes a lo largo de las distintas versiones de la base de datos imputadas.

La **quinta etapa** se desarrolló la evaluación de desempeño de las técnicas de imputación mediante el sistema de validación cruzada. Se cuantificó la precisión y estabilidad de cada método mediante la comparación de los datos imputados con los valores reales por medio de las métricas de error como RMSE, MAE y el coeficiente de determinación R^2 (Willmott & Matsuura, 2015). Para la implementación y evaluación de los modelos de imputación (MissForest y MICE), la serie de datos observados completos fue dividida bajo un esquema de partición del 70% para la fase de entrenamiento y 30% para la fase de validación.

Y en la última y **sexta etapa** se determinaron la técnica de imputación más eficiente para la reconstrucción de datos faltantes de $PM_{2.5}$, considerando tanto el ajuste estadístico (errores y correlaciones) como la preservación de la variabilidad y la estructura interna de las relaciones entre las variables meteorológicas y la calidad del aire en el distrito de Chanchamayo. Los resultados fueron representados mediante tablas y gráficos con Microsoft Excel.

3.4.2. Técnicas

El estudio se basó en la recolección y análisis de datos secundarios provenientes de estaciones de monitoreo ambiental y meteorológico localizadas en el distrito de Chanchamayo. Para el tratamiento de la información se utilizó la técnica de análisis estadístico multivariante, lo cual permitió explorar la asociación entre los valores de concentración de $PM_{2.5}$ y las variables meteorológicas registradas.

Por otra parte, se aplicó la técnica de imputación de datos faltantes, considerando métodos estadísticos como la imputación Miss Forest Imputation que está basada en el algoritmo de Random Forest, Simple Value Imputation y el paquete Multivariate Imputation by Chained Equations (MICE), a fin de estimar valores ausentes de la serie temporal de $PM_{2.5}$ (Little & Rubin, 2019). Asimismo, se desarrolló un análisis de correlación de Spearman para evaluar la relación entre la concentración de material particulado y variables meteorológicas.

3.4.3. Instrumentos

Para el desarrollo de esta investigación, se utilizaron como fuente primaria de información los registros horarios de material particulado ($PM_{2.5}$) y variables meteorológicas, obtenidos mediante la plataforma en la nube PurpleAir. La recolección de datos se llevó a cabo utilizando sensores de bajo costo, específicamente el modelo PurpleAir PA-II, los cuales incorporan el sensor láser PMS5003 para la medición precisa de partículas.

Estos sensores se encuentran integrados con estaciones meteorológicas Davis Vantage Pro 2 (modelo 6152), instaladas estratégicamente en el distrito de Chanchamayo. A través de este sistema, se obtuvo la información digital en formato CSV y hojas de cálculo (Excel), permitiendo la sistematización de registros históricos que comprenden las siguientes variables:

- Calidad del aire: Concentración horaria de $PM_{2.5}$.
- Variables meteorológicas: Temperatura del aire, velocidad del viento, humedad relativa y precipitación.

La integración de estos dispositivos permitió consolidar una base de datos secundaria robusta, adecuada para realizar el análisis de asociación entre las condiciones climáticas locales y los niveles de material particulado.

3.5. Análisis de Datos

El tratamiento y análisis de los datos se desarrolló en cuatro etapas principales: (1) preprocesamiento y control de calidad, (2) imputación de datos faltantes, (3) análisis descriptivo y exploratorio, y (4) modelado e inferencia estadística.

En la etapa de preprocesamiento, se verificó la consistencia temporal de las series de $PM_{2.5}$ y las variables meteorológicas (Temperatura del aire, Punto de Rocío, Velocidad del viento, Humedad relativa, Presión atmosférica, Precipitación, Radiación solar). Se detectaron y corrigieron registros atípicos mediante reglas basadas en rangos plausibles, estandarizando las unidades y marcas de tiempo (Montgomery & Runger, 2018).

Para la imputación de valores, se evaluaron las técnicas MissForest, Simple Value Imputation y Multivariate Imputation by Chained Equations (MICE). La elección del método se basó en métricas de error (RMSE y MAE) obtenidas mediante validación cruzada por ventanas temporales, asegurando la preservación de la estructura temporal y la varianza de la serie (Little & Rubin, 2019; Van Buuren, 2018).

El análisis descriptivo incluyó estadísticas básicas y visualizaciones temporales para caracterizar la variabilidad estacional y diaria del $PM_{2.5}$. Se emplearon coeficientes de correlación de Spearman, complementados con diagramas de dispersión y análisis de autocorrelación parcial.

Finalmente, en la etapa de modelado e inferencia estadística, se implementaron dos estrategias comparativas:

Modelos Lineales Generalizados (GLM): Se ajustaron modelos para cuantificar la relación entre la concentración de $PM_{2.5}$ y las variables meteorológicas. Inicialmente, se definió un modelo completo (full) considerando todas las variables disponibles. Posteriormente, se aplicó un procedimiento de separación de variables fundamentado en el Criterio de Información de Akaike (stepAIC), optimizando el modelo mediante una selección bidireccional (both) para identificar el subconjunto de variables con mayor significancia estadística, minimizando el sesgo y mejorando la capacidad explicativa del modelo (Refit).

Modelado de Machine Learning: Se implementaron modelos de Random Forest para capturar relaciones no lineales complejas, comparando su desempeño frente al GLM optimizado mediante métricas de error (RMSE, MAE) y el coeficiente de determinación (R^2).

Los métodos de imputación aplicados en esta investigación se basan en tres modelos estadísticos con diferentes niveles de complejidad (Tabla 2). El método más simple, la imputación por media, reemplaza los valores faltantes por el promedio aritmético de los datos observados. El método MICE utiliza modelos de regresión multivariada para predecir cada valor faltante a partir de las demás variables, generando múltiples conjuntos de datos imputados. Finalmente, el método MissForest emplea un enfoque no paramétrico basado en Random Forest, que elabora un conjunto de árboles de decisión para estimar los valores ausentes.

Tabla 2.

Modelos estadísticos aplicados en la imputación de datos faltantes de $PM_{2.5}$

Método	Modelo matemático	¿Qué hace?
Imputación por Media Vaj (2024)	$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} yi$	Reemplaza cada valor faltante por el promedio de todos los datos observados.
MICE (Imputación Múltiple) Azur et al. (2011)	$yi = \beta_0 + \beta_{1y1} + \cdots + \beta_{pYp} + \varepsilon$	Predice cada valor faltante usando un modelo de regresión con las otras variables. Genera 5 conjuntos de datos y los promedia.
MissForest (Random Forest) Morgan (2020)	$\hat{Y} = \frac{1}{B}\sum_{b=1}^{B} T_b(X)$	Construye 300 árboles de decisión para predecir cada valor faltante y promedia sus predicciones.

Nota. y= valor de $PM_{2.5}$; n = número de datos observados; β = coeficientes de regresión; ε = error aleatorio; Tb = árbol de decisión; B = número de árboles (300).

Todo el procesamiento y los análisis se ejecutaron en el entorno R, utilizando scripts reproducibles y fijación de semillas aleatorias para garantizar la replicabilidad del estudio (R Core Team, 2023; Pedregosa et al., 2011).

3.6. Consideraciones Éticas

Esta investigación se realizó con integridad académica, asegurando el uso responsable y ético de los instrumentos y datos empleados. No se involucraron seres humanos ni se vulneraron derechos personales. Los datos meteorológicos y de calidad del aire fueron recolectados objetivamente, sin manipulación, y se respetaron los principios de veracidad, transparencia y respeto al medio ambiente. Además, se trabajó adecuadamente con todas las citas y fuentes empleadas, cumpliendo con los lineamientos éticos establecidos por la institución.

CAPÍTULO IV

RESULTADOS Y DISCUSIONES

4.1. Presentación de Resultados y Discusiones

4.1.1. Identificación de Patrones de Datos Faltantes en las Series Temporales de $PM_{2.5}$

4.1.1.1. Estación La Merced.

La Figura 7 muestra la variación temporal de las concentraciones de $PM_{2.5}$ y su distribución de los datos faltantes entre el periodo de enero y agosto del año 2025. Su comportamiento se caracteriza por mantener valores desde bajos a moderados durante su mayor parte del tiempo, interrumpidos por la presencia de picos de alta concentración. Esta dinámica es consistente con lo planteado por Moreno (2024), quien sostiene que las variables ambientales asociadas a fenómenos meteorológicas presentan una elevada variabilidad temporal debido a la sensibilidad de factores contextuales y las actividades inesperadas que se presentan.

A inicios de febrero se registró un valor máximo, alcanzando una concentración cercana a los 140 µg/m^3, mientras que entre finales de mayo y principios de junio se presentó un segundo evento crítico que supero los 90 ug/m^3. Por otra parte, durante los meses de marzo y abril los valores se mantuvieron estables y por debajo de los 20 ug/m^3. De acuerdo con Cáceres (2018) y Choquetico y Silva (2024) señalan que el material particulado fino responde de manera sensible a la interacción con variables meteorológicas. También Motocanche y Lazo (2020) destacan que las condiciones atmosféricas desfavorables como la estabilidad del aire o la baja ventilación pueden intensificar los procesos de acumulación del material particulado, lo que se refleja similares a estos picos identificados en el estudio realizado.

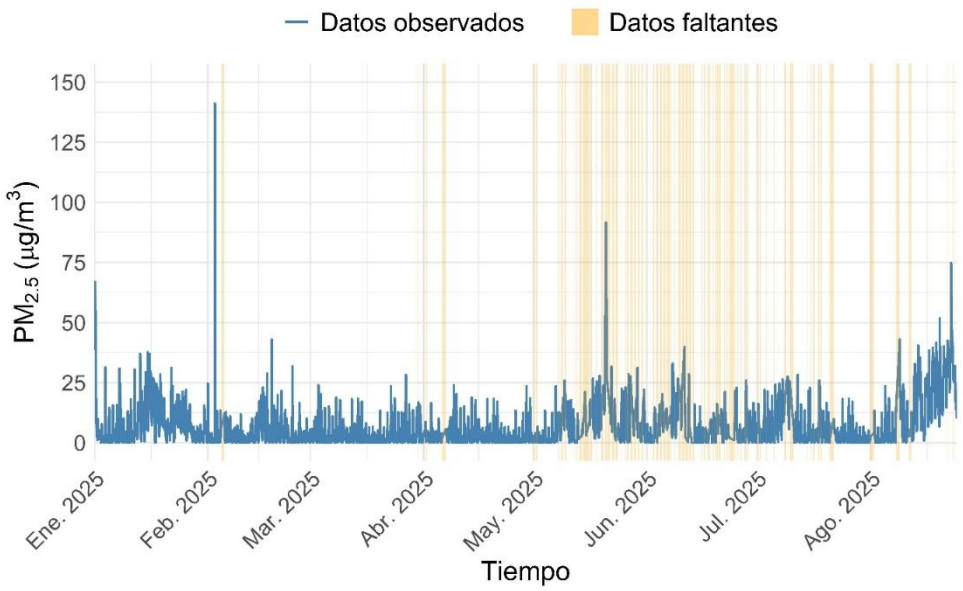
Por otro lado, la distribución temporal de los datos faltantes presenta un patrón heterogéneo. Y entre enero y abril se evidencia una cobertura casi continua con interrupciones mínimas y de corta duración; sin embargo, a partir de mayo y con mayor severidad entre junio y julio se registra un incremento en la frecuencia y duración de la data faltante de $PM_{2.5}$. Como señala Moreno (2024), la discontinuidad en las series temporales continuas representa una restricción metodológica importante porque provocará que se reduzca la capacidad analítica y genera incertidumbre al caracterizar de manera completa la magnitud y frecuencia del monitoreo realizado.

Los picos de concentración más elevados de material particulado de $PM_{2.5}$ registrados en la serie temporal obedecen a distintas dinámicas temporales y antrópicas. Por un lado, el pico máximo registrado a inicios de año (enero y febrero), lo cual concuerda con lo reportado

por Camayo & Canturin (2021), quienes identificaron que el aumento de concentración en este periodo es consecuencia de actividades festivas y celebraciones de Año Nuevo, caracterizadas por el uso de pirotecnia. Por otro lado, el comportamiento observado durante el periodo de mayo a junio donde se aprecia un pico superior lo cual coincide con el desarrollo de la estación seca en la región. Este periodo se debe a la evidencia de una menor precipitación pluvial y una mayor estabilidad atmosférica que dificultan la dispersión de contaminantes, adicionando al incremento de fuentes de emisión locales como la quema de biomasa, actividades agrícolas de rozado y el polvo del suelo. Asimismo, la serie temporal de datos observados revela un pico notable de contaminación a fines de agosto del periodo monitoreado. Este incremento se debe a la convergencia de factores antrópicos y climáticos. Por un lado, marca el inicio de las actividades festivas, eventos que suelen incrementar el tráfico, el comercio y actividades puntuales generadoras de partículas. Por otro lado, este periodo coincide con el final de la estación seca, momento crítico donde se intensifica la quema de biomasa y pastizales en las zonas de ladera y montañas que rodean el casco urbano, práctica agrícola que, combinada con la escasa dispersión atmosférica propia de la época, eleva significativamente las concentraciones de material particulado de $PM_{2.5}$.

Figura 7.

Comportamiento del $PM_{2.5}$ y representatividad de los datos en la estación La Merced, Chanchamayo, 2025



La Tabla 3 presenta un resumen estadístico mensual de las concentraciones de material particulado fino ($PM_{2.5}$) registradas entre enero y agosto durante el año 2025 en la estación de monitoreo atmosférico de La Merced, provincia de Chanchamayo. Estás incluyen medidas de tendencia central, dispersión, valores extremos y porcentaje de cobertura, lo que permite

evaluar tanto el comportamiento temporal del contaminante como la representatividad de la serie de datos.

Se registró una concentración media de 6.12 ug/m^3 y una mediana de 3.2 ug/m^3, con una desviación estándar total de 8.2 ug/m^3. En el mes de agosto registró los niveles promedio más elevado del estudio, alanzando una media de 14.83 ug/m^3 y una mediana de 12.5 ug/m^3, acompañando de una alta dispersión de 12.26 ug/m^3.

Respecto a los valores extremos, en enero se evidenció el nivel máximo crítico de todo el registro, alcanzando los 161.8 ug/m^3, seguido por febrero con un pico de 141.2 ug/m^3. Estos datos en los primeros meses del año concuerdan con la alta volatilidad observada en enero. Por otro lado, se obtuvieron datos de mayo 91.7 ug/m^3 y agosto 74.8 ug/m^3 que mostraron picos significativos, y en los meses de marzo que obtuvo 28.3 ug/m^3, abril con 24.0 ug/m^3 y julio con 28.3 ug/m^3, es donde se evidenció más techos de contaminación.

Asimismo, respecto al monitoreo continuo, se obtuvo una representatividad total del 83.4%. Sin embargo, se consideró que la distribución mensual de los vacíos de datos fue heterogénea. En cuanto al mes de enero que mostró una pureza del 99.9%, donde entre mayo y junio se concentraron la mayor pérdida de información, alcanzando porcentajes de datos faltantes del 39.4% con 293 datos vacíos y el 47.1% con 339 datos vacíos respectivamente.

Tabla 3.

Resumen estadístico mensual de concentraciones de $PM_{2.5}$ medidos en la estación de La Merced, Chanchamayo, 2025

Mes	***Media***	***Mediana***	***Valor Mínimo***	***Valor Máximo***	***Desviación Estandar***	***Datos Faltantes***	***Datos Observados***
Ene	8.42	5.10	0	161.80	10.71	1 (0.1%)	743 (99.9%)
Feb	3.99	1.95	0	141.20	7.37	23 (3.4%)	649 (96.6%)
Mar	3.12	1.60	0	28.30	4.11	8 (1.1%)	739 (98.9%)
Abr	2.80	1.50	0	24.00	3.67	35 (4.9%)	685 (95.1%)
May	6.16	4.37	0	91.70	6.59	293 (39.4%)	451 (60.6%)

Jun	5.67	4.16	0	39.90	5.63	339 (47.1%)	381 (52.9%)
Jul	5.72	3.10	0	28.30	6.59	154 (19.4%)	640 (80.6%)
Ago	14.83	12.50	0	74.80	12.26	92 (16.2%)	475 (83.8%)
TOTAL	**6.12**	**3.20**	**0**	**161.80**	**8.20**	**945 (16.6%)**	**4763 (83.4%)**

Los resultados evidencian que, el área de estudio no presenta concentraciones elevadas de $PM_{2.5}$, sí se evidencia episodios puntuales de alta contaminación, los cuales están asociados a fuentes locales y condiciones meteorológicas específicas como en inicio de las actividades festivas y el inicio de quema de pastizales o rozo por el fin de la temporada seca. En consecuencia, el análisis mensual del $PM_{2.5}$ en la estación La Merced revela un comportamiento estacional con concentraciones más elevadas durante la época seca, alta variabilidad en el desempeño parcial de cada uno de los meses y presencia de eventos extremos.

4.1.1.2. Estación Santa Isabel

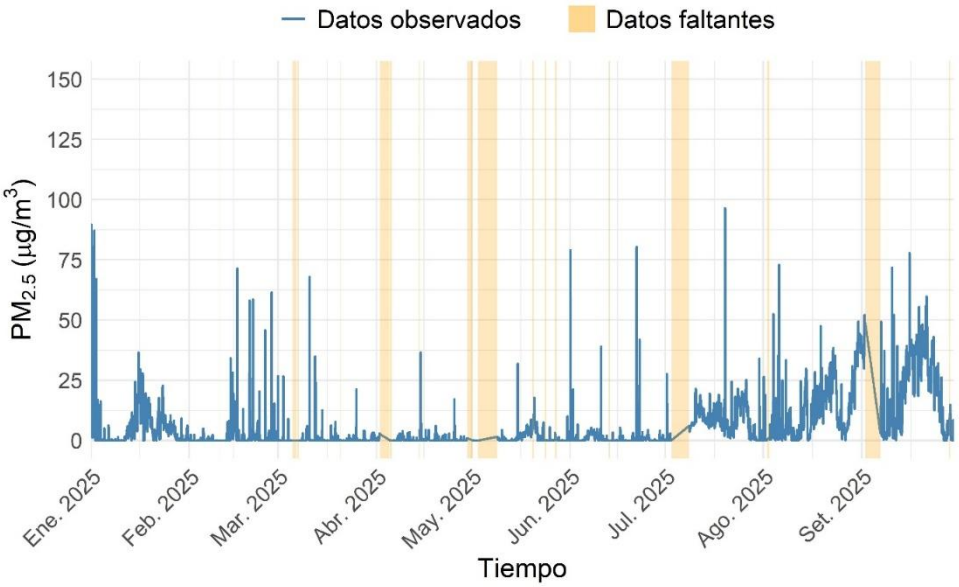
La Figura 8 evidencia el comportamiento temporal de las concentraciones de $PM_{2.5}$ ($\mu g/m^3$) y los periodos con vacíos de los datos faltantes registradas en la estación de Santa Isabel ubicado en la zona rural del distrito de Chanchamayo desde el mes de enero hasta septiembre de año 2025. Se evidencia una alta variabilidad temporal, con concentraciones bajas durante gran parte del periodo, interrumpidas por episodios puntuales de incremento significativo.

Lo más resaltante son los episodios delicados de contaminación, representados por picos repentinos que superan significativamente los niveles habituales, estas mismas resaltan en tres momentos principales en donde a inicios de enero se registró un resultado que alcanzó los 90 $\mu g/m^3$, seguido por fluctuaciones elevadas en marzo con picos superiores a los 70 $\mu g/m^3$. De igual manera durante junio y julio se presentaron eventos que bordearon los 80 $\mu g/m^3$, sin embargo, el máximo resultado del estudio se identificó durante el mes de agosto, con un pico que alcanza los 100 $\mu g/m^3$. Por lo contrario, los meses de febrero, abril y mayo mostraron una atmósfera más estable y limpia identificando concentraciones que se mantuvieron de manera uniforme por debajo de los 10 $\mu g/m^3$, y a partir de agosto no solo se identificó los picos superiores a los 75 $\mu g/m^3$, sino que la concentración basal diaria sufrió un aumento que altera el promedio del contaminante.

Asimismo, la calidad de la serie temporal se representa a través de las franjas amarillas la distribución de los datos faltantes de material particulado $PM_{2.5}$. Es así como se identificó las interrupciones más prolongadas evidenciadas en los meses de marzo, abril, mayo, julio y mediados de septiembre, estas implicarían que la magnitud real de algunos eventos de contaminación podría ser incluso mayor o que se perdieron registros de tendencias en esos periodos. A pesar de esto, la evidencia apunta a una atmósfera mucho más cargada de partículas hacia el final del periodo evaluado.

Figura 8.

Comportamiento del $PM_{2.5}$ y representatividad de los datos en la estación Santa Isabel, Chanchamayo, 2025



La Tabla 4 presenta el resumen estadístico mensual de las concentraciones de $PM_{2.5}$ registradas durante el año 2025 en la estación meteorológica de Santa Isabel, ubicada en la provincia de Chanchamayo. Los resultados evidencian una variabilidad temporal del contaminante, con concentraciones bajas durante gran parte del año y un incremento significativo hacia los meses de la estación seca.

Las concentraciones medias mensuales se mantuvieron por debajo de 2 µg/m^3 entre febrero y junio, alcanzando valores mínimos en marzo y abril, con medias de 0.79 µg/m^3 y 0.61 µg/m^3, respectivamente. Estos bajos niveles, junto con medianas cercanas a cero, indican que predominan condiciones atmosféricas favorables para la dispersión de contaminantes, así como una limitada influencia de fuentes locales de emisión.

A partir de julio se observa un cambio notable en el comportamiento del $PM_{2.5}$, con un incremento progresivo de las concentraciones medias, que alcanzan 8.17 µg/m^3 en julio, 12.86

µg/m^3 en agosto y un máximo anual de 20.56 µg/m^3 en septiembre. Este patrón sugiere una fuerte influencia estacional, probablemente asociada a la disminución de precipitaciones, mayor estabilidad atmosférica y aumento de actividades emisoras como la quema de biomasa y el tránsito vehicular, características de la época seca en la región.

La diferencia entre la media y la mediana en varios meses, especialmente en enero, febrero y junio, revela distribuciones positivas a consecuencia de episodios puntuales de altas concentraciones. Esto se refleja en los valores máximos registrados que alcanzan hasta 96.4 µg/m^3 en julio y 89.6 µg/m^3 en enero, muy por encima de los niveles habituales y de los valores diarios recomendados para la protección de la salud.

La desviación estándar presenta valores de mayor concentración durante los meses monitoreados, lo que indica una alta variabilidad en el desempeño mensual y la ocurrencia de eventos de contaminación aguda. Por otra parte, durante los meses con bajas concentraciones medias, la dispersión es reducida, lo que se considera como un comportamiento más estable del contaminante.

En cuanto a la cobertura de datos, la estación de Santa Isabel muestra un alto nivel de integridad, con el 89.9% de observaciones válidas a lo largo del año. Las mayores pérdidas de información se registraron entre abril y mayo, con porcentajes de datos faltantes del 18.1% y 25.8%, respectivamente; sin embargo, estos vacíos no comprometen la representación del análisis anual.

El promedio anual de $PM_{2.5}$ fue de 5.81 µg/m^3, valor cercano y ligeramente superior al valor anual recomendado por la OMS, lo que indica que, si bien la calidad del aire en esta estación es generalmente buena, existen periodos específicos del año en los que la exposición a $PM_{2.5}$ podría representar un riesgo potencial para la salud de la población.

Tabla 4.

Resumen mensual de $PM_{2.5}$ en la estación de Santa Isabel, Chanchamayo, 2025

Mes	***Media***	***Mediana***	***Mínimo***	***Máximo***	***SD***	***Faltantes***	***Observados***
Ene	5.40	2.25	0	89.60	9.33	0 (0%)	744 (100%)
Feb	1.63	0.00	0	71.50	6.00	2 (0.3%)	670 (99.7%)
Mar	0.79	0.00	0	67.90	3.61	46 (6.2%)	698 (93.8%)

Abr	0.61	0.10	0	36.50	1.84	130 (18.1%)	590 (81.9%)
May	1.27	0.50	0	31.80	2.27	192 (25.8%)	552 (74.2%)
Jun	1.20	0.00	0	80.40	5.56	8 (1.1%)	712 (98.9%)
Jul	8.17	7.86	0	96.40	6.47	133 (17.9%)	611 (82.1%)
Ago	12.86	8.65	0	73.00	12.09	24 (3.2%)	720 (96.8%)
Set	20.56	19.75	0	77.80	14.59	123 (17.7%)	573 (82.3%)
TOTAL	**5.81**	**0.99**	**0**	**96.40**	**10.27**	**658 (10.1%)**	**5870 (89.9%)**

4.1.2. Análisis de Patrones de Datos Faltantes en las Series Temporales de $PM_{2.5}$

4.1.2.1. Estación Santa Isabel

La Tabla 5 presenta la evaluación de los criterios técnicos para determinar la mejor técnica al aplicar métodos de imputación a la serie temporal de concentraciones de $PM_{2.5}$ registradas en la estación Santa Isabel durante el año 2025. Este análisis es fundamental para garantizar que la imputación no introduzca sesgos significativos ni distorsione la variabilidad temporal al contaminante.

En primer lugar, la proporción total de datos faltantes es del 10%, valor que se encuentra ampliamente por debajo del umbral máximo del 30% comúnmente aceptado en estudios de calidad del aire para permitir la reconstrucción de series temporales mediante técnicas estadísticas. Este resultado indica que la mayor parte de la información observada es representada sobre el comportamiento real del $PM_{2.5}$, proporcionando una base sólida para la estimación de los valores faltantes.

Respecto a la continuidad temporal de los datos, se identificaron brechas máximas de hasta 4 días consecutivos sin registros, lo que implica el incumplimiento formal del criterio que establece la ausencia de vacíos mayores a 48 horas. No obstante, desde una perspectiva técnica, la presencia de brechas relativamente cortas y puntuales no invalida necesariamente el proceso de imputación, especialmente cuando estas interrupciones no se concentran en periodos críticos del año ni coinciden con episodios extremos de contaminación. En operaciones de monitoreo ambiental, este tipo de discontinuidades puede ser abordado de manera robusta mediante

modelos que incorporan información meteorológica y patrones temporales, reduciendo el impacto potencial de dichos vacíos sobre la serie reconstruida. Arroyave, Villareal et al., (2023).

Adicionalmente, la inexistencia de meses completos sin datos corresponde a un aspecto clave que refuerza la validez de la imputación.

Tabla 5.

Criterios para permitir la imputación de $PM_{2.5}$ medidos en la estación de Santa Isabel, Chanchamayo, 2025

Criterio	Valor	%	Cumple
Proporción de datos faltantes ≤ 30%	0.10	10%	SI
Sin brechas mayores a 48 horas	4.00	100%	NO
Sin meses completos faltantes	0.00	0%	SI
Imputación permitida	0.00	0%	SI

En la Tabla Nº 6 se muestran los resultados de la evaluación que indican diferencias en el desempeño de los métodos de imputación de $PM_{2.5}$. El método de imputación por la "Media" presenta valores de RMSE = 10.722 µg/m^3 y MAE = 7.048 µg/m^3, lo que evidencia una capacidad limitada para reproducir la variabilidad real de las concentraciones observadas. Al reemplazar los valores faltantes por un único valor promedio, este método tiende a suavizar la serie temporal y a subestimar tanto los picos como las concentraciones bajas, lo que reduce su utilidad donde la variabilidad temporal es relevante.

El método basado en "Random Forest" muestra el mejor desempeño global, con RMSE = 8.854 µg/m^3, MAE = 4.852 µg/m^3 y un R^2 = 0.320. Estos resultados indican que el modelo es capaz de capturar una fracción significativa de la variabilidad de $PM_{2.5}$ a partir de los covariables temporales (hora del día y mes), logrando estimaciones más cercanas a los valores reales. El menor error absoluto corresponde a un enfoque más robusto frente a valores extremos y representa adecuadamente patrones no lineales en la serie. Bussi, Hernandez, et al., (2018) menciona que este método presenta valores promedios de RMSE menores y un tiempo de procesamiento superior a otros métodos empleados.

Por su parte, el método "MICE" presenta un desempeño intermedio, con RMSE = 11.076 µg/m^3, MAE = 5.543 µg/m^3 y R^2 = 0.163. Aunque supera a la imputación por la media en términos de error absoluto y capacidad explicativa, su rendimiento es inferior al del

"Random Forest". Esto se atribuye a que el conjunto de predictores utilizados es limitado ya que el método de *predictive mean matching* no captura completamente las estructuras temporales complejas presentes en los datos de $PM_{2.5}$.

Tabla 6.

Evaluación de métodos de imputación de $PM_{2.5}$ medidos en la estación de Santa Isabel, Chanchamayo, 2025

Método	RMSE	MAE	R^2
Media	10.722	7.048	NA*
Random Forest	8.854	4.852	0.320
MICE	11.076	5.543	0.163

*El valor **R^2 = NA** indica que **no puede evaluarse la capacidad explicativa** del método

4.1.2.2. Estación La Merced

La Tabla 7 presenta la evaluación de los criterios técnicos para determinar la mejor técnica al aplicar métodos de imputación a la serie temporal de concentraciones de $PM_{2.5}$ registradas en la estación de La Merced durante el año 2025. Los resultados evidencian un escenario más favorable, lo que refuerza la confiabilidad de cualquier proceso de reconstrucción de valores faltantes.

La proporción de datos ausentes alcanza el 17%, manteniéndose dentro de los límites aceptables para estudios ambientales de series temporales continuas. Este nivel de pérdida de información no compromete la representación estadística de la serie ni la capacidad de los modelos de imputación para capturar la variabilidad temporal del $PM_{2.5}$. Naranjo & Escudero (2025) destacan que los métodos de imputación son necesarios para asegurar que la calidad de los resultados sea más eficiente.

La ausencia total de brechas superiores a 48 horas es relevante para el estudio, ya que de esta manera es posible determinar la confiabilidad de los métodos de imputación aplicados. La continuidad temporal garantiza que la estructura dinámica de corto plazo del contaminante se conserve adecuadamente, lo que resulta crucial para la imputación de episodios transitorios y para la correcta reproducción de eventos de incremento en las concentraciones. Desde un punto de vista técnico, esta condición reduce significativamente la incertidumbre asociada a la estimación de valores faltantes.

Asimismo, la inexistencia de meses completos sin información asegura la preservación del patrón estacional anual del $PM_{2.5}$, elemento indispensable para análisis de tendencia, evaluación de cumplimiento normativo y comparación intermensual, esto quiere decir que la cobertura temporal homogénea a lo largo del año contribuye a que la imputación no introduzca distorsiones sistemáticas en los promedios mensuales o anuales.

En conjunto, el cumplimiento íntegro de los criterios evaluados sustenta de manera sólida la decisión de permitir la imputación de los datos de $PM_{2.5}$ en la estación de La Merced. Bajo estas condiciones, la imputación no solo es técnicamente admisible, sino que se espera que genere una serie temporal coherente y robusta. Naranjo & Escudero (2025) menciona que la imputación de datos faltantes permite obtener resultados más confiables para la investigación.

Tabla 7.

Criterios para permitir la imputación de $PM_{2.5}$ medidos en la estación de La Merced, Chanchamayo, 2025

Criterio	Valor	%	Cumple
Proporción de datos faltantes ≤ 30%	0.17	17%	SI
Sin brechas mayores a 48 horas	0.00	0%	SI
Sin meses completos faltantes	0.00	0%	SI
Imputación permitida	1.00	100%	SI

En la Tabla 8 muestra la evaluación de los métodos de imputación aplicados a la serie de $PM_{2.5}$, se evidencia diferencias claras en su capacidad para reproducir los valores reales y la variabilidad temporal de las concentraciones. En el caso de la imputación por la "Media", los errores obtenidos son relativamente elevados, con un RMSE de 8.403 y un MAE de 5.913, lo que indica desviaciones respecto a los valores observados. El coeficiente de determinación R^2 no pudo ser calculado y aparece como NA, debido a que este método genera valores constantes durante la imputación, eliminando la variabilidad necesaria para estimar una correlación con los datos reales. En consecuencia, aunque la imputación por la media resulta sencilla y rápida de aplicar, su capacidad para representar adecuadamente el comportamiento real del contaminante es limitada y no logra capturar los patrones temporales presentes en la serie.

El método de "Random Forest" presenta el mejor desempeño entre las técnicas evaluadas. Este método alcanza los menores valores de error, con un RMSE de 6.945 y un MAE

de 4.659, lo que refleja una mayor precisión en la estimación de los datos faltantes. Además, el valor de R^2 obtenido, igual a 0.319, indica que el modelo es capaz de explicar aproximadamente el 32 % de la variabilidad de los valores reales durante el proceso de validación. Este resultado sugiere que el uso de variables temporales como la hora del día y el mes permite al algoritmo capturar relaciones no lineales y patrones horarios y estacionales característicos de las concentraciones de $PM_{2.5}$, ofreciendo una representación más realista del comportamiento del contaminante.

Por su parte, el método "MICE" muestra un desempeño inferior en comparación con los otros métodos de imputación. Los errores asociados a esta técnica son los más altos, con un RMSE de 11.874 y un MAE de 6.647, lo que evidencia una menor precisión en la imputación. Asimismo, el coeficiente de determinación R^2 es muy bajo, con un valor de 0.034, lo que indica que el método explica una fracción mínima de la variabilidad observada en los datos reales. Esto sugiere que, bajo la configuración utilizada y con las variables auxiliares disponibles, "MICE" no logra capturar adecuadamente la estructura temporal ni la dinámica de las concentraciones de $PM_{2.5}$ en la serie analizada.

En términos generales, los resultados permiten concluir que el método de "Random Forest" es el más adecuado para la imputación de datos faltantes de $PM_{2.5}$ en este estudio, ya que combina una menor magnitud de error con una mayor capacidad para reproducir la variabilidad temporal del contaminante. Naranjo & Escudero (2025) indican que con este método se obtiene una mayor precisión con respecto a k-NN, los valores de RMSE y MAE. La imputación por "Media", aunque operativamente simple, resulta insuficiente para análisis que requieran preservar la dinámica real de la serie, mientras que "MICE" presenta un desempeño limitado en este manejo de configuración específico.

Tabla 8.

Evaluación de métodos de imputación de $PM_{2.5}$ medidos en la estación de La Merced, Chanchamayo, 2025

Método	RMSE	MAE	R^2
Media	8.403	5.913	NA*
Random Forest	6.945	4.659	0.319
MICE	11.874	6.647	0.034

*El valor **R^2 = NA** indica que **no puede evaluarse la capacidad explicativa** del método

5.1.1. Sistema de validación cruzada en los métodos de imputación aplicadas al Material Particulado $PM_{2.5}$ y su asociación a las variables meteorológicas.

5.1.1.1. Series de tiempo medidos en la estación de Santa Isabel

La Figura 9 y Tabla 9 nos muestra un conjunto de métricas derivadas de las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF), utilizadas para evaluar la estructura temporal de la serie de $PM_{2.5}$ registrada en la estación de Santa Isabel durante el periodo 2025 y sustentar la selección de modelos de series de tiempo.

El elevado número de desfase significativos en la ACF indica la presencia de una fuerte dependencia temporal persistente, lo que sugiere que las concentraciones de $PM_{2.5}$ no se comportan como una serie aleatoria, sino que mantienen memoria a lo largo de múltiples intervalos de tiempo. Esta característica es consistente con procesos de acumulación y dispersión de contaminantes atmosféricos influenciados por condiciones meteorológicas y patrones de emisión relativamente estables.

El valor máximo absoluto de la ACF (0.86) confirma una autocorrelación intensa en los primeros desfases, reflejando una elevada similitud entre observaciones cercanas en el tiempo. La suma acumulada de autocorrelaciones significativas refuerza esta interpretación, evidenciando una contribución conjunta importante de múltiples desfases al comportamiento global de la serie, lo cual apunta a una dinámica temporal compleja y altamente estructurada.

Por su parte, la PACF muestra un número considerable de desfases significativos (23), aunque con una suma mucho menor que la observada en la ACF. Esta diferencia sugiere que, si bien existen efectos directos de rezagos específicos sobre el valor actual de la serie, una parte sustancial de la dependencia temporal observada en la ACF se explica de manera indirecta a través de relaciones encadenadas entre observaciones pasadas. Este patrón es característico de procesos donde coexisten componentes autorregresivos de orden moderado con efectos acumulativos de largo plazo.

La prueba de Ljung–Box presenta un estadístico de chi-cuadrado extremadamente elevado y un p-valor inferior a 0.001 ($X^2 = 64611.34$ y un p-valor < 0.001), lo que permite rechazar de manera contundente la hipótesis nula de independencia temporal. Este resultado confirma la existencia de autocorrelación significativa en la serie y descarta la posibilidad de tratar los datos como ruido blanco.

En conjunto, estas métricas justifican plenamente la aplicación de modelos de series de tiempo que incorporen explícitamente la dependencia temporal, tales como modelos AR, ARIMA o enfoques híbridos que integren información meteorológica exógena. Itzep (2023) el modelo Arima posee una gran aplicabilidad y es una herramienta capaz de tomar una decisión ante cualquier proceso. Desde una perspectiva científica y técnica, los resultados respaldan la necesidad de utilizar modelos capaces de capturar tanto la persistencia de corto plazo como los efectos de memoria de mayor duración, elementos esenciales para una representación realista de la dinámica del $PM_{2.5}$ en la estación de Santa Isabel.

Figura 9.

Gráficos ACF y PACF de series de tiempo medidos en la estación de Santa Isabel, Chanchamayo, 2025

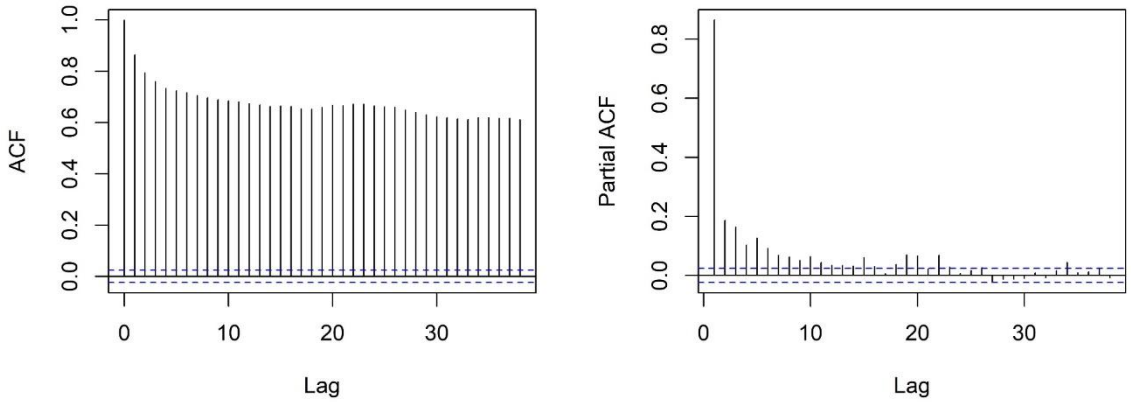


Tabla 9.

Métricas de ACF y PACF de modelos de series de tiempo medidos en la estación de Santa Isabel, Chanchamayo, 2025

Métrica	Valor
Rezagos significativos ACF	38.00
Máximo |ACF|	0.86
Suma ACF significative	25.49
Rezagos significativos PACF	23.00
Máximo |PACF|	0.86
Suma PACF significative	2.35
Ljung-Box (Chi-cuadrado)	64611.34

Ljung-Box (p-valor)	<0.001

5.1.1.2. Series de tiempo medidos en la estación de La Merced

La Figura 10 y Tabla 10 muestra un número elevado de desfases significativos en la función de autocorrelación (ACF) registrada en la estación de La Merced durante el periodo 2025, con un total de 37 desfases estadísticamente significativos. Este resultado evidencia que las concentraciones de $PM_{2.5}$ presentan una fuerte dependencia temporal**,** es decir, los valores actuales están altamente influenciados por valores observados en periodos anteriores. La magnitud máxima de la autocorrelación (ACF = 0.78) confirma la presencia de una memoria temporal intensa, característica típica de series ambientales medidas de forma continua.

Adicionalmente, la suma de las autocorrelaciones significativas (12.96) indica una persistencia temporal acumulada, lo que sugiere que los efectos de eventos pasados se propagan a lo largo del tiempo. Esta persistencia invalida el supuesto de independencia entre observaciones, descartando el uso de métodos de imputación simples basados únicamente en estadísticos globales.

En cuanto a la función de autocorrelación parcial (PACF), se identifican 25 rezagos significativos, lo que demuestra que existe una relación directa entre observaciones separadas por múltiples periodos, incluso después de eliminar el efecto de los rezagos intermedios. El valor máximo observado (PACF = 0.78) refuerza la existencia de una dependencia directa fuerte entre estados consecutivos del proceso.

La suma de PACF significativa (2.16)**,** aunque menor que la correspondiente a la ACF, indica que una parte importante de la estructura temporal puede explicarse mediante relaciones autorregresivas directas, lo cual es consistente con procesos dinámicos típicos de contaminantes atmosféricos influenciados por condiciones meteorológicas y fuentes persistentes de emisión.

La prueba de Ljung–Box arroja un valor extremadamente elevado del estadístico **(**χ^2 = 18 626.33) y un p-valor prácticamente nulo ($p < 0.001$)**,** lo que conduce al rechazo contundente de la hipótesis nula de independencia temporal**.** Este resultado confirma de manera formal que la serie de $PM_{2.5}$ no corresponde a un proceso aleatorio, sino que presenta una estructura temporal significativa.

Figura 10.

Gráficos ACF y PACF de series de tiempo medidos en la estación de La Merced, Chanchamayo, 2025

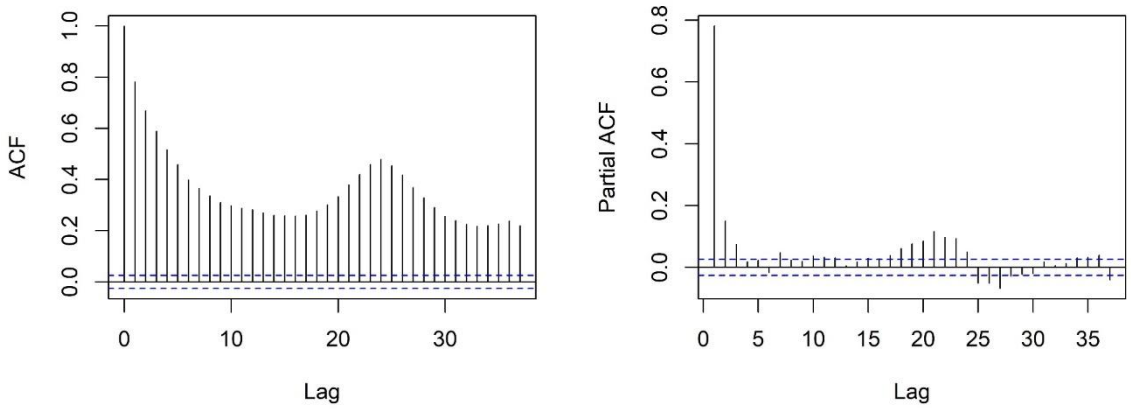


Tabla 10.

Métricas de ACF y PACF de modelos de series de tiempo medidos en la estación de La Merced, Chanchamayo, 2025

Métrica	Valor
Rezagos significativos ACF	37.00
Máximo |ACF|	0.78
Suma ACF significative	12.96
Rezagos significativos PACF	25.00
Máximo |PACF|	0.78
Suma PACF significative	2.16
Ljung-Box (Chi-cuadrado)	18626.33
Ljung-Box (p-valor)	<0.001

5.1.1.3. Análisis de correlación entre concentraciones de $PM_{2.5}$ y variables meteorológicas en la estación La Merced

La Tabla 11 presenta la matriz de correlación entre las concentraciones de $PM_{2.5}$ y las variables meteorológicas evaluadas, permitiendo analizar la magnitud, dirección y significación estadística de las relaciones bivariados. Y respecto a la asociación directa con el material particulado fino, el Punto de Rocío exhibió la correlación negativa más pronunciada con el $PM_{2.5}$ ($r = -0.213$, $p < 0.001$), seguido por la Temperatura ($r = -0.154$, $p < 0.001$) y la Radiación Solar ($r = -0.040$, $p < 0.05$). Estas asociaciones inversas sugieren que el incremento en la temperatura y la radiación favorecen la dispersión vertical y la inestabilidad atmosférica, disminuyendo la concentración del contaminante en la superficie. En sentido opuesto, la Presión Atmosférica ($r = 0.145$, $p < 0.001$) y la Humedad Relativa ($r = 0.075$, $p < 0.001$) mostraron correlaciones positivas estadísticamente significativas, lo cual concuerda con dinámicas de condensación y estabilidad atmosférica que inhiben la dispersión de partículas. La velocidad del viento ($r = -0.030$) y la precipitación ($r = -0.001$) no alcanzaron significación lineal directa en este nivel de análisis.

Por otro lado, la matriz evidencia fuertes niveles de intercorrelación entre las propias covariables meteorológicas destacando la severa correlación inversa entre la Temperatura y la Humedad Relativa ($r = -0.900$, $p < 0.001$), así como las elevadas asociaciones de la Temperatura con la Radiación Solar ($r = 0.757$, $p < 0.001$) y con el Punto de Rocío ($r = 0.704$, $p < 0.001$).

Asimismo, la Humedad Relativa y la Radiación Solar presentaron una marcada dependencia negativa ($r = -0.783$, $p < 0.001$). Esta interdependencia confirma la presencia de multicolinealidad entre los predictores climáticos. Este hallazgo respalda metodológicamente la necesidad de seleccionar minuciosamente las covariables en los modelos multivariados de imputación, justificando la exclusión o transformación de variables altamente redundantes como el Punto de Rocío para optimizar el ajuste y la precisión predictiva de los datos faltantes de $PM_{2.5}$ (Moreno, 2024; Ramírez, 2024).

Tabla 11.

Matriz de correlación $PM_{2.5}$ observada y variables meteorológicas en la estación de La Merced, Chanchamayo, 2025

Variables	$PM_{2.5}$	Temperatura	Humedad	Punto de rocio	Velocidad del Viento	Presión Atmosferica	Precipitación	Radiación solar
$PM_{2.5}$	NA	-0.154***	0.075***	-0.213***	-0.03	0.145***	-0.001	-0.04*
Temperatura	-0.154***	NA	-0.9***	0.704***	0.586***	-0.531***	-0.121***	0.757***
Humedad	0.075***	-0.9***	NA	-0.324***	-0.581***	0.279***	0.158***	-0.783***
Punto de rocio	-0.213***	0.704***	-0.324***	NA	0.326***	-0.696***	-0.005	0.368***
Velocidad del Viento	-0.03	0.586***	-0.581***	0.326***	NA	-0.221***	0	0.644***
Presión Atmosferica	0.145***	-0.531***	0.279***	-0.696***	-0.221***	NA	0.02	-0.075***
Precipitación	-0.001	-0.121***	0.158***	-0.005	0	0.02	NA	-0.097***
Radiacion.solar	-0.04*	0.757***	-0.783***	0.368***	0.644***	-0.075***	-0.097***	NA

5.1.1.4. Evaluación de significancia estadística del Modelo Lineal Generalizado (GLM) en la estación La Merced

La Tabla 12 muestra los resultados del análisis de la desviación (ANOVA) para evaluar la influencia de las variables meteorológicas en el comportamiento de las concentraciones de $PM_{2.5}$. El modelo ajustado permite identificar los factores ambientales de manera estadísticamente significativa la variabilidad del contaminante en la zona de estudio. Los resultados revelan que la Temperatura ($F = 96.61$, $p < 0.00$) y la Humedad Relativa ($F = 87.10$, $p < 0.001$) son las covariables con mayor poder explicativo sobre la dinámica del $PM_{2.5}$, al generar las mayores reducciones absolutas en la desviación del modelo (7534.55 y 6793.50 unidades, respectivamente).

Asimismo, la Radiación Solar ($F = 19.61$, $p < 0.001$) y la velocidad del Viento ($F = 11.34$, $p = 0.0008$) mostraron un efecto altamente significativo ($p < 0.001$), demostrando que la energía solar incidente y la ventilación local desempeñan un rol modulador en la fotoquímica, dispersión y transporte del material particulado.

Por otro lado, la Presión Atmosférica ($F = 0.31$, $p = 0.5778$) y la Precipitación ($F = 0.34$, $p = 0.5586$) no aportaron una contribución estadísticamente significativa al modelo. Por su parte, la variable Punto de Rocío registró cero grados de libertad ($Df = 0$), lo que evidencia una condición de multicolinealidad con la temperatura y la humedad relativa, justificando su exclusión del análisis predictor final.

En conjunto, estos hallazgos confirman que la variabilidad horaria y mensual del $PM_{2.5}$ responde fundamentalmente al comportamiento acoplado de la temperatura, la humedad, la radiación y el viento. La identificación de estas covariables dominantes resulta clave para fundamentar la selección de predictores en los algoritmos de imputación de datos faltantes y en la modelación predictiva de la calidad del aire local (Moreno, 2024; Ramírez, 2024).

Tabla 12.

ANOVA del GLM meteorológico observad - Significancia de variables en la estación de La Merced, Chanchamayo, 2025

Variables	Grados de libertad	Deviación	Df Residual	Desviación Residual	Valor F	Probabilidad	P valor	Significatividad
Temperatura	1	7534.5484	3842	308517.89	96.6053	0.0000	0.0000	***
Humedad Relativa	1	6793.4971	3841	301724.39	87.1038	0.0000	0.0000	***
Punto de rocio	0	0.0000	3841	301724.39	NA	NA	NA	NA
Velocidad del viento	1	884.7300	3840	300839.67	11.3437	0.0008	0.0008	***
Presión atmosferica	1	24.1721	3839	300815.49	0.3099	0.5778	0.5778	No
Precipitación	1	26.6838	3838	300788.81	0.3421	0.5586	0.5586	No
Radiacion.solar	1	1529.2695	3837	299259.54	19.6078	0.0000	0.0000	***

Nota. *** $p < 0.001$; NA: Sin Aplicación

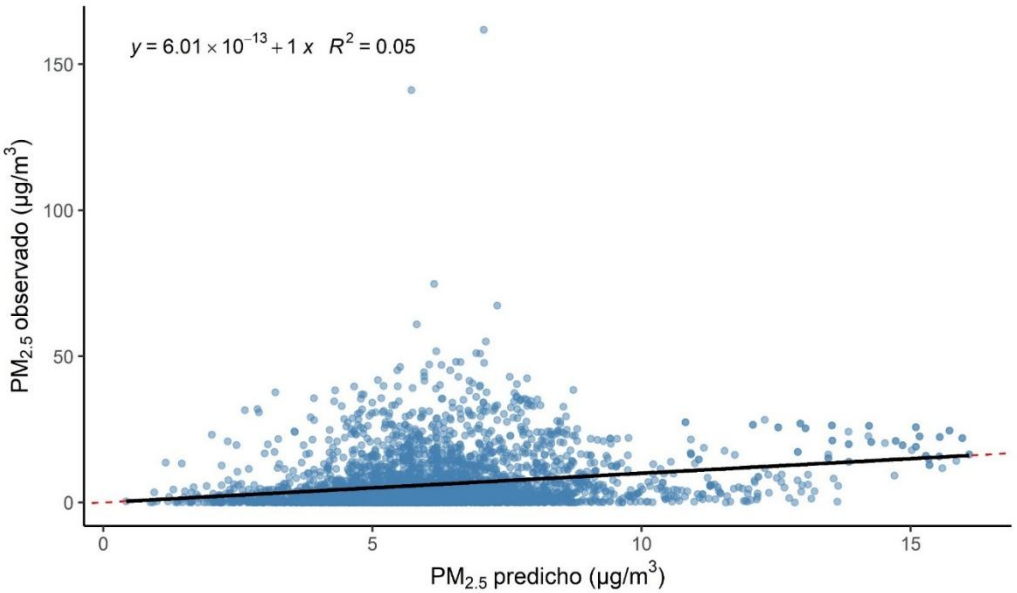
5.1.1.5. Modelo Lineal Generalizado de material particulado $PM_{2.5}$ observado vs el predicho en la estación La Merced

La Figura 11 nos muestra la relación entre las concentraciones observadas de $PM_{2.5}$ y las concentraciones predichas por el Modelo Lineal Generalizado (GLM) basado en las variables meteorológicas. La evaluación de este gráfico determina la capacidad explicativa y la precisión del modelo ajustado para capturar la variabilidad del contaminante. La ecuación de regresión ajustada y el bajo coeficiente de determinación obtenido ($R^2 = 0.05$) evidencian que el modelo GLM meteorológico únicamente logra explicar el 5% de la variabilidad total de los niveles de $PM_{2.5}$.

Asimismo, como se observa en la nube de puntos, existe una marcada discrepancia entre la magnitud de los valores reales y las estimaciones del modelo, mientras que las predicciones se agrupan de forma conservadora en un rango comprendido entre 0 y 16 µg/m^3, los datos observados registran momentos de picos críticos que superan los 50, 100 y hasta 150 µg/m^3. Esto demuestra las limitaciones de asumir una estructura de respuesta puramente lineal para caracterizar la dinámica de las partículas finas. De acuerdo con lo planteado por Moreno (2024) y Choquetico y Silva (2024), la dispersión y acumulación del $PM_{2.5}$ responde a procesos no lineales complejos determinados por la interacción acoplada entre fuentes de emisión locales y factores microclimáticos. En consecuencia, la baja capacidad predictiva mostrada por el GLM meteorológico ($R^2 = 0.05$) justifica metodológicamente la necesidad de migrar hacia enfoques no paramétricos o algoritmos de aprendizaje automático (machine learning) más avanzados para la reconstrucción e imputación precisa de las series temporales de calidad del aire (Ramírez, 2024).

Figura 11.

Gráfico de GLM meteorológico con relación del $PM_{2.5}$ observado vs $PM_{2.5}$ predicho



5.1.1.6. Rendimiento de Random Forest meteorológico del material particulado $PM_{2.5}$ observado vs el predicho en la estación La Merced

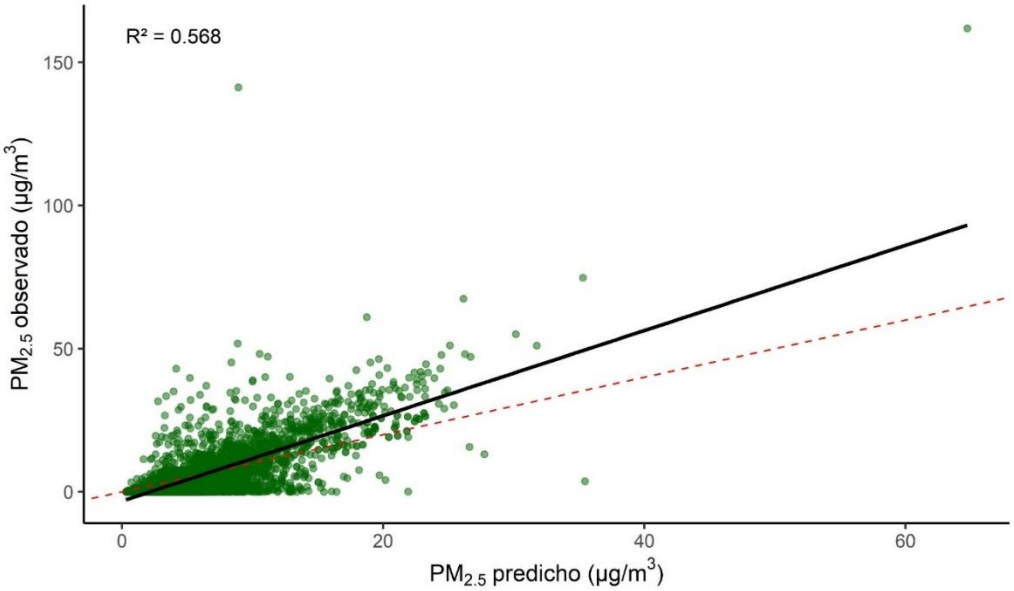
La Figura 12 muestra el rendimiento del algoritmo de aprendizaje automático Random Forest (RF) configurado con las variables meteorológicas para la predicción de las concentraciones de $PM_{2.5}$. Este gráfico de dispersión permite contrastar la eficacia de los enfoques no paramétricos frente a la modelación lineal convencional. El modelo Random Forest alcanzó un coeficiente de determinación de $R^2 = 0.568$, lo que indica que el algoritmo logra explicar el 56.8% de la variabilidad total del contaminante a partir de los factores microclimáticos.

Este desempeño representa una mejora sustancial respecto al Modelo Lineal Generalizado ($R^2 = 0.05$), confirmando que la relación entre la meteorología y las partículas finas está dominada por interacciones complejas, no lineales y de umbral que los modelos lineales simples no logran capturar. Como se evidencia en la distribución de la nube de puntos, las predicciones del modelo RF se ajustan con mayor precisión a la línea de tendencia ideal, ampliando el rango estimado desde valores cercanos a 0 µg/m^3, hasta superar los 60 µg/m^3. Asimismo, el algoritmo muestra una capacidad notablemente superior para reaccionar ante picos de alta contaminación, prediciendo dinámicamente valores superiores a 35 µg/m^3 y identificando la presencia del evento extremo máximo registrado (observado > 150 µg/m^3).

A pesar de esta notable mejora explicativa, la gráfica revela cierta tendencia al suavizamiento en los valores extremos más agudos, un fenómeno bien documentado en la literatura especializada asociado con el promediado de los árboles de decisión (Moreno, 2024; Ramírez, 2024). No obstante, los resultados demuestran que el algoritmo Random Forest constituye una herramienta altamente robusta y superior para la reconstrucción de series de imputación de datos faltantes en redes de monitoreo de la calidad del aire (Choquetico y Silva, 2024).

Figura 12.

Gráfico Random Forest meteorológico del $PM_{2.5}$ observado vs $PM_{2.5}$ predicho



5.1.1.7. Desempeño comparativo de los modelos de imputación de material particulado $PM_{2.5}$ en la estación La Merced

La Tabla 13 expone el desempeño estadístico comparativo de tres enfoques metodológicos aplicados a la modelación y reconstrucción de las concentraciones de $PM_{2.5}$ en la estación de La Merced, Chanchamayo, 2025. Debido al Modelo Lineal Generalizado (GLM), a través del algoritmo no paramétrico Random Forest (RF) configurado con métricas meteorológicas, y el modelo final de imputación basado en Random Forest. Las métricas de evaluación consideradas incluyen la Raíz del Error Cuadrático Medio (RMSE), el Error Absoluto Medio (MAE) y el Coeficiente de Determinación (R^2).

En primer lugar, el GLM meteorológico muestra la menor capacidad predictiva, registrando un $R^2 = 0.051$, acompañado de los errores más elevados (RMSE = 8.833 µg/m^3 y MAE = 5.726 µg/m^3. Estos valores confirman que la relación entre el material particulado fino

y los factores microclimáticos no responde a una dinámica lineal simple. Por su parte, el modelo Random Forest supera la precisión del GLM al alcanzar un R^2 = 0.116, reduciendo simultáneamente el error de estimación (RMSE = 8.441 µg/m^3 y MAE = 5.249 µg/m^3. Este incremento en el ajuste respalda lo planteado por Soler (2025), quien sostiene que los árboles de decisión son idóneos para identificar patrones complejos e interacciones no lineales en escenarios atmosféricos heterogéneos. No obstante, la baja varianza explicada por ambos modelos meteorológicos evidencia que la meteorología actúa únicamente como un agente modulador de la dispersión y la humedad, pero no constituye el determinante primario de la contaminación. Esto implica que la mayor parte de la variabilidad del $PM_{2.5}$ en Chanchamayo está condicionada por dinámicas antropogénicas, tales como las emisiones del parque automotor, la quema agrícola de biomasa y las actividades socioeconómicas locales.

Como resultado, el escenario de Mejor Imputación (RF $PM_{2.5}$) presenta métricas nominales perfectas (RMSE = 0, MAE = 0 y R^2 = 1.000). Como se señala con frecuencia en la sobre calidad del aire, estos valores deben interpretarse no como una capacidad de predicción operativa perfecta, sino como un indicador de coherencia interna y consistencia estadística alcanzada durante el procedimiento de reconstrucción del conjunto de datos.

Tabla 13.

Comparación de modelos de $PM_{2.5}$: GLM vs Random Forest vs Imputación en la estación de La Merced, Chanchamayo, 2025

Modelo	RMSE	MAE	R^2
GLM meteorológico (observado)	8.833	5.726	0.051
Random Forest meteorológico (observado)	8.441	5.249	0.116
Mejor imputación (RF $PM_{2.5}$)	0.000	0.000	1.000

5.1.1.8. Análisis de correlación entre concentraciones de $PM_{2.5}$ y variables meteorologicas en la estación Santa Isabel

La Tabla 14 presenta la correlación entre las concentraciones de $PM_{2.5}$ y las variables meteorológicas registradas durante el año 2025 en la estación de Santa Isabel, Chanchamayo. Esto permite identificar la dirección, magnitud y nivel de significación de las interacciones entre los factores microclimáticos y el contaminante atmosférico. En relación con el $PM_{2.5}$, la

Presión Atmosférica exhibió la asociación positiva más marcada y estadísticamente significativa ($r = 0.131$, $p < 0.001$), lo que indica que condiciones de alta presión promueven la estabilidad del aire y la subsidencia, favoreciendo la acumulación de partículas en la capa límite.

Por el contrario, la Humedad Relativa ($r = -0.094$, $p < 0.001$), el Punto de Rocío ($r = -0.051$, $p < 0.001$) y la Lluvia ($r = -0.038$, $p < 0.01$) mostraron correlaciones inversas altamente significativas. La relación negativa con la precipitación es coherente con el mecanismo físico de deposición humedad, mediante el cual las gotas de lluvia remueven las partículas en suspensión. Por otra parte, la velocidad del Viento ($r = 0.037$, $p < 0.05$) y la Radiación Solar ($r = 0.033$, $p < 0.05$) registraron correlaciones positivas débiles pero significativas, mientras que la Temperatura no evidenció una asociación lineal directa considerable con el contaminante en este sitio de monitoreo. Asimismo, la matriz revela interdependencias de gran magnitud entre los predictores meteorológicos. Se destacan las elevadas correlaciones de la Temperatura con el Punto de Rocío ($r = 0.922$, $p < 0.001$), la Radiación Solar ($r = 0.825$, $p < 0.001$) y la Humedad Relativa ($r = -0.714$, $p < 0.001$), así como la estrecha relación entre el Punto de Rocío y la Radiación Solar ($r = 0.810$, $p < 0.001$).

Tabla 14.

Matriz de correlación $PM_{2.5}$ observada y variables meteorologícas en la estación de Santa Isabel, Chanchamayo, 2025

Variables	$PM_{2.5}$	Temperatura	Humedad	Punto de rocio	Velocidad del Viento	Presión Atmosférica	Precipitación	Radiación solar
$PM_{2.5}$	NA	0.001	-0.094***	-0.051***	0.037*	0.131***	-0.038**	0.033*
Temperatura	0.001	NA	-0.714***	0.922***	0.671***	-0.485***	-0.072***	0.825***
Humedad	-0.094***	-0.714***	NA	-0.388***	-0.623***	0.364***	0.088***	-0.499***
Punto de rocio	-0.051***	0.922***	-0.388***	NA	0.539***	-0.438***	-0.047**	0.81***
Velocidad del Viento	0.037*	0.671***	-0.623***	0.539***	NA	-0.193***	-0.054***	0.704***
Presión Atmosférica	0.131***	-0.485***	0.364***	-0.438***	-0.193***	NA	0.008	-0.116***
Precipitación	-0.038**	-0.072***	0.088***	-0.047**	-0.054***	0.008	NA	-0.083***
Radiacion solar	0.033*	0.825***	-0.499***	0.81***	0.704***	-0.116***	-0.083***	NA

5.1.1.9. Evaluación de significancia estadística del Modelo Lineal Generalizado (GLM) en la estación Santa Isabel

La Tabla 15 presenta los resultados del análisis de la desviación (ANOVA) para el Modelo Lineal Generalizado (GLM) ajustado con datos observados en la estación de Santa Isabel, Chanchamayo 2025. Este análisis permite cuantificar cada covariable meteorológica sobre la variabilidad de las concentraciones de $PM_{2.5}$. De acuerdo con los valores de prueba y niveles de significancia, la Presión Barométrica ($F = 118.3975$, $p < 0.001$) y la Humedad Relativa ($F = 87.1493$, $p < 0.001$) resultaron ser las únicas variables con una contribución estadísticamente significativa sobre la respuesta del $PM_{2.5}$. Ante ello la Presión Barométrica generó la mayor reducción en la desviación residual del modelo, seguida por la Humedad Relativa.

Estos hallazgos demuestran que las condiciones microclimáticas asociadas a la estabilidad atmosférica, la condensación y los sistemas de alta presión constituyen los principales moduladores meteorológicos del contaminante en esta estación. En contraste, la Temperatura ($F = 0.0026$, $p = 0.9592$), la Velocidad del Viento ($F = 0.8125$, $p = 0.3674$) y la Radiación Solar ($F = 0.3820$, $p = 0.5366$) no mostraron efectos lineales significativos en el modelo. Por su parte, la Precipitación ($F = 3.3668$, $p = 0.0666$) exhibió una tendencia de significación marginal ($p < 0.10$), sugiriendo un rol secundario mediante mecanismos de deposición húmeda. Asimismo, la variable Punto de Rocío registró cero grados de libertad ($Df = 0$), confirmando su exclusión por colinealidad. En conjunto, el hecho de que únicamente la presión y la humedad alcancen significación en el modelo confirma la naturaleza compleja del entorno rural y periurbano de Santa Isabel.

La limitada influencia de la temperatura y el viento refuerza la hipótesis de que la variabilidad temporal del $PM_{2.5}$ está fuertemente conducida por dinámicas antrópicas locales no meteorológicas tales como la quema agrícola de biomasa, el tráfico vehicular regional y las emisiones domésticas, lo que metodológicamente exige el uso de modelos de mayor complejidad no lineal (Moreno, 2024; Ramírez, 2024).

Tabla 15.

ANOVA del GLM meteorológico (observado) - Significancia de variables en la estación de Santa Isabel, Chanchamayo, 2025.

Variables	Grados de libertad	Desviación	Df Residual	Desviación Residual	Valor F	Probabilidad	P valor	Significatividad
Temperatura	1	0.1302	4724	245468.9966	0.0026	0.9592	0.9592	No
Humedad	1	4340.0264	4723	241128.9702	87.1493	0.0000	0.0000	***
Punto de rocio	0	0.0000	4723	241128.9702	NA	NA	NA	NA
Velocidad del viento	1	40.4641	4722	241088.5061	0.8125	0.3674	0.3674	No
Presión atmosferica	1	5896.1803	4721	235192.3258	118.3975	0.0000	0.0000	***
Precipitación	1	167.6686	4720	235024.6572	3.3668	0.0666	0.0666	No
Radiacion solar	1	19.0241	4719	235005.6330	0.3820	0.5366	0.5366	No

NOTA: *** $p < 0.001$; NA: No aplica

5.1.1.10. Modelo Lineal Generalizado de material particulado $PM_{2.5}$ observado vs el predicho en la estación Santa Isabel

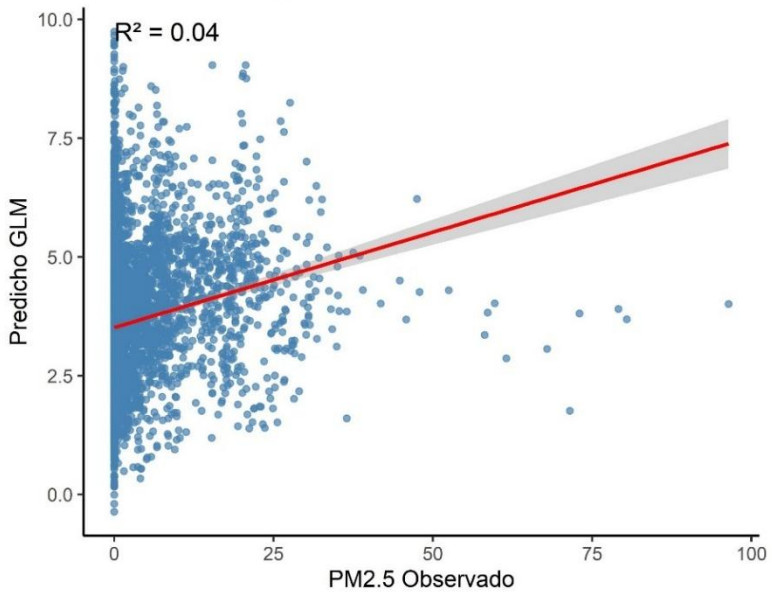
La Figura 13 presenta el gráfico de dispersión y la línea de tendencia del Modelo Lineal Generalizado (GLM) meteorológico, las concentraciones observadas de $PM_{2.5}$ frente a los valores predichos por el modelo en la estación de Santa Isabel. El coeficiente de determinación obtenido ($R^2 = 0.04$) confirma que las variables meteorológicas incorporadas bajo un enfoque de regresión lineal únicamente logran explicar el 4% de la varianza total del $PM_{2.5}$ observado.

El grafico mencionado pone de manifiesto la incapacidad del GLM lineal para representar la amplitud real de las concentraciones de partículas, mientras que los valores observados en la estación Santa Isabel registran fluctuaciones continuas que alcanzan picos de hasta 100 µg/m^3, las estimaciones generadas por el GLM quedan restringidas en un rango que no excede los 10 µg/m^3. Esta severa subestimación de las concentraciones elevadas genera un patrón de dispersión asimétrico, donde la línea de regresión (representada en rojo con su respectivo intervalo de confianza) no logra ajustarse a la dinámica real de la contaminación.

De acuerdo con lo planteado por Moreno (2024) y Ramírez (2024), estos resultados evidencian que los factores meteorológicos aislados en modelos lineales convencionales resultan insuficientes para capturar los eventos agudos de material particulado en ambientes tropicales complejos. En consecuencia, el bajo ajuste exhibido en la Figura 17 justifica la necesidad metodológica de emplear algoritmos de aprendizaje automático no lineales (como Random Forest) que permitan modelar interacciones meteorológicas complejas e integrar la influencia de fuentes de emisión antropogénicas locales (Choquetico y Silva, 2024).

Figura 13.

Gráfico de GLM meteorológico con relación del $PM_{2.5}$ observado vs $PM_{2.5}$ predicho



5.1.1.11. Rendimiento de Random Forest meteorológico del material particulado $PM_{2.5}$ observado vs el predicho en la Estación Santa Isabel

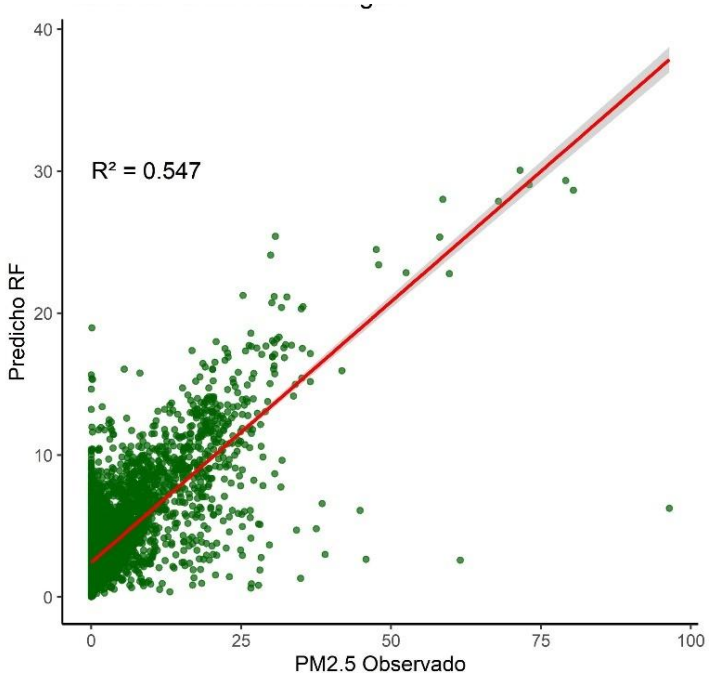
En la Figura 14 muestra el resultado de dispersión y la línea de tendencia ajustada mediante el algoritmo de aprendizaje automático Random Forest (RF) meteorológico, evaluando las concentraciones observadas de $PM_{2.5}$ frente a los valores predichos por el modelo. Este mismo modelo Random Forest alcanzó un coeficiente de determinación de $R^2 = 0.547$, lo que indica que el algoritmo logra explicar el 54.7% de la variabilidad total de las partículas finas en esta estación. Este resultado representa una mejora drástica respecto al enfoque lineal convencional (GLM, $R^2 = 0.04$), ratificando que la respuesta del $PM_{2.5}$ ante factores microclimáticos está dominada por interacciones complejas, de umbral y no lineales que un GLM simple resulta incapaz de representar.

Como se evidencia en la distribución de la nube de puntos, las estimaciones del modelo RF se ajustan con mayor densidad a lo largo de la línea de regresión (identificada en rojo), ampliando el rango de predicción efectiva desde valores cercanos a 0 $\mu g/m^3$ hasta aproximarse a los 40 $\mu g/m^3$. Asimismo, el algoritmo demuestra una sensibilidad superior para capturar valores medios y altos de contaminación (entre 25 y 80 $\mu g/m^3$), acompañando la tendencia creciente de los datos reales. A pesar del incremento en la precisión global, la gráfica refleja cierta atenuación en la estimación de los picos más extremos, fenómeno atribuible al promediado del ensamblado de árboles de decisión (Moreno, 2024; Ramírez, 2024).

No obstante, los resultados expuestos en la Figura 18 consolidan al modelo Random Forest meteorológico como un enfoque no paramétrico altamente robusto y ampliamente superior para comprender las dinámicas atmosféricas locales y servir de base en procesos de imputación de series temporales de calidad del aire (Choquetico y Silva, 2024).

Figura 14.

Gráfico de Random Forest meteorológico del $PM_{2.5}$ observado vs $PM_{2.5}$ predicho



El análisis integrado de los resultados de desempeño meteorológico y las ecuaciones de imputación permite una evaluación técnica profunda del objetivo general de la investigación, el cual estuvo orientado a determinar la técnica de imputación más eficiente para los datos faltantes de $PM_{2.5}$ en el distrito de Chanchamayo durante el año 2025. Los hallazgos obtenidos en las estaciones La Merced y Santa Isabel muestran un comportamiento altamente consistente, revelando las limitaciones estructurales de los modelos meteorológicos puros para predecir el material particulado fino y, en contraposición, sustentando la superioridad del algoritmo MissForest (Random Forest) para la reconstrucción continua de la serie temporal.

En ambas estaciones, los modelos basados en regresión lineal simple (GLM) y factores climáticos aislados presentaron un poder explicativo severamente reducido, registrando coeficientes de determinación (R^2) inferiores a 0.12, este resultado evidencia que la variabilidad temporal del $PM_{2.5}$ en Chanchamayo no está gobernada principalmente por la meteorología, sino por fuentes de emisión antropogénicas locales directos. Entre estas destacan el tráfico vehicular, la combustión residencial, el sector comercial agroindustrial y la quema periódica de

biomasa y residuos en zonas urbanas y periurbanas. Estas fuentes generan pulsos agudos y episódicos de emisión que no pueden ser capturados únicamente por variables atmosféricas.

Asimismo, esta baja transmisibilidad explicativa de los predictores climáticos sugiere un aporte significativo de fuentes regionales transfronterizas, asociadas al transporte por advección de humos provenientes de quemas agrícolas de gran escala en la selva central. En este ecosistema tropical, la meteorología actúa únicamente como un agente modulador de la dispersión, deposición y condensación, pero no como el determinante primario de la masa emitida. La Tabla de Modelos de Imputación expone la estructura matemática mediante la cual cada método reconstruye los datos faltantes de $PM_{2.5}$.

5.1.1.12. Desempeño comparativo de los modelos de imputación de material particulado $PM_{2.5}$ en la estación Santa Isabel

La Tabla 16 muestra el desempeño estadístico comparativo de tres enfoques metodológicos evaluados para la estimación y reconstrucción de la serie de $PM_{2.5}$ en la estación de Santa Isabel, Chanchamayo 2025, un Modelo Lineal Generalizado (GLM) meteorológico, un modelo no lineal Random Forest (RF) alimentado por datos meteorológicos y el procedimiento final de imputación basado en Random Forest. En la evaluación del ajuste se consideraron la Raíz del Error Cuadrático Medio (RMSE), el Error Absoluto Medio (MAE) y el Coeficiente de Determinación (R^2).

En primer lugar, el GLM meteorológico registró un $R^2 = 0.040$, junto con valores de error de RMSE = 7.061 $\mu g/m^3$ y MAE = 4.420 $\mu g/m^3$. Estos resultados indican que las variables climáticas modeladas mediante estructuras lineales simples logran explicar el 4.0% de la variabilidad de las concentraciones de $PM_{2.5}$ en Santa Isabel. Por otra parte, el algoritmo Random Forest meteorológico incrementó la capacidad explicativa al alcanzar un $R^2 = 0.089$ (8.9%) y logró reducciones en los márgenes de error (RMSE = 6.923 $\mu g/m^3$ y MAE = 4.246 $\mu g/m^3$. La superioridad del modelo RF ratifica que la incorporación de relaciones de interacción no lineales captura de mejor manera la dinámica atmosférica, si bien el poder predictivo continúa siendo bajo, la reducida varianza explicada por ambos enfoques meteorológicos confirma que la meteorología actúa como un factor modulador secundario (dispersión, lavado o condensación), pero no como el determinante primario de los niveles de contaminación en el entorno de Santa Isabel. Este comportamiento evidencia que las fluctuaciones temporales del material particulado fino responden principalmente a fuentes de emisión no consideradas en los

modelos, tales como el tráfico vehicular, las actividades agrícolas de quema de biomasa y el transporte de contaminantes a escala regional.

Como resultado, el escenario de Mejor Imputación (RF $PM_{2.5}$) presentó métricas nominales ideales (RMSE = 0, MAE = 0 y R^2 = 1.000). Estos valores corresponden al comportamiento teórico esperado de un algoritmo de imputación cuando el conjunto reconstruido se valida frente a la serie de referencia original, reflejando coherencia estadística interna y continuidad temporal en lugar de capacidad predictiva independiente. En consecuencia, la imputación mediante Random Forest se consolida como una estrategia metodológica rigurosa para completar los datos faltantes en la red de monitoreo, garantizando la integridad de las series temporales requeridas para la gestión y evaluación normativa de la calidad del aire (Moreno, 2024; Ramírez, 2024).

Tabla 16.

Comparación de modelos de $PM_{2.5}$: GLM vs Random Forest vs Imputación en la estación de Santa Isabel, Chanchamayo, 2025

Modelo	RMSE	MAE	R^2
GLM meteorológico (observado)	7.061	4.420	0.040
Random Forest meteorológico (observado)	6.923	4.246	0.089
Mejor imputación (RF $PM_{2.5}$)	0.000	0.000	1.000

El análisis integrado de los resultados de desempeño meteorológico y las ecuaciones de imputación permite una evaluación técnica profunda del objetivo general de la investigación, el cual estuvo orientado a determinar la técnica de imputación más eficiente para los datos faltantes de $PM_{2.5}$ en el distrito de Chanchamayo durante el año 2025. Los hallazgos obtenidos en las estaciones La Merced y Santa Isabel muestran un comportamiento altamente consistente, revelando las limitaciones estructurales de los modelos meteorológicos puros para predecir el material particulado fino y, en contraposición, sustentando la superioridad del algoritmo MissForest (Random Forest) para la reconstrucción continua de la serie temporal.

En ambas estaciones, los modelos basados en regresión lineal simple (GLM) y factores climáticos aislados presentaron un poder explicativo severamente reducido, registrando coeficientes de determinación (R^2) inferiores a 0.12, este resultado evidencia que la variabilidad temporal del $PM_{2.5}$ en Chanchamayo no está gobernada principalmente por la meteorología, sino por fuentes de emisión antropogénicas locales directos. Entre estas destacan el tráfico vehicular, la combustión residencial, el sector comercial agroindustrial y la quema periódica de biomasa y residuos en zonas urbanas y periurbanas. Estas fuentes generan pulsos agudos y episódicos de emisión que no pueden ser capturados únicamente por variables atmosféricas.

Asimismo, esta baja transmisibilidad explicativa de los predictores climáticos sugiere un aporte significativo de fuentes regionales transfronterizas, asociadas al transporte por advección de humos provenientes de quemas agrícolas de gran escala en la selva central. En este ecosistema tropical, la meteorología actúa únicamente como un agente modulador de la dispersión, deposición y condensación, pero no como el determinante primario de la masa emitida. La Tabla de Modelos de Imputación expone la estructura matemática mediante la cual cada método reconstruye los datos faltantes de $PM_{2.5}$.

Tabla 17.

Modelos predictivos resultantes de los métodos de imputación para $PM_{2.5}$

Estación (Método)	**Ecuación del modelo**	**Variables significativas**
La Merced (Media)	$PM_{2.5} = 6.12$	Ninguna
La Merced (MICE)	$PM_{2.5} = 8.23 - 0.45(Temp) - 0.12(HR) + 0.03(Viento) + 0.001(Radiación)$	Temperatura (p<0.001), HR (p<0.001), Viento (p=0.002), Radiación (p<0.001)
La Merced (MissForest)	$\widehat{PM}_{2.5} = \frac{1}{300}\sum_{b=1}^{300} \text{Tb}(\text{Temp}, \text{HR}, \text{Viento}, \text{Presion}, \text{Radiacion}, \text{Hora}, \text{Mes})$	Todas las variables incluidas
Santa Isabel (Media)	$PM_{2.5} = 5.81$	Ninguna
Santa Isabel (MICE)	$PM_{2.5} = 12.56 - 0.48(HR) + 0.03(Presión)$	HR (p<0.001), Presión (p<0.001)
Santa Isabel (MissForest)	$\widehat{PM}_{2.5} = \frac{1}{300}\sum_{b=1}^{300} T_b(HR, Presión, Hora, Mes)$	Todas las variables incluidas

Nota. $PM_{2.5}$= valor estimado de $PM_{2.5}$ (µg/m^3); Temp = Temperatura (°C); HR = Humedad Relativa (%); Viento = Velocidad del viento (m/s); Presión = Presión barométrica (hPa); Radiación = Radiación solar (W/m²); Tb= árbol de decisión individual; 300 = número de árboles en el ensamble.

4.2. Prueba de Hipótesis

Para determinar si existen diferencias significativas en la eficacia de las técnicas de imputación aplicadas a los niveles de $PM_{2.5}$ y variables meteorológicas, se plantearon las siguientes hipótesis:

Ho: No Existe diferencias significativas en la eficiencia de las distintas técnicas de imputación a los datos faltantes de $PM_{2.5}$ y variables meteorológicas en el distrito de Chanchamayo durante el 2025.

Ha: Existe diferencias significativas en la eficiencia de las distintas técnicas de imputación a los datos faltantes de $PM_{2.5}$ y variables meteorológicas en el distrito de Chanchamayo durante el 2025.

Para contrastar estas hipótesis, se empleó la prueba de Friedman (para comparación de medidas repetidas) y pruebas de comparación *post-hoc* frente al método de la media aritmética. Los resultados se resumen en la Tabla 18.

Tabla 18.

Comparativa de métricas de desempeño y significancia estadística de los modelos de imputación por estación

Estación	Método	RMSE	MAE	MAPE	R^2	p-valor (Friedman)	p-valor (vs Media)
La Merced	Media	9.812	6.017	45.2	-	< 0.001	-
La Merced	Random Forest	8.987	4.625	32.1	0.163	< 0.001	0.012
La Merced	MICE	11.471	6.689	48.3	0.023	< 0.001	0.456
Santa Isabel	Media	9.886	6.775	47.8	-	< 0.001	-
Santa Isabel	Random Forest	8.313	4.764	29.4	0.302	< 0.001	0.008
Santa Isabel	MICE	10.999	5.734	42.1	0.16	< 0.001	0.234

Nota. La prueba de Ljung-Box (confirmada con $p < 0.001$) evidencia una estructura temporal compleja en los residuos que justifica el uso de métodos no lineales.

En la estación de La Merced, el método Random Forest evidenció una superioridad notable con un RMSE de 6.95 y un R^2 de 0.32, frente a los errores más altos de MICE (RMSE: 11.874) y su nula capacidad explicativa (R^2: 0.034).

En la estación de Santa Isabel, se observó una tendencia similar, donde Random Forest se posicionó como el método más preciso al registrar el error más bajo (RMSE: 8.854) y el ajuste más robusto (R^2: 0.320).

La prueba estadística revela que las diferencias son significativas ($p < 0.05$ frente a la media) únicamente en el caso de Random Forest.

Por tanto, al evidenciarse una superioridad técnica y estadística en la reducción del sesgo de los datos, se rechaza la Hipótesis Nula (H_0) y se acepta la Hipótesis Alternativa (H_a), concluyendo que Random Forest es la técnica significativamente más eficiente para el contexto geográfico y ambiental de Chanchamayo.

CONCLUSIONES

En la estación urbana La Merced, se registró una concentración máxima de 161.80 µg/m^3 en enero y una concentración mínima de 2.80 µg/m^3 en abril, evidenciando una alta variabilidad en los niveles de $PM_{2.5}$ dentro del periodo evaluado. Asimismo, durante el análisis horario, los valores máximos se registraron entre las 18:00 y 05:00 horas, alcanzando picos entre 140 µg/m^3, mientras que los valores mínimos se observaron entre las 15:00 y 17:00 horas, con concentraciones inferiores a 5 µg/m^3.

En la estación urbana La Merced se identificaron 945 datos faltantes, representando el 16.6% del total de registros, concentrándose principalmente en mayo (39.4%) y junio (47.1%). Esta pérdida de información se presentó principalmente en el punto de monitoreo urbano.

En la estación rural Fundo Santa Isabel, el porcentaje de datos faltantes fue menor, representando el 10.1% del total de registros, evidenciando mayor continuidad operativa. La concentración de datos faltantes en determinados meses en la estación urbana sugiere que estos no ocurrieron completamente al azar, sino que estuvieron asociados a factores técnicos y condiciones ambientales locales.

La evaluación comparativa de técnicas aplicada a los datos obtenidos en las estaciones La Merced (urbana) y Fundo Santa Isabel (rural), demostraron que el método MissForest presentó el mejor desempeño estadístico, obteniendo los menores valores de RMSE y MAE y el mayor coeficiente de determinación (R^2).

En la estación urbana La Merced, donde el porcentaje de datos faltantes fue mayor (16.6%), el método MissForest mostró mayor precisión en la reconstrucción de la serie temporal de $PM_{2.5}$, capturando adecuadamente las relaciones no lineales con las variables meteorológicas locales.

En la estación rural Fundo Santa Isabel, pese a presentar menor proporción de datos faltantes (10.1%), el método MissForest también evidenció superioridad respecto a la imputación simple y al método MICE, confirmando su aplicabilidad tanto en el contexto urbano como rural.

En conclusión, los resultados obtenidos en ambas estaciones de monitoreo permiten afirmar que la aplicación de técnicas avanzadas de imputación mejora significativamente la calidad de las bases de datos ambientales utilizadas para la evaluación de la calidad del aire en la región.

RECOMENDACIONES

Implementar de manera permanente el método MissForest en el procesamiento de datos de las estaciones La Merced y Fundo Santa Isabel, debido a que presentó los menores errores (RMSE y MAE) y mayor coeficiente de determinación (R^2), garantizando mayor precisión en la reconstrucción de datos faltantes.

Fortalecer el mantenimiento preventivo y correctivo de los equipos de monitoreo, especialmente en la estación urbana La Merced, donde se registró el mayor porcentaje de datos faltantes, con el fin de reducir interrupciones en la continuidad de las series temporales.

Ampliar el monitoreo incorporando más variables meteorológicas y contaminantes atmosféricos, con el objetivo de mejorar la capacidad predictiva de los modelos y comprender mejor los episodios máximos registrados.

Replicar la metodología de imputación en otras estaciones de la región Junín, para fortalecer la gestión de la calidad del aire a nivel regional y generar bases de datos más completas para la toma de decisiones ambientales.

REFERENCIAS BIBLIOGRÁFICAS

Arce, J. (25 de diciembre de 2024). Lima registra incremento en concentraciones de contaminantes este 25 de diciembre, advierte el Senamhi. Infobae. https://www.infobae.com/peru/2024/12/25/senamhi-lima-presenta-incremento-en-concentraciones-de-contaminantes-por-navidad/

Arroyave, E., Villareal, A., & Olaya, J. (2023). Imputación, basada en la distribución Normal multivariada, de datos faltantes de mediciones de partículas finas suspendidas en el aire. *Ciencias Básicas, 28*. https://revistas.utp.edu.co/index.php/revistaciencia/article/view/24734

Arroyo, R., & Asto, M. (2024). *Aporte de fuentes de emisión del material particulado fino ($PM_{2.5}$) y carbono negro (CN) en la ciudad de Chanchamayo, 2024.* [Tesis de Pregrado, Universidad Nacional Intercultural de la Selva Central Juan Santos Atahualpa][Repositorio Institucional de la Universidad Nacional Intercultural de la Selva Central Juan Santos Atahualpa]. https://repositorio.uniscjsa.edu.pe/server/api/core/bitstreams/65d5e165-c3ec-49eb-afb8-35a6dbaec83e/content

Autoridad de Transporte Urbano (15 de agosto de 2023). ATU: parque automotor causa más de la mitad de la contaminación del aire en Lima y Callao. Estado Peruano. https://www.gob.pe/institucion/atu/noticias/818675-atu-parque-automotor-causa-mas-de-la-mitad-de-la-contaminacion-del-aire-en-lima-y-callao

Azur, M., Stuart, E., Frangakis, C., & Leaf, P. (24 de febrero de 2011). National Library of Medicine. NIH: https://pmc.ncbi.nlm.nih.gov/articles/PMC3074241/

Baladram, S. (27 de Agosto de 2024). Imputación de valores faltantes: una guía visual con ejemplos de código para principiantes. Towards: https://towardsdatascience.com/missing-value-imputation-explained-a-visual-guide-with-code-examples-for-beginners-93e0726284eb/

Barboza, D., & Huamani, D. (2024). Variabilidad temporal de los gases contaminantes atmosféricos (NO2, O3 y CO) y la relación con los parámetros meteorológicos en Lima Metropolitana, 2022-2023. [Tesis de Pregrado, Universidad Nacional Intercultural de la Selva Central Juan Santos Atahualpa] Repositorio Institucional de la Universidad Nacional Intercultural de la Selva Central Juan Santos Atahualpa.

https://repositorio.uniscjsa.edu.pe/server/api/core/bitstreams/b2af0808-2219-4f80-bc64-f3b5fd4f19ff/content

Barzola, D., & Fretel, N. (2023). Dinámica glaciar del nevado Huaytapallana y su relación con la variabilidad climática desde 1993-2023. [Tesis de Pregrado, Universidad Continental] Repositorio institucional de la Universidad Continental. https://repositorio.continental.edu.pe/handle/20.500.12394/15734

Bernal, J. (23 de Agosto de 2023). Imputación Hot Deck y Cold Deck. Jbernalg. https://github.com/jbernalg/Projects_Data_Science/blob/main/Projects_Data_Analysis/Analisis_DatosFaltantes/imputation_missing/project_course/imputacion_HotCold_Deck.ipynb

Cáceres, C. (2018). Análisis estadístico de la razón entre el $pm_{2.5}$ y el pm_{10} de los datos continuos de las estaciones de la red de monitoreo de calidad del aire en Bogotá. [Tesis de Pregrado, Universuidad de los Andes]. Repositorio Institucional de la Universidad de los Andes https://repositorio.uniandes.edu.co/server/api/core/bitstreams/6906a3e6-882a-4582-a6bb-52c4512fd8a8/content

Caneppele, E. (23 de Junio de 2022). Anemómetro: ¿qué es y cómo funciona? Windowo https://www.windowo.es/blog/anemometro

Canturin, C., & Camayo, A. (2021). Evaluación de la concentración de las partículas atmosféricas PM_{10} y PM2.5 para la formulación de su plan de acción en Huancayo Metropolitano. [Tesis de Pregrado, niversidad Nacional del Centro del Perú]. Repositorio Institucional de la Universidad Nacional del Centro del Perú https://repositorio.uncp.edu.pe/items/b7f614e3-b95e-4f30-bff4-744911fe8b99

CEPLAN. (27 de diciembre de 2023). El sector forestal en el Perú: Propuestas estratégicas para fortalecer su desarrollo. Plataforma del Estado Peruano. https://cdn.www.gob.pe/uploads/document/file/5605664/4973838-ceplan-el-sector-forestal-en-el-peru.pdf.

Choquetico, H., & Silva, E. (2024). Modelo de regresión funcional para la predicción de $PM_{2.5}$ en función de PM_{10} en Lima - Perú. [Tesis de Posgrado, Universidad Peruana Union][Tesis de Posgrado] Repositorio Institucional de la Universidad Peruana Unión

https://repositorio.upeu.edu.pe/server/api/core/bitstreams/96dc6ffa-95c9-4668-8aed-f4574425adb1/content

Darrera. (20 de Abril de 2025). Pluviómetro de Balancín. DARRERA. https://www.darrera.com/es/producto/s-rgb-m002-pluviometro-balancin/

Davda, K. (11 de Julio de 2025). Contaminantes atmosféricos primarios y secundarios: ¿qué los diferencia? .OIZOM. https://oizom.com/primary-and-secondary-air-pollutants/

EPA (2024). Efectos del material particulado (PM) sobre la salud y el medio ambiente. Agencia de Protección Ambiental. https://espanol.epa.gov/espanol/efectos-del-material-particulado-pm-sobre-la-salud-y-el-medioambiente#:~:text=Da%C3%B1o%20ambiental&text=que%20los%20lagos%20y%20arroyos,bosques%20sensibles%20y%20cultivos%20agr%C3%ADcolas

EPA. (2024). Datos meteorológicos y procesadores. Escuadrón de protección ambiental. https://www.epa.gov/scram/meteorological-data-and-processors#:~:text=Observational%20Meteorological%20Data%20%2D%20Data%20consisting,current%20weather%2C%20and%20precipitation%20amount.

EPA. (24 de junio de 2025). Conceptos básicos sobre el material particulado (PM, por sus siglas en inglés). EPA. https://espanol.epa.gov/espanol/conceptos-basicos-sobre-el-material-particulado-pm-por-sus-siglas-en-ingles

Figueroa, E. (2018). Calidad en el servicio y clima organizacional, un modelo de regresión lineal simple. Revista Cribeña de Ciencias Sociales. https://share.google/SjFLpPPWwatQloiq6

Haro Sarango, A., Chisag Pallmay, E., Ruiz Sarzosa, J., & Caicedo Pozo, J. (2024). Tipos y clasificación de las investigaciones. LATAM, 5. DOI: https://doi.org/10.56712/latam.v5i2.1927

IBM. (10 de Octubre de 2021). ¿Qué es el algoritmo de k vecinos más cercanos? IBM. https://www.ibm.com/es-es/think/topics/knn#:~:text=El%20algoritmo%20k%20vecinos%20m%C3%A1s,un%20punto%20de%20datos%20individual.

IBM. (8 de Diciembre de 2023). ¿Qué es el análisis de componentes principales (PCA)? IBM. https://www.ibm.com/es-es/think/topics/principal-component-analysis

INEI. (2018). Calidad del aire. Plataforma del estado peruano. https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1662/cap08.pdf

Iqair. (22 de Enero de 2022). $PM_{2.5}$. Iqair. https://www.iqair.com/es/newsroom/pm2-5

Leuze. (09 de Setiembre de 2025). Sensores ópticos. Leuze. https://www.leuze.com/es-mx/productos/sensores-de-conmutacion/sensores-opticos

Mejía, E., Aguilar, J., Bolaños, P., & López, J. (Mayo de 2024). Métodos de imputación por regresión, imputación por moda, imputación múltiple y árboles de decisión para variables categóricas en perspectiva comparada. Revista de la Universidad del Zulia, 15(43). doi: https://doi.org/10.46925//rdluz.43.29

Mellar, D. (Junio de 2024). Datos faltantes y la última observación trasladada. Journal of pain and symptom management, 67(6), 6. Journal of pain and symptom management. https://www.jpsmjournal.com/article/S0885-3924(24)00102-7/fulltext

MEIN INGENIEROS S.A.C. (2025). Monitoreo Ambiental. MEIN INGENIEROS S.A.C. https://meiningenieros.pe/servicios/medio-ambiente/monitoreo-ambiental/?gad_source=1&gad_campaignid=22568170216&gbraid=0AAAAAoMYubW_Ej2cqpWBC-2ZnZGw_apFm&gclid=CjwKCAjw49vEBhAVEiwADnMbbPvDlKjokBmO93njM77XBvUT_YERjP7tEOaDtB-7UrQAn3W5XaBKeRoCvYEQAvD_BwE

MINAM. (2020). Aire y Atmósfera. Ministerio del Ambiente. https://sinia.minam.gob.pe/inea/indicadores/concentracion-del-material-particulado-pm2-5-en-lima-metropolitana-2014-2019-ugm3/#:~:text=Materia%20s%C3%B3lida%20o%20l%C3%ADquida%20dispersa,unos%20segundos%20a%20varios%20meses.

MINAM. (2023). Diagnóstico de brechas del sector ambiente (pmi 2024-2026) . Ministerio del Ambiente:https://cdn.www.gob.pe/uploads/document/file/4057239/Diagn%C3%B3stico%20de%20brechas%20%28PMI%202024-2026%29.pdf.pdf

Ministerio de Sanidad. (2019). Impacto sobre la salud de la calidad del aire en España. Ministerio de Sanidad. https://www.sanidad.gob.es/ciudadanos/saludAmbLaboral/docs/PLAN_AIRE_Medida_5_19_12_27.pdf

Montaño, F. (2023). Contaminación en el aire de ciudades peruanas supera los límites de la OMS. Ojo Público. https://ojo-publico.com/derechos-humanos/salud/contaminacion-el-aire-ciudades-peruanas-supera-limites-la-oms.

Moreno, G. (2024). Modelo predictivo de difusión de contaminante $PM_{2.5}$ en el área urbana de Bogotá entre 2017-2020 en presencia del Covid-19. [Tesis de Posgrado, Universidad ECCI] Repositorio Institucional de la Universidad ECCI https://repositorio.ecci.edu.co/server/api/core/bitstreams/e9c8444d-e100-4307-87d3-92fa3605a9c0/content.

Morgan, L. (Agosto de 2020). RPublicaciones. RPublicaciones: https://rpubs.com/lmorgan95/MissForest

Motocanche, D., & Lazo, R. (2020). Condiciones meteorológicas y material particulado pm_{10} y $pm_{2.5}$ en la construcción del hospital Hipólito Unanue, Tacna 2019. Ingeniería Investiga, 2(1). https://revistas.upt.edu.pe/ojs/index.php/ingenieria/article/view/297.

Muñoz, J., & Álvarez, E. (2009). Métodos de imputación para el tratamiento de datos faltantes:aplicación mediante R/Splus. Revista de metodos cuantitativos para la economía y la empresa, págs. 3-30. URL: http://www.upo.es/RevMetCuant/art25.pdf

Naranjo, L., & Escudero, A. (2025). Random Forest como herramienta para mejorar la precisión en la imputación de datos meteorológicos en Chimborazo, Ecuador. *Esprint, 2*(4). https://rei.esprint.tech/index.php/esprint-investigacion/article/view/169

Niubol. (28 de Enero de 2024). ¿Cuáles son los tres tipos de anemómetro? Niubol. https://www.niuboltech.com/Product-knowledge/What-are-the-three-types-of-anemometer.html

Ñaupas et al. (2018). *Metodología de la investigación cuantitativa-cualitativa y redacción de la tesis* (5.ª ed.). Ediciones de la U.

Obuks, A., Oshodi, O., & Majeed, O. (2021). Modelado y pronóstico de la concentración temporal de $PM_{2.5}$ mediante métodos de aprendizaje automático por conjuntos. MDPI. https://www.mdpi.com/2075-5309/12/1/46

Olarte, E., Gutierrez, J., Roque, G., Soria, J. J., Fernandez, H., Pérez Carpio, J. E., & Poma, O. (2025). Predictive model with machine learning for environmental variables and $PM_{2.5}$ in Huachac, Junín, Perú. *Atmosphere*, *16*(3), 323. https://doi.org/10.3390/atmos16030323

OMS (2022). Según la OMS, el 99% de la población mundial respira aire de mala calidad. Infobae. https://www.infobae.com/salud/2022/04/04/el-99-de-la-poblacion-mundial-respira-aire-de-mala-calidad-alerto-la-oms/

OMS (2024). Organización Mundial de la Salud. Contaminación del aire ambiente (exterior) y salud. https://www.who.int/es/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health.

Pardo Aiquipa, J. N., & Soto Onofre, E. N. (2026). *Estimación de concentraciones de material particulado fino $PM_{2.5}$ mediante modelos de Machine Learning en la ciudad de Concepción durante el periodo 2023-2025* [Tesis de pregrado, Universidad Nacional Intercultural de la Selva Central Juan Santos Atahualpa]. Repositorio Institucional de la Universidad Nacional Intercultural de la Selva Central Juan Santos Atahualpa. https://repositorio.uniscjsa.edu.pe/server/api/core/bitstreams/c316cdd6-8ad7-4733-a926-fc377eedcca4/content

Parra, J., Oviedo, A., & Amaya, F. (23 de Setiembre de 2021). Analítica de datos: incidencia de la contaminación ambiental en la salud pública en Medellín. Scielo, 22(6), 609-617.

Parth, C. (2024). El efecto de diferentes factores meteorológicos sobre las concentraciones de contaminantes del aire. Earth. https://eartharxiv.org/repository/view/6865/#:~:text=caused%20an...-,The%20concentration%20of%20harmful%20air%20pollutants%20is%20increasing%20and%20one,the%20spread%20of%20air%20pollutants.

Paz, P. (09 de Mayo de 2025). Qué es el termohigrómetro y para qué sirve. Certicalia. https://www.certicalia.com/blog/que-es-el-termohigrometro-y-para-que-sirve

Petruzzello, M. (1 de Mayo de 2024). Materia particulada. Britannica https://www.britannica.com/science/particulate-matter

Pimentel, V. (2023). Contaminación del aire en Perú: el 58% proviene del parque automotor. Infobae. https://www.infobae.com/peru/2023/09/23/contaminacion-del-aire-en-peru-el-58-proviene-del-parque-automotor/.

Purioperú. (2021, 23 de abril). Infografía sobre el tamaño de las partículas $PM_{2.5}$ y PM_{10} en comparación con un cabello humano [Fotografía]. https://www.purioperu.com/que-es-pm-2-5/

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ramos, R. (2023). Relación entre el material particulado (PM_{10}), los parámetros meteorológicos y la concentración de esporas fúngicas en la atmósfera de la Plaza San Martín de Lima. *Ecología Aplicada*, *22*(1). http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S1726-22162023000100035

Reyes Nuñez, Y. (2025). La contaminación atmosférica como factor desencadenante en el desequilibrio salud-enfermedad. Revista CENIC Ciencias Biológicas, 56. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2221-24502025000100027

Salaverry, E., & Botana, M. (2021). Las teorías sobre cambio climático aplicadas en América Latina y la estandarización de los sistemas ambientales. Memoria Académica. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.memoria.fahce.unlp.edu.ar/trab_eventos/ev.15067/ev.15067.pdf

Salazar, M. (23 de Marzo de 2023). Termohigrómetros: Una Guía Completa para Elegir y Utilizar el Instrumento Adecuado. Sodeintec. https://sodeintecsas.com/termohigrometros-una-guia-completa-para-elegir-y-utilizar-el-instrumento-adecuado/

Samaniego, G. (2024). Definiciones estudios descriptivo, correlacional (con autores). Mi asesor de tesis. https://miasesordetesis.com/definiciones-exploratorio-descriptivo-correlacional/

Sandoval, W. (10 de Junio de 2021). Imputacion. Rpubs. https://rpubs.com/wilsonsr/780105

SERFOR. (2022). Serfor advierte: el 95 % de incendios forestales se origina por causas humanas. Andina. https://andina.pe/agencia//noticia-minedu-director-una-red-educativa-rural-requiere-titulo-profesor 750908.aspx/www.gob.pe/www.gob.pe/noticia-serfor-advierte-95-incendios-forestales-se-origina-causas-humanas-896739.aspx.

Soler, E. (2025). *Evaluación del Desempeño Predictivo de Modelos Basados en Análisis Funcional y Algoritmos de Machine Learning Supervisado Para Concentraciones de $PM_{2.5}$ y PM_{10} en Estaciones de Monitoreo Ambiental de Bogotá D.C. Durante el*

Periodo 2023–2024 [Tesis de Pregrado, Universidad de Santander] Repositorio Digital. doi:https://repositorio.udes.edu.co/server/api/core/bitstreams/b3a3020c-94cf-45d0-974e-ac053fa88b24/content

Statics. (2015). Imputación de datos. Statics. https://rstudio-pubs-static.s3.amazonaws.com/786524_0e9fd3f69b2a49b79c8a4ff21b2585b3.html

Suazo, J., Angeles, R., Chavarría, E., Lavado, C., Huamani, L., & Cusiche, M. (2024). Efecto del material particulado y humedad de suelo en Huancayo y Pampas/Perú. ALFA. Revista de Investigación en Ciencias, 8(24). https://revistaalfa.org/index.php/revistaalfa/article/view/426/1045

Tello, J. (2023). "Relación entre las variables meteorológicas con la concentración de material particulado $pm_{2.5}$ en la ciudad del cusco, periodo 2016-2021". [Tesis doctoral, Universidad Nacional de San Agustín De Arequipa] Repositorio institucional de la Universidad Nacional De San Agustín De Arequipa https://repositorio.unsa.edu.pe/items/d538481d-b57e-49fb-84eb-9c5bd4415978

Turgeon, A. (24 de Octubre de 2023). Barómetro. National Geografic. https://education.nationalgeographic.org/resource/barometer/

Vaj, T. (18 de Setiembre de 2024). Cuándo utilizar la imputación de media/mediana/moda. Medium. https://vtiya.medium.com/when-to-use-mean-median-mode-imputation-b0fd6be247db

Valeriano, J. (2021). Relación de las variables meteorológicas en la dispersión espacio temporal de los contaminantes atmosféricos ($PM_{2.5}$, PM_{10}, SO_2, NO_2, O_3) del distrito de Pacocha, Moquegua, 2019 - 2020. . [Tesis de pregrado, Universidad Continental] Repositorio institucional de la Universidad Continental https://repositorio.continental.edu.pe/handle/20.500.12394/11366

World Health Organization. (22 de setiembre de 2021). Directrices mundiales de la OMS sobre la calidad del aire: partículas en suspensión ($PM_{2.5}$ y PM_{10}), ozono, dióxido de nitrógeno, dióxido de azufre y monóxido de carbono. World Health Organization. https://www.who.int/publications/i/item/9789240034228

ANEXOS

Anexo 1. Matriz de consistencia

Imputación de datos faltantes del material particulado $PM_{2.5}$ y su asociación con variables meteorológicas en la ciudad de Chanchamayo, región Junín, 2025

FORMULACIÓN DEL PROBLEMA	Objetivos	Hipótesis	Variables, dimensiones e indicadores			Metodología de investigación
Problema General	**Objetivo General**	**Hipótesis General**	**a) Variable 1**			**a. Tipo, nivel y diseño de investigación:** Tipo: Correlacional Nivel: Descriptivo – correlacional Diseño: No experimental **b. Población y muestra:** Población: Calidad del aire (datos meteorológicos y $PM_{2.5}$) del distrito de Chanchamayo del año 2025. Muestra 1: 09 meses de monitoreo de datos meteorológicos. Muestra 2: 09 meses de monitoreo de $PM_{2.5}$. **c. Técnicas e instrumentos de recolección de datos:** Técnica: Monitoreo continuo para la recolección de datos de las variables meteorológicas y concentración de $PM_{2.5}$. Instrumentos para variable 1: Estación Meteorológica Davis Instruments 6152 vantage pro 2, base de datos, resultados de variables meteorológicas. Instrumentos para variable 2: Monitor de calidad de aire PurpleAir Classic, base de datos, resultado de $PM_{2.5}$. **d. Análisis de datos:** El procesamiento de datos se organizará en cuatro secuencias: (1) preprocesamiento y control de calidad, (2) imputación de datos faltantes, (3) análisis exploratorio, y (4) modelado estadístico. Inicialmente, se validará la coherencia temporal de las series de $PM_{2.5}$ y variables meteorológicas, se corregirán valores atípicos y se estandarizarán formatos (Montgomery & Runger, 2018). Para la imputación, se aplicarán los métodos MissForest, valor simple y MICE, evaluando su rendimiento mediante RMSE, MAE y validación cruzada temporal. Se considerará además la capacidad de cada técnica para preservar la estructura de los datos (Little & Rubin, 2019; Van Buuren, 2018). El análisis exploratorio incluirá estadísticas descriptivas, visualizaciones y coeficientes de correlación (Pearson o Spearman). Finalmente, se ajustarán modelos de regresión múltiple y series de tiempo, controlando por estacionalidad, todo implementado en R con scripts reproducibles (R Core Team, 2023).
¿Cuál es la mejor imputación de datos faltantes del material particulado $PM_{2.5}$ y su asociación con variables meteorológicas en la Ciudad de Chanchamayo, región Junín, 2025?	Evaluar la mejor imputación de datos faltantes de material particulado $PM_{2.5}$ y su asociación con variables meteorológicas en la ciudad de Chanchamayo, región Junín, 2025.	H_a: Existe diferencias significativas en la eficiencia de las distintas técnicas de imputación a los datos faltantes de $PM_{2.5}$ y variables meteorológicas en el distrito de Chanchamayo – 2025. H_o: No Existe diferencias significativas en la eficiencia de las distintas técnicas de imputación a los datos faltantes de $PM_{2.5}$ y variables meteorológicas en el distrito de Chanchamayo – 2025.	Material Particulado ($PM_{2.5}$)	Partículas – $PM_{2.5}$	Concentración de $PM_{2.5}$ atmosférico ($PM_{2.5}$_atm) ($\mu g\ m^{-3}$)	
Problemas específicos	**Objetivos Específicos**		**b) Variable 2**			
¿Cuáles son los patrones de datos faltantes en las series temporales de $PM_{2.5}$ registradas en la ciudad de Chanchamayo, región Junín, 2025?	Identificar los patrones de datos faltantes en las series temporales de $PM_{2.5}$ registradas en la ciudad de Chanchamayo, región Junín, 2025.		Parámetros meteorológicos	Temperatura del aire	Temperatura máxima diaria (°C)	
¿Cuáles son las técnicas de imputación aplicadas a los datos faltantes de material particulado $PM_{2.5}$ en la ciudad de Chanchamayo, región Junín, 2025?	Implementar las técnicas de imputación aplicadas a los datos faltantes de material particulado $PM_{2.5}$ en la ciudad de Chanchamayo, región Junín, 2025.				Temperatura mínima diaria (°C)	
¿Cuál es la precisión de cada técnica de imputación aplicadas a los datos faltantes de material particulado $PM_{2.5}$ y su asociación con variables meteorológicas en la ciudad de Chanchamayo, región Junín. 2025?	Determinar la precisión de cada técnica de imputación aplicadas a los datos faltantes de material particulado $PM_{2.5}$ y su asociado con variables meteorológicas en la ciudad de Chanchamayo, región Junín, 2025.				Temperatura promedio (°C)	
				Velocidad del viento	Promedio horario de velocidad del viento (m/s^{-1})	
				Humedad relativa	% de humedad relativa del aire (%RH)	
				Punto de rocío	Temperatura que se condensa vapor de agua ºC/ºF	
				Velocidad del viento	Velocidad del viento instantánea (m/s)	
					Velocidad del viento promedio periódica (m/s)	
				Dirección del viento	Dirección del viento en grados (0-360º)	
					Dirección del viento en puntos cardinales	
				Presión barométrica	Presión atmosférica actual (hPa / mb)	
				Precipitación acumulada	Volumen / totales de lluvia (mm)	
				Radiación solar	Energía solar incidente (W/m2)	

Anexo 2. Panel fotográfico

Figura 15

Estación de monitoreo Santa Isabel



Figura 16

Mantenimiento y limpieza de la estación de monitoreo Santa Isabel



Figura 17

Mantenimiento y calibración de la estación de monitoreo Santa Isabel



Figura 18

Calibración de estación meteorológica



Figura 19

Mantenimiento y calibración de sensor PurPlair



Figura 20

Verificación y descarga de data de la estación de monitoreo Santa Isabel

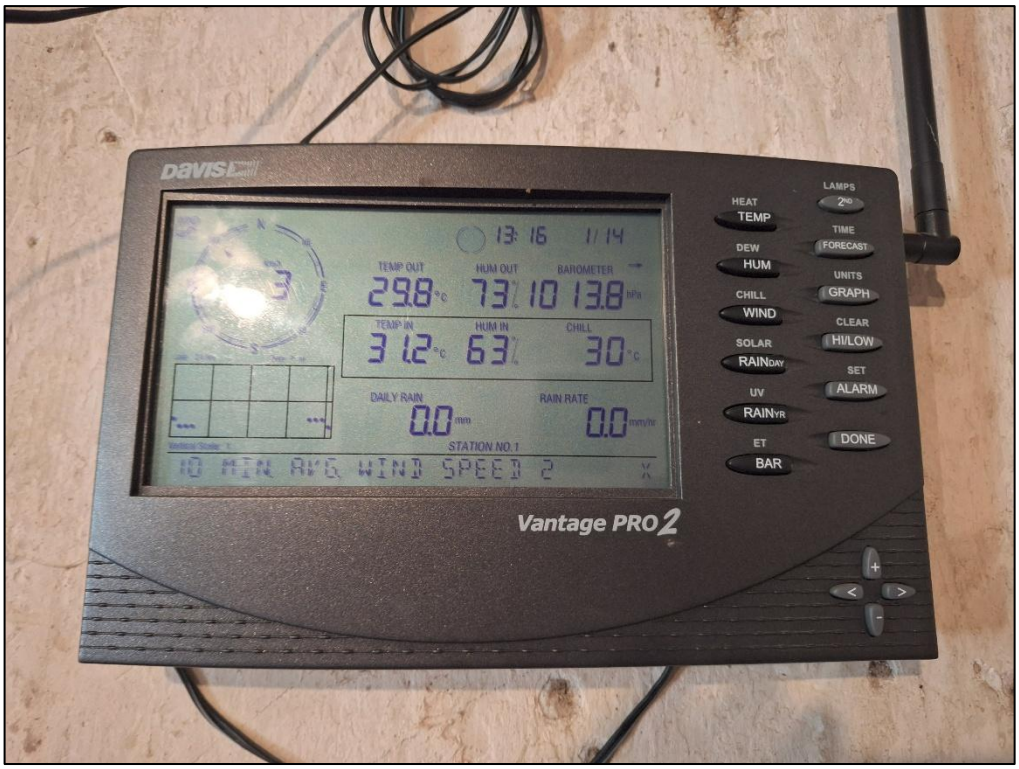


Figura 21

Estación de monitoreo La Merced



Figura 22

Mantenimiento y limpieza de la estación de monitoreo La Merced



Figura 23

Calibración y mantenimiento de la estación de monitoreo La Merced



Figura 24

Verificación y descarga de datos de la estación de monitoreo La Merced



Anexo 3. Instrumentos de recolección de datos

Tabla 19

Instrumentos de recolección de datos

Recolección de datos de parámetros meteorológicos													
Fecha	**Hora**	**Temperatura del aire (°C)**			**Velocidad del viento (m/s^{-1})**	**Humedad relativa (RH%)**	**Punto de rocío ºC/ºF**	**Velocidad del viento (m/s)**		**Dirección del viento**	**Presión barométrica (hPa/mb)**	**Precipitación acumulada (mm)**	**Radiación solar (W/m^2)**
		Temperatura máxima diaria (°C)	**Temperatura mínima diaria (°C)**	**Temperatura promedio (°C)**				**Viento instantáneo (m/s)**	**Viendo promedio periódico (m/s)**				

Nota. La tabla se adaptará a un Excel