CORRELACIÓN ENTRE COBERTURA VEGETAL Y NIVELES DE CONTAMINACIÓN DEL AIRE EN LOS ALREDEDORES DE CALI: UN ENFOQUE BASADO EN ANÁLISIS DE DATOS SATELITALES E INTELIGENCIA ARTIFICIAL.

Alejandro Villarreal Monsalve
Carlos Andrés Osorio Serna
Nicolás Méndez Gutiérrez

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director(a)
Yady Tatiana Solano Correa

Codirector(a)
Mario Patiño

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, 1 DE DICIEMBRE DEL 2025

TABLA DE CONTENIDO

Lista de figuras

Lista de tablas

INTRODUCCIÓN

La contaminación atmosférica en entornos urbanos representa un desafío crítico de salud pública a nivel global, siendo el incremento de material particulado (PM_{10} y $PM_{2.5}$) un factor de deterioro ambiental en ciudades en rápida expansión. En Santiago de Cali, el crecimiento urbano, el tráfico vehicular y las actividades industriales han generado un deterioro progresivo de la calidad del aire, con estaciones de monitoreo como Obrero y La Ermita registrando picos de concentración. Frente a esta problemática, la infraestructura verde urbana (IVU), conformada por la cobertura vegetal, es reconocida como un mecanismo natural de mitigación, actuando como filtro y modulador de la dispersión de contaminantes. No obstante, existe una brecha de conocimiento crítica en el contexto local: la ausencia de estudios que integren la información de monitoreo in-situ con el análisis detallado de la dinámica espacio-temporal de la vegetación, a la alta resolución requerida para entornos urbanos fragmentados, utilizando herramientas avanzadas de Inteligencia Artificial y Big Earth Data.

Para abordar esta brecha, el presente proyecto se planteó como Objetivo General:

Determinar la correlación entre la cobertura vegetal y los niveles de contaminación del aire en los alrededores de la ciudad de Cali utilizando técnicas avanzadas de aprendizaje automático e inteligencia artificial para la clasificación y análisis de imágenes satelitales y datos medioambientales, contribuyendo a la generación de conocimiento que facilite la toma de decisiones informadas en materia ambiental y urbana.

La metodología se estructuró en tres fases principales. Primero, se construyó una base de datos robusta mediante un *Data Pipeline* de tipo ETL, integrando imágenes satelitales PlanetScope con series temporales de contaminantes y variables meteorológicas del SVCASC. Segundo, se aplicaron algoritmos de *Machine Learning* (Random Forest, XGBoost, SVM) para la clasificación de coberturas y la modelación de la relación entre vegetación e índices de contaminación. Finalmente, se validaron los resultados mediante análisis estadísticos avanzados y una discusión biofísica de las implicaciones ambientales.

El documento se organiza en capítulos que reflejan las fases del proyecto. El Capítulo 4 detalla la Arquitectura de Datos y el proceso ETL, incluyendo la justificación de los sensores PlanetScope y la técnica de imputación avanzada con Random Forest Regressor. El Capítulo 5 aborda el Modelado Computacional y la evaluación de las arquitecturas de aprendizaje automático, determinando el modelo óptimo para la clasificación de coberturas. El Capítulo 6 presenta el Análisis Estadístico de la correlación espacio-temporal, arrojando el coeficiente global de asociación. Finalmente, el Capítulo 7 ofrece una Discusión y Validación de los hallazgos a la luz de la literatura internacional, y el Capítulo 8 presenta las conclusiones y los trabajos futuros.

1. DEFINICIÓN DEL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

En las últimas décadas, el impacto de la contaminación atmosférica se ha convertido en un tema crítico a nivel global debido a sus efectos adversos sobre la salud pública y el medio ambiente. Las áreas urbanas enfrentan desafíos particulares, ya que las fuentes de contaminación, como el tráfico vehicular y las actividades industriales, contribuyen significativamente a la concentración de partículas contaminantes, incluyendo PM10 y PM2.5. En este contexto, la vegetación urbana emerge como una solución natural con el potencial de mitigar estos impactos, al actuar como un filtro para los contaminantes atmosféricos. Sin embargo, la falta de información precisa y actualizada sobre la interacción entre cobertura vegetal y calidad del aire dificulta la implementación de estrategias efectivas que aprovechen esta relación para el beneficio de las comunidades urbanas.

La calidad del aire en los alrededores de Cali, entendiéndolo como diferentes puntos dentro de la ciudad, se ha convertido en una creciente preocupación debido a la emisión de partículas contaminantes originadas principalmente por actividades vehiculares e industriales [1]. Estas partículas, en especial PM10 y PM2.5, representan riesgos significativos para la salud pública, afectando especialmente a comunidades cercanas a áreas de alta densidad de tráfico y actividades industriales. Diversos estudios sugieren que la vegetación, particularmente los árboles, actúa como un filtro natural al captar y retener parte de estos contaminantes, contribuyendo así a mejorar la calidad del aire en zonas urbanas y periurbanas [2]. Sin embargo, en los alrededores de Cali, la distribución y cantidad de vegetación es desigual, lo que plantea un desafío para identificar y aprovechar efectivamente estas áreas verdes en la mitigación de la contaminación atmosférica.

La ausencia de un mapeo detallado y actualizado sobre la correlación entre la cobertura vegetal y los niveles de contaminación del aire en esta región limita la capacidad de los entes gubernamentales y la Corporación Autónoma Regional del Valle del Cauca (CVC) para diseñar políticas públicas efectivas y sostenibles. El conocimiento profundo de estas correlaciones permitiría mejorar la planificación urbana y promover estrategias de sostenibilidad ambiental, que a su vez beneficiarían la calidad de vida y el bienestar de las comunidades afectadas.

1.2 FORMULACIÓN DEL PROBLEMA

En este contexto, surge la necesidad de profundizar en la relación entre la vegetación y la calidad del aire, lo que nos lleva a plantearnos la siguiente pregunta de investigación: **¿Qué tan significativa es la correlación entre la densidad de vegetación y los niveles de contaminación del aire en distintas zonas de los alrededores de Cali?**

En este marco de análisis, resulta esencial abordar interrogantes clave que permitan guiar el

desarrollo del proyecto. A continuación, se presentan las preguntas de sistematización que orientarán la investigación:

1. ¿Cómo se puede construir una base de datos que integre información satelital, datos de estaciones meteorológicas y niveles de contaminación para analizar la calidad del aire en relación con la cobertura vegetal?
2. ¿Qué técnicas avanzadas de modelación, basadas en inteligencia artificial, son las más efectivas para clasificar la cobertura vegetal y cuantificar niveles de contaminación del aire para posteriormente determinar la correlación entre estas en los alrededores de Cali?
3. ¿Como podría validarse la correlación de la cobertura vegetal en la calidad del aire en los alrededores de Cali?

2. OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Determinar la correlación entre la cobertura vegetal y los niveles de contaminación del aire en los alrededores de la ciudad de Cali utilizando técnicas avanzadas de aprendizaje automático e inteligencia artificial para la clasificación y análisis de imágenes satelitales y datos medioambientales, contribuyendo a la generación de conocimiento que facilite la toma de decisiones informadas en materia ambiental y urbana.

2.2 OBJETIVOS ESPECÍFICOS

1. Desarrollar una base de datos integrando información satelital de cobertura vegetal, datos de estaciones meteorológicas y niveles de contaminación atmosférica, con una temporalidad de uno a tres años.
2. Aplicar métodos de aprendizaje automático para clasificar coberturas vegetales y cuantificar niveles de contaminación de aire para determinar la correlación entre las variables en los alrededores de Cali.
3. Validar la correlación entre la cobertura vegetal y la calidad del aire en los alrededores de Cali para respaldar la toma de decisiones en el manejo ambiental.

3. MARCO TEÓRICO Y ANTECEDENTES

3.1 MARCO TEÓRICO

3.1.1 CALIDAD DEL AIRE

La contaminación atmosférica es un fenómeno complejo que afecta diferentes escalas espaciales: global, mesoescala y microescala. Cada una tiene características y consecuencias específicas para la salud humana, el medio ambiente y los ecosistemas [3].
La contaminación atmosférica escala global tiene efectos acumulativos y de largo alcance la cual impacta sistemas planetarios como la atmósfera y el clima, esto contribuido por el calentamiento global y el cambio climático debido a que con esto se alteran ecosistemas, aumentando la frecuencia de eventos meteorológicos extremos, colocando en riesgo el acceso al agua y a no poder tener una seguridad alimentaria. Como principal contaminante incluyen los gases de efecto invernadero como el dióxido de carbono (CO_2) y el metano (CH_4) y las partículas de carbono negro (BC).
Es importante resaltar que, aunque la contaminación atmosférica escala global no afecta directamente a la salud humana, sí tiene repercusiones indirectas, ya que afecta la capa de ozono estratosférico por sus compuestos halogenados como los clorofluorocarbonos (CFC) aumentando así, la radiación ultravioleta los cuales generan problemas de salud y daños ecológicos.
La contaminación atmosférica de mesoescala tiene efectos desde decenas hasta miles de kilómetros, es decir este puede impactar hasta regiones diferentes alejadas de las fuentes emisoras, ya que viajan partículas originada por las emisiones de dióxido de azufre (SO_2) y óxidos de nitrógeno (NOX) que se transforman en ácidos sulfúrico (H_2SO_4) y nítrico (HNO_3) generando así lluvias ácidas, acidificando suelos y cuerpos de agua, con graves consecuencias ecológicas.
Otro caso es el ozono troposférico (O_3), un contaminante secundario que no es emitido directamente, sino que se forma a partir de reacciones fotoquímicas entre los NOX y los compuestos orgánicos volátiles (COVs). Este fenómeno se da a sotavento de zonas urbanas e industriales y puede impactar regiones alejadas de las fuentes emisoras, afectando tanto la salud humana como los ecosistemas.
La contaminación atmosférica en la microescala es una de las principales causas del deterioro de la calidad del aire, impactando directamente en la zonas urbanas, rurales e industriales ya que son las fuentes en donde se originan estas emisiones, esto debido a que, a son fuentes puntuales y localizadas las cuales generan contaminantes específicos que afectan la salud humana y el medio ambiente.
A continuación, se mencionan algunas de las principales fuentes de la contaminación que

se encuentran en la microescala:

- Tráfico rodado: Se produce mediante los vehículos a motor y se presentan en zonas rurales altamente pobladas generando partículas PM_{10} y $PM_{2,5}$, partículas ultrafinas (UFP), óxidos de nitrógeno (NOX), monóxido de carbono (CO) y compuestos orgánicos volátiles (COVs), entre otros. Las partículas mencionadas anteriormente incrementan el riesgo para la salud respiratoria y cardiovascular de las personas.
- Fuentes domésticas e institucionales: Uso de combustibles fósiles las cuales generan gases esto se da mediante el uso de sistemas de calefacción y/o cocción que utilizan combustibles.
- Actividades industriales y centrales térmicas: Emisiones de metales pesados, amoníaco (NH_3) y partículas ultrafinas (UFP) que afectan la calidad del aire local y regional.
- Agricultura y ganadería: La aplicación de fertilizantes y el manejo de desechos animales liberan amoníaco (NH_3) y otras sustancias que contribuyen al deterioro de la calidad del aire.
- Construcción y demolición: Estas actividades liberan polvo y partículas suspendidas, generando riesgos inmediatos para las personas cercanas a las obras.

3.1.2 MEDICIÓN CALIDAD DEL AIRE

La medición de la calidad del aire se realiza mediante un proceso esencial en donde se evalúa la concentración de contaminantes atmosféricos, identificando fuentes para diseñar estrategias que permitan mitigar aquellos efectos que impactan en la salud humana y ecosistemas. Estas mediciones se realizan por medio de tecnologías especializadas y metodologías estandarizadas en las cuales se cuantifican diversos contaminantes.

En Colombia la medición de la calidad del aire está regulada por el Ministerio de Ambiente y Desarrollo Sostenible, bajo el marco de la Resolución 2254 de 2017 [4], la cual establece los estándares de calidad del aire y los lineamientos para el monitoreo. Esta medición se realiza mediante Redes de Monitoreo de Calidad del Aire (RCMA) gestionadas por las autoridades ambientales regionales y locales.

- MEDICIÓN DE LA CALIDAD DEL AIRE EN LA CIUDAD DE CALI

En Santiago de Cali, el Sistema de Vigilancia de Calidad del Aire de Santiago de Cali (SVCASC) opera bajo la administración y coordinación del Grupo de Calidad del Aire del Departamento Administrativo de Gestión del Medio Ambiente (DAGMA) [5]. Su objetivo principal es medir constantemente aquellos contaminantes y variables meteorológicas en

diferentes puntos estratégicos garantizando el aire de la ciudad mediante estrategias de mejoramiento ambiental.

Acciones del SVCASC:

- Monitoreo y mantenimiento de equipo: Actualmente se cuenta con 9 estaciones Tabla 1 Estaciones Meteorológicas de Cali.)de monitoreo que miden contaminantes del aire como el material particulado (PM_{10} y $PM_{2.5}$), ozono troposférico (O_3), dióxido de azufre (SO_2), dióxido de nitrógeno (NO_2), sulfuro de hidrógeno (H_2S). Además de variables meteorológicas como la velocidad y dirección del viento, temperatura, humedad relativa, radiación solar, presión barométrica y precipitación.
- Calibración y medición: Se calibran semanalmente con el fin de tener mediciones más acertadas, los datos generados se analizan y evalúan estadísticamente antes de publicar la información.
- Gestión y reportes: Se generan informes en el que se describe el comportamiento de la calidad del aire y se mantiene actualizado el inventario de emisiones atmosféricas de manera mensual y anual para acceso público.
- Control y Seguimiento de Fuentes Móviles: mediante alianza con secretaría de tránsito se realizan controles operativos de inspección vehicular para validar el cumplimiento de la norma de emisión atmosférica, adicional se supervisa mediante los Centro de Diagnósticos Automotor CDA dando certificaciones técnico-mecánicos.

Tabla 1 Estaciones Meteorológicas de Cali.

ID	Estación	Barrio - Zona	Variables que mide
1	La Flora	La Flora - Norte	Material Particulado PM_{10}, Sulfuro de Hidrógeno H_2S
2	ERA - Obrero	Obrero - Centro	Material Particulado PM_{10}
3	Transitoria EDB-Navarro	El Poblado - Oriente	Material Particulado PM_{10}, Sulfuro de Hidrógeno H_2S
4	Base Aérea	La Base - Nororiente	Material Particulado $PM_{2.5}$, Ozono Troposférico O_3, Dióxido de Azufre SO_2
5	Pance	Pance - Rural	Ozono Troposférico O_3
6	Univalle	Meléndez - Sur	Material Particulado $PM_{2.5}$, Ozono Troposférico O_3, Dióxido de Nitrógeno NO_2

7	Compartir	Compartir - Oriente	Material Particulado $PM_{2.5}$, Ozono Troposférico O_3
8	La Ermita	La Ermita - Centro	Material Particulado PM_{10}
9	Cañaveralejo	Cañaveralejo - Suroriente	Material Particulado PM_{10}, Dióxido de Azufre SO_2

3.1.3 TELEDETECCIÓN

La teledetección es la ciencia y tecnología orientada a obtener información de la superficie terrestre sin requerir contacto físico directo con el objeto o área observada, mediante la captura y análisis de la energía electromagnética reflejada o emitida por los elementos presentes en la superficie [6]. Este proceso se basa en la interacción entre la radiación solar y los diferentes tipos de cobertura, cuyos patrones espectrales pueden ser registrados por sensores remotos instalados en plataformas satelitales, aéreas o UAV, generando imágenes multiespectrales o hiperespectrales que permiten identificar, caracterizar y monitorear fenómenos ambientales en distintas escalas espaciales y temporales [7]. La teledetección constituye un insumo fundamental en estudios de dinámica urbana y análisis de cobertura vegetal, ya que posibilita el cálculo de índices espectrales como el NDVI, la detección de cambios en el territorio y la integración con modelos de aprendizaje automático para la clasificación temática y predicción de variables ambientales [8]. En el contexto de ciudades latinoamericanas en expansión, la teledetección satelital ofrece ventajas como alta resolución temporal, acceso a series históricas y cobertura completa del territorio, lo que la convierte en una herramienta esencial para evaluar la relación entre vegetación y contaminación atmosférica, como se plantea en el presente proyecto aplicado.

3.1.3.1 IMAGEN SATELITAL

Una imagen satelital es el resultado de capturar la radiación electromagnética emitida o reflejada por la superficie terrestre, esto mediante sensores que son instalados en satélites artificiales. Estas imágenes se transmiten a estaciones terrestres con el fin de procesarlas y analizarlas en diversos contextos. Una imagen se compone de la unión de bandas espectrales, según la energía recibida en distintas longitudes de onda del espectro electromagnético, algunas de estas bandas son el espectro visible, infrarrojo cercano (NIR), infrarrojo medio (SWIR), entre otras; Un sensor puede tener la capacidad de capturar cada banda de manera independiente, siendo que la información en cada una de estas se puede utilizar para diferentes análisis. La forma en que el sensor registra los datos es en una matriz de píxeles, donde cada píxel representa un área específica de la superficie terrestre y su valor numérico es la radiancia detectada en esa porción de la superficie [9].

Existen varios tipos de sensores que se pueden diferenciar según ciertas características:

- Resolución espacial: Es el tamaño del área representada por cada pixel en la imagen, se puede tener una alta resolución si el área capturada es muy detallada (5m, 10 m por pixel) o una baja resolución si el área capturada por pixel es muy grande (1 km por pixel).
- Resolución espectral: Cantidad y rango de bandas espectrales que el sensor puede captar (Multiespectrales, Hiperespectrales).
- Resolución temporal: Frecuencia en la que el sensor captura imágenes en una misma ubicación (cada 1 día, 5 días, etc.)
- Resolución radiométrica: capacidad de un sensor para detectar variaciones en la intensidad de radiación. Expresada en bits.

En el marco de este proyecto se emplearon imágenes satelitales PlanetScope, las cuales poseen una resolución espacial aproximada de 3.7 m^2 por pixel y cuatro bandas espectrales (azul, verde, rojo y NIR) [10], características que permiten una identificación detallada de coberturas vegetales en entornos urbanos. La alta frecuencia temporal de este sensor posibilitó seleccionar escenas libres de nubosidad para los periodos de estudio, garantizando la consistencia temporal requerida para el análisis multitemporal. La elección de PlanetScope se fundamentó en su capacidad para capturar variaciones espaciales de vegetación en áreas fragmentadas como las que rodean a las estaciones de monitoreo de Cali, constituyéndose en la fuente primaria de información para las etapas de clasificación descritas en capítulos posteriores.

3.1.4 PROCESAMIENTO DE IMÁGENES

El procesamiento de imágenes es una disciplina que busca aplicar distintas operaciones o algoritmos sobre una imagen digital con el fin de reconocer datos o patrones de interés. Una imagen es la representación de un objeto o espacio real (Tres dimensiones) en un plano (Dos dimensiones). Al considerar la imagen como un plano se puede establecer una función $f_{(x, y)}$ donde (x, y) son las coordenadas del plano y f es el valor de luminosidad (en escala de grises) en ese punto. Es importante mencionar que para poder procesar la imagen por computador no se tiene como tal una función continua $f_{(x, y)}$ sino un conjunto finito de puntos discretos, píxeles, que toman un valor según su nivel de luminosidad, color u otros atributos [11].

3.1.4.1 ETAPAS DEL PROCESAMIENTO

1. Captura
   Proceso donde se toma una imagen digital para ser procesada por una computadora.
2. Preprocesamiento

   Conjunto de técnicas de filtrado aplicadas sobre cada píxel de la imagen, para corregir imperfecciones u optimizarla (Eliminar ruido, suavizar, etc.).
3. Segmentación

   Divide la imagen en regiones o segmentos según color, luminosidad u otros patrones.
4. Extracción de características

   Proceso de describir y reconocer elementos de interés presentes en la segmentación, dependiendo de sus características geométricas.
5. Identificación de objetos

   Proceso de reconocimiento automatizado. Se apoya en algoritmos de toma de decisiones y el trabajo realizado en las etapas anteriores para poder reconocer características de interés automáticamente.

3.1.5 CLASIFICACIÓN DE IMÁGENES SATELITALES

La clasificación de imágenes satelitales es un procedimiento fundamental en teledetección que permite asignar a cada píxel una categoría temática de acuerdo con su comportamiento espectral. En este proyecto se emplea la clasificación supervisada, en la cual el modelo aprende a partir de muestras previamente etiquetadas para diferenciar tipos de cobertura como áreas urbanas, bosques, pastizales, suelos desnudos y cuerpos de agua. Este proceso se basa en el análisis de firmas espectrales, entendidas como el conjunto de valores de reflectancia que caracterizan a cada tipo de superficie en distintas longitudes de onda [12]. Los algoritmos de aprendizaje automático como Random Forest, Support Vector Machines y XGBoost han demostrado un desempeño sobresaliente en la clasificación de coberturas y usos del suelo en imágenes satelitales, debido a su capacidad para manejar relaciones no lineales, trabajar con un número limitado de muestras de entrenamiento y mantener una alta exactitud [13] [14]. La validación de la clasificación se realiza mediante métricas como la exactitud global y el coeficiente Kappa, que permiten cuantificar el nivel de acuerdo entre las etiquetas predichas y las observadas en datos de referencia, asegurando que los mapas obtenidos representen de manera confiable la

distribución espacial de la cobertura vegetal. Este insumo es esencial para analizar la evolución temporal de la vegetación y su relación con los niveles de contaminación atmosférica registrados en las estaciones seleccionadas.

En el contexto de este proyecto, la clasificación supervisada se seleccionó debido a la disponibilidad de muestras de entrenamiento obtenidas sobre áreas de referencia previamente validadas. Este enfoque permitió asignar una etiqueta temática a cada píxel de las imágenes PlanetScope, diferenciando categorías como vegetación, suelo desnudo, áreas urbanizadas y cuerpos de agua. Los algoritmos empleados, entre ellos Random Forest, XGBoost y SVM, han demostrado un desempeño consistente en la clasificación de coberturas en entornos urbanos debido a su capacidad para manejar relaciones no lineales y evitar el sobreajuste en escenarios con un número limitado de muestras de entrenamiento. La precisión de la clasificación se evaluó mediante exactitud global y coeficiente Kappa, lo cual aseguró que los mapas generados representaran de manera confiable la distribución espacial de la cobertura vegetal, insumo fundamental para el análisis de correlación desarrollado en las secciones posteriores.

3.1.5.1 ÍNDICES DE VEGETACIÓN

En teledetección, a partir de una imagen, un índice de vegetación es una operación matemática entre valores de reflectancia en distintas longitudes de onda que tienen cambios significativos dependiendo de la cobertura vegetal; Se asume que, el valor obtenido representa de alguna manera la cantidad de vegetación en el área de la imagen (Píxel o píxeles) donde se realizó la operación [15]. A continuación, se definen los principales índices de vegetación:

1. NDVI (Normalized Difference Vegetation Index) - Índice Normalizado de Vegetación: Es el índice más utilizado para evaluar la densidad y el vigor de la vegetación viva. Valores altos (cercanos a 1) indican vegetación densa y saludable. Valores bajos (cercanos a 0 o negativos) indican suelos desnudos, agua u otras superficies no vegetadas.

$$NDVI = \frac{NIR - R}{NIR + R} \quad [1]$$

Donde:

NIR: Reflectancia en el infrarrojo cercano.

R: Reflectancia en la banda roja.

2. GNDVI (Green Normalized Difference Vegetation Index) - Índice de Vegetación de Diferencia Normalizada Verde: Similar al NDVI, pero utiliza la banda verde en lugar de la roja, lo que mejora la detección de clorofila en ciertos casos y es más sensible a niveles intermedios de clorofila.

$$GNDVI = \frac{NIR - G}{NIR + G} \quad [2]$$

Donde:

NIR: Reflectancia en el infrarrojo cercano.

G: Reflectancia en la banda verde.

3. RVI (Ratio Vegetation Index): Es un índice más simple que el NDVI, basado en el cociente de reflectancia. Indica la proporción entre la reflectancia en el infrarrojo cercano y la roja. Valores altos indican mayor cantidad de vegetación.

$$RVI = \frac{NIR}{R} \quad [3]$$

Donde:

NIR: Reflectancia en el infrarrojo cercano.

R: Reflectancia en la banda roja.

3.1.5.2 INTELIGENCIA ARTIFICIAL

La inteligencia artificial (IA) es un campo interdisciplinario enfocado en el desarrollo de teorías, métodos y tecnologías para simular y expandir las capacidades cognitivas humanas. Desde su formalización en 1956 por John McCarthy [16], ha evolucionado a través de hitos como el aprendizaje automático (Machine Learning, ML) y el aprendizaje profundo (Deep Learning, DL), que han permitido resolver problemas complejos en áreas como visión por computadora, procesamiento del lenguaje natural y sistemas autónomos [17].

El ML se basa en técnicas que permiten a las máquinas aprender patrones a partir de datos sin necesidad de programar explícitamente reglas para cada tarea. Entre sus principales métodos se encuentran los algoritmos supervisados, como los árboles de decisión y las máquinas de soporte vectorial (SVM), y los algoritmos no supervisados, como el agrupamiento (clustering) con k-means. También incluye técnicas de aprendizaje por refuerzo, como Q-Learning y políticas basadas en gradientes, que permiten aprender estrategias mediante retroalimentación de recompensas. Estas metodologías se han aplicado con éxito en tareas como clasificación, predicción y optimización en dominios como la salud, la educación y la industria. En el marco de este proyecto, el interés no se centra en todas las ramas de la inteligencia artificial, sino específicamente en aquellas técnicas de aprendizaje automático que permiten integrar datos satelitales y ambientales para modelar la relación entre cobertura vegetal y calidad del aire.

En el contexto de este proyecto, la inteligencia artificial se aplica mediante técnicas de aprendizaje automático supervisado para dos propósitos principales: la clasificación de coberturas vegetales y la modelación de la relación entre índices de vegetación y niveles de contaminación atmosférica. Modelos como Random Forest y XGBoost fueron seleccionados debido a su capacidad para manejar múltiples variables predictoras, su resistencia al sobreajuste y su eficacia en la interpretación de importancia de características [14]. Las máquinas de soporte vectorial (SVM), configuradas mediante funciones kernel no lineales, permiten capturar patrones complejos entre características espectrales y concentraciones de partículas en contextos donde las fronteras de decisión son estrechas o los datos están parcialmente solapados [18]. Adicionalmente, se implementó una red neuronal artificial con el fin de explorar interacciones no lineales de mayor complejidad entre las variables ambientales; no obstante, su desempeño depende de la disponibilidad de volúmenes adecuados de datos y de la optimización cuidadosa de hiperparámetros. La evaluación sistemática de estos modelos mediante validación cruzada y métricas de desempeño asegura que los patrones identificados entre cobertura vegetal y calidad del aire sean estadística y computacionalmente robustos, aportando evidencia sólida para la toma de decisiones en gestión ambiental urbana.

Fundamento Teórico de la Imputación RF:

El algoritmo opera construyendo una multitud de árboles de decisión no correlacionados durante el entrenamiento. Para el problema de imputación, se trató cada variable con datos faltantes (por ejemplo, $PM_{2.5}$ en el tiempo t) como la variable objetivo (Y), utilizando el resto de variables disponibles en ese mismo instante t (Temperatura, Viento, Hora, Día de la semana, Otros Contaminantes) como predictores (X). El modelo predice el valor faltante promediando las salidas de todos los árboles individuales en el bosque, lo que

reduce la varianza del error y previene el sobreajuste (overfitting).

La predicción Y^ para un dato faltante se obtiene mediante la agregación:

$$G\hat{Y} = \frac{1}{K}\sum\_\{k=1\}^{\{K\}h}(X) \quad [9]$$

Donde K es el número total de árboles en el bosque y $h_k(X)$ es la predicción individual del k-ésimo árbol.

3.1.6 CORRELACIÓN DE VARIABLES

La El análisis de correlación constituye un componente fundamental en la inferencia estadística ambiental, permitiendo cuantificar la magnitud y la dirección de la relación entre dos variables aleatorias continuas o discretas. En el contexto de la interacción atmósfera-biósfera, donde coexisten variables biofísicas (como los índices de vegetación) y concentraciones de contaminantes, la evaluación de la dependencia estadística requiere un enfoque multidimensional que trascienda la linealidad simple [19].

A continuación, se describen los métodos estadísticos seleccionados para evaluar la hipótesis de interdependencia entre la infraestructura verde y la calidad del aire:

Coeficiente de Correlación de Pearson

El coeficiente de correlación producto-momento de Pearson es la medida paramétrica estándar para evaluar el grado de asociación lineal entre dos variables cuantitativas con distribución normal conjunta. Este coeficiente, denotado como *r*, normaliza la covarianza de las variables por el producto de sus desviaciones estándar, resultando en un valor adimensional en el intervalo [-1, 1] [20].
Matemáticamente, para una muestra de tamaño *n* con pares de datos (x_i, y_i), se define como:

$$r = \frac{\sum_{i=1}^{n}(Xi - X)(Yi - Y)}{\sqrt{\sum_{i=1}^{n}(Xi - X)^2}\sqrt{(Yi - Y)^2}} \quad [4]$$

$$r = \Sigma 1ni = 1(Xi - X)(Yi - Y)\sqrt{\Sigma 1ni = 1(Xi - X)2}\sqrt{(Yi - Y)2} \quad [5]$$

donde:

Xi e Yi: son los valores individuales de las variables X e Y.

X y Y: son las medias aritméticas de las variables X e Y, respectivamente.

n: es el número total de observaciones.

Sin embargo, el uso exclusivo de Pearson presenta limitaciones en estudios ambientales, dado que es altamente sensible a valores atípicos (*outliers*) y asume homocedasticidad y linealidad, condiciones que raramente se cumplen estrictamente en series temporales de contaminación urbana afectadas por picos agudos [21].

Coeficiente de correlación de rangos de Spearman

Dada la naturaleza estocástica y a menudo no paramétrica de los datos de calidad del aire (con distribuciones sesgadas a la derecha), se incorpora el coeficiente de correlación de Spearman (ρ). Este método no paramétrico evalúa la relación monótona entre variables, basándose en el rango u orden de los datos en lugar de sus valores crudos, lo que lo hace robusto frente a *outliers* y no requiere el supuesto de normalidad [22].

Correlación Cruzada

Los fenómenos ambientales poseen una dimensión temporal intrínseca; el efecto de la vegetación o de variables meteorológicas (como el lavado por lluvia) sobre la calidad del aire puede no ser inmediato. La función de correlación cruzada (CCF) permite identificar la dependencia entre dos series temporales desplazadas en el tiempo, detectando retardos (*lags*) significativos [23].

3.1.7 MÉTRICAS DE VALIDACIÓN

Además del coeficiente de correlación de Pearson, ampliamente utilizado para evaluar relaciones lineales entre variables cuantitativas [24], en este trabajo se consideran otras métricas en contextos específicos, como la correlación de Spearman [25], que resulta adecuada para analizar relaciones monótonas no lineales a partir de rangos, y la correlación de Kendal [26], útil en estudios con datos ordinales o con presencia de valores atípicos. Sin embargo, la interpretación de la correlación se complementa con métricas de

validación de modelos, particularmente cuando se emplean técnicas de aprendizaje automático para cuantificar la relación entre cobertura vegetal y calidad del aire. La evaluación del desempeño de los modelos utilizados en este proyecto requiere la incorporación de métricas de validación que permitan medir de manera objetiva la precisión y confiabilidad de los resultados. En tareas de clasificación de coberturas vegetales, se emplean indicadores como la exactitud global (accuracy), que refleja el porcentaje total de aciertos; la precisión (precision) y el recall, que permiten analizar el comportamiento del modelo frente a falsos positivos y falsos negativos; y la medida F1, que combina precisión y recall para ofrecer un balance entre ambos extremos, especialmente útil en contextos con clases desbalanceadas [12] [18]. Adicionalmente, la matriz de confusión permite identificar patrones de error entre categorías con firmas espectrales similares, mientras que el coeficiente Kappa corrige la exactitud por el efecto del azar, proporcionando una estimación más robusta del desempeño del modelo [27]. Para los modelos de regresión aplicados a la estimación de contaminantes atmosféricos se utilizan métricas como el error absoluto medio (MAE), el error cuadrático medio (RMSE) y el coeficiente de determinación (R^2), las cuales cuantifican la diferencia entre valores predichos y observados y permiten evaluar el ajuste del modelo. El uso combinado de estas métricas garantiza que los resultados obtenidos no solo reflejen asociaciones estadísticas, sino también un desempeño computacional confiable, fortaleciendo la validez de las conclusiones sobre la relación entre cobertura vegetal y calidad del aire.

3.2 INTRODUCCIÓN AL PARADIGMA DE CLASIFICACIÓN ESPECTRAL SUPERVISADA

La transformación de los datos crudos de observación de la Tierra en información temática procesable constituye el núcleo algorítmico y metodológico de la presente investigación. Tras la consolidación de la arquitectura de datos y la ejecución de una rigurosa ingeniería de características (ETL) detallada en los capítulos precedentes, el presente Capítulo 5 aborda la fase crítica de modelado computacional. En este estadio, el objetivo fundamental trasciende la mera representación radiométrica de la superficie terrestre —expresada hasta este punto en valores de reflectancia de superficie (BOA) calibrados física y atmosféricamente— para construir una abstracción matemática robusta. Esta abstracción debe ser capaz de discernir, con alta precisión y consistencia estadística, las complejas categorías semánticas que componen el mosaico urbano de Santiago de Cali.

La complejidad inherente a los entornos urbanos neotropicales plantea desafíos significativos de optimización para los algoritmos de clasificación tradicionales y paramétricos. A diferencia de los paisajes templados o las zonas agrícolas homogéneas, la heterogeneidad espacial de una ciudad como Cali se caracteriza por una yuxtaposición caótica de materiales sintéticos y biológicos en escalas sub-píxel. En un espacio geográfico reducido, interactúan superficies de alta impedancia térmica (concreto, asfalto, polímeros

de techos) con elementos de biomasa activa (vegetación arbórea, herbácea), suelos desnudos y cuerpos de agua eutrofizados. Esta configuración genera dos fenómenos estadísticos adversos que dificultan la separabilidad lineal en el hiperespacio espectral:

1. **Alta varianza intra-clase:** Una misma categoría semántica, como "Urbano", posee firmas espectrales multimodalmente diversas (e.g., la reflectancia de una teja de arcilla nueva difiere drásticamente de un asfalto envejecido o concreto meteorizado), lo que impide su agrupación en clústeres compactos [28].
2. **Sutil varianza inter-clase:** Categorías funcionalmente distintas pueden converger radiométricamente (e.g., pastos senescentes vs. suelo desnudo laterítico), generando solapamientos en las bandas del espectro visible que requieren dimensiones adicionales para su desambiguación [29].

Además, la presencia de factores estocásticos como las sombras proyectadas por la infraestructura vertical ("cañones urbanos"), la variabilidad fenológica de la vegetación tropical y la contaminación atmosférica introducen un nivel de ruido aleatorio que los modelos lineales simples son incapaces de resolver adecuadamente. En consecuencia, este capítulo documenta de manera exhaustiva el diseño experimental, la fundamentación matemática profunda, la implementación técnica y la evaluación comparativa de cinco arquitecturas de aprendizaje automático (*Machine Learning*) de distinta naturaleza inductiva: métodos de ensamble basados en árboles (*Random Forest* y *XGBoost*), modelos basados en vectores de soporte (*SVM*), redes neuronales artificiales (*Perceptrón Multicapa - MLP*) y modelos lineales generalizados (*Regresión Logística*).
El propósito de esta evaluación multi-modelo no es meramente identificar el algoritmo con mayor exactitud predictiva global (*Accuracy*), sino deconstruir su comportamiento interno para comprender cómo las variables espectrales (bandas visibles e infrarrojas) y espaciales (textura Gabor) interactúan para explicar la distribución de la cobertura vegetal. Estudios recientes sugieren que, si bien algoritmos como XGBoost han demostrado superioridad en tareas de regresión, su complejidad no siempre justifica la ganancia marginal sobre métodos más estables como Random Forest en tareas de clasificación con datos ruidosos [30]. Por tanto, la selección del modelo óptimo se sustenta en una búsqueda exhaustiva de hiperparámetros (*Grid Search*) y una validación cruzada estratificada de k pliegues, garantizando que los mapas de cobertura derivados posean la validez estadística necesaria para inferir causalidad en los modelos de calidad del aire subsiguientes.

3.3 FUNDAMENTACIÓN TEÓRICA Y MATEMÁTICA DE LOS ALGORITMOS EVALUADOS

Para justificar la selección y configuración de las arquitecturas sometidas a prueba, es

imperativo profundizar en los principios matemáticos que gobiernan su proceso de inferencia. La naturaleza de los datos satelitales PlanetScope —multiespectrales, de alta resolución espacial (3 m) pero limitada resolución espectral (4 bandas)— exige algoritmos capaces de explotar relaciones no lineales, manejar la colinealidad entre las variables predictoras (e.g., la alta correlación de Pearson entre la banda Roja y la Verde) y resistir el ruido de etiqueta inherente a los procesos de muestreo supervisado.

3.3.1 RANDOM FOREST: EL PARADIGMA DEL ENSAMBLE Y LA REDUCCIÓN DE VARIANZA

El algoritmo *Random Forest* (RF), propuesto seminalmente por [31], opera bajo el principio de *Bagging* (*Bootstrap Aggregating*). En el contexto de la clasificación de imágenes satelitales, RF aborda uno de los problemas fundamentales de los árboles de decisión individuales (CART): su alta varianza y tendencia al sobreajuste (*overfitting*) cuando se enfrentan a datos con geometrías de decisión complejas o ruidosas.

Mecanismo Matemático de Construcción

Sea $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ el conjunto de entrenamiento donde $x_i \in R^p$ representa el vector de características espectrales (R, G, B, NIR, NDVI, Gabor) y y_i la etiqueta de clase discreta. RF genera un ensamble de B árboles de decisión independientes $\{T_1, \ldots, T_{B\backslash}\}$. Cada árbol T_b se entrena sobre una muestra D_b obtenida mediante muestreo con reemplazo (*bootstrap*) del conjunto original D. Este proceso de remuestreo asegura que cada estimador base "vea" una versión ligeramente perturbada de la realidad estadística, introduciendo diversidad esencial en el modelo global.

La innovación crítica de RF, que lo distingue del *Bagging* simple, radica en la inyección de aleatoriedad adicional durante la división de los nodos (Random Subspace Method). En lugar de buscar la mejor división determinista entre todas las p variables predictoras disponibles, cada nodo del árbol evalúa solo un subconjunto aleatorio de m variables, donde típicamente se define $m \approx \sqrt{\{p\}}$ para tareas de clasificación. Esta estrategia descorrelaciona estadísticamente los árboles del bosque, asegurando que el ensamble capture diversas estructuras subyacentes de los datos y no se vea dominado por una única variable predictora fuerte (como podría ser el NDVI), mitigando así el riesgo de sesgo estructural [29].

Criterio de División: Impureza de Gini

La pureza de los nodos durante el crecimiento del árbol se maximiza minimizando una función de impureza. En este estudio, dada la naturaleza categórica de la variable objetivo, se emplea el Índice Gini. Para un nodo t con una distribución de clases estimada $p(k|t)$, el índice Gini se define como:

$$I_{G(t)} = 1 - \backslash sum_{k=1}^{\{K\}} \; p(k|t)^2 \quad [10]$$

Donde K es el número de clases de cobertura (5 en este diseño experimental). El algoritmo busca recursivamente la división óptima θ (par variable y umbral de corte) que maximice la ganancia de información, definida como la reducción ponderada de la impureza:

$$\Delta I(\theta, t) = I_G(t) - \{N_L\}/\{N\} I_G(t_L) - \{N_R\}/\{N\} I_G(t_R) \quad [11]$$

Donde N_L y N_R son el número de muestras en los nodos hijos izquierdo y derecho, respectivamente, y N es el total de muestras en el nodo padre. Este proceso de particionamiento recursivo continúa hasta alcanzar un criterio de parada, como la profundidad máxima (*max_depth*) o un número mínimo de muestras por hoja, hiperparámetros cuya optimización es crucial para balancear la capacidad de generalización del modelo.

Inferencia por Votación Mayoritaria

Para clasificar un nuevo píxel x', el ensamble agrega las predicciones de todos los árboles individuales mediante votación mayoritaria (moda estadística):

$$\hat{y} = moda\left\{ T_{b(x')\backslash} \right\}_{\{b=1\}}^{\{B\}} \quad [12]$$

Esta agregación reduce teóricamente la varianza del estimador global sin incrementar significativamente el sesgo. Esta propiedad hace que RF sea particularmente robusto frente al ruido de tipo "sal y pimienta" común en imágenes urbanas y eficiente en la gestión de *outliers* espectrales o píxeles mixtos que confundirían a clasificadores basados en máxima verosimilitud.

3.3.2 XGBOOST: OPTIMIZACIÓN SECUENCIAL MEDIANTE BOOSTING DE GRADIENTE

El algoritmo *Extreme Gradient Boosting* (XGBoost), descrito formalmente por [32], representa una evolución sofisticada del paradigma de ensamble. A diferencia de RF, que construye árboles en paralelo e independientes para reducir varianza, XGBoost opera secuencialmente bajo el marco de *Boosting*. La premisa fundamental es que cada nuevo árbol se entrena específicamente para predecir y corregir los errores residuales (la diferencia negativa del gradiente de la función de pérdida) de los árboles precedentes,

transformando iterativamente un conjunto de aprendices débiles (*weak learners*) en un predictor fuerte.

Función Objetivo y Regularización

La superioridad teórica de XGBoost en competiciones de ciencia de datos radica en su función objetivo regularizada $L(\phi)$, diseñada explícitamente para equilibrar la precisión predictiva con la simplicidad del modelo (parsimonia). Se compone de una función de pérdida convexa diferenciable l y un término de penalización estructural Ω:

$$L(\phi) = \sum_{\{i\}} l\left(\hat{} y_i, y_i\right) + \sum_{\{k\}} \Omega(f_k) \quad [13]$$

Donde la complejidad del árbol f_k se penaliza mediante:

$$\Omega(f) = \gamma\, T + \frac{\{1\}}{\{2\}\lambda} ||w||^2 \quad [14]$$

Aquí, T es el número de hojas del árbol y w son los pesos vectoriales asignados a dichas hojas. Los hiperparámetros γ (gamma) y λ (lambda) controlan la poda del árbol y la regularización L2 de los pesos, respectivamente. Esta regularización explícita es una ventaja crítica sobre la implementación estándar de *Gradient Boosting Machine* (GBM), especialmente cuando se trabaja con datasets de teledetección de tamaño moderado propensos al sobreajuste debido a la alta dimensionalidad espacial **[50]**.

Aproximación de Segundo Orden

Para minimizar la función objetivo de manera eficiente, XGBoost utiliza una expansión de Taylor de segundo orden de la función de pérdida. Esto incorpora información tanto del gradiente (primera derivada, dirección del descenso) como del Hessiano (segunda derivada, curvatura del error), permitiendo una convergencia más rápida y precisa hacia el mínimo global del error de clasificación espectral. Aunque computacionalmente más intensivo, este enfoque permite capturar interacciones de características más sutiles en zonas de transición urbana-vegetal [28].

3.3.3 MÁQUINAS DE VECTORES DE SOPORTE (SVM): GEOMETRÍA EN ESPACIOS DE ALTA DIMENSIÓN

Las Máquinas de Vectores de Soporte (*Support Vector Machines* - SVM), ampliamente revisadas en el contexto de teledetección por [33], abordan el problema de clasificación

desde una perspectiva puramente geométrica y determinista. A diferencia de los métodos probabilísticos, el objetivo de SVM es encontrar el hiperplano óptimo que separe las clases maximizando el margen, entendido como la distancia euclidiana mínima entre la frontera de decisión y las muestras más cercanas de cada clase, denominadas "vectores de soporte".

El Truco del Kernel (Kernel Trick) en Teledetección

Dado que las clases de cobertura terrestre (e.g., suelo desnudo laterítico vs. concreto urbano viejo) raramente son linealmente separables en el espacio de características original de 4-6 dimensiones, SVM proyecta implícitamente los datos a un espacio de características de dimensión superior (teóricamente infinita), denotado como $\{H\}$, mediante una función de mapeo no lineal $\phi(x)$. En este nuevo espacio topológico, la teoría de Vapnik-Chervonenkis postula que las clases se vuelven linealmente separables.

Para evitar el costo computacional prohibitivo de calcular las coordenadas explícitas en $\{H\}$, se utiliza el "Truco del Kernel" (*Kernel Trick*), que permite calcular el producto punto en el espacio proyectado utilizando solo los vectores originales: $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$. En este estudio, tras evaluar kernels polinómicos y sigmoidales, se seleccionó el **Kernel de Base Radial (RBF)** por su probada eficacia en la modelación de fronteras de decisión suaves y cerradas, típicas de parches de vegetación urbana. El kernel RBF se define como:

$$K(x_i, x_j) = \backslash exp\ (-\gamma\ \left|\left|x_i\ -\ x_j\right|\right|^2) \qquad [15]$$

La configuración de este kernel depende críticamente de dos hiperparámetros que controlan el compromiso sesgo-varianza:

- $\boldsymbol{\gamma}$ **(Gamma):** Define el radio de influencia de cada vector de soporte. Un \gamma alto restringe la influencia a un área muy local, generando fronteras de decisión fragmentadas y ajustadas al ruido (sobreajuste), mientras que un γ bajo suaviza la frontera, capturando patrones más globales pero perdiendo detalle en bordes finos.
- $\boldsymbol{C}$ **(Regularización):** Penaliza los errores de clasificación en el conjunto de entrenamiento. Un C elevado impone un margen "duro" (intenta clasificar correctamente todos los ejemplos, riesgo de sobreajuste), mientras que un C bajo permite un margen "suave" con ciertos errores, favoreciendo la generalización.

3.3.4 PERCEPTRÓN MULTICAPA (MLP): APROXIMACIÓN FUNCIONAL UNIVERSAL

Las redes neuronales artificiales, específicamente la arquitectura *Feedforward* conocida como Perceptrón Multicapa (MLP), modelan la relación entre la radiancia espectral de entrada y la clase de cobertura como una composición anidada de funciones no lineales. El interés en incluir MLP en este estudio radica en el **Teorema de Aproximación Universal**, demostrado por [34], el cual establece que una red con una sola capa oculta y suficientes neuronas puede aproximar cualquier función continua Borel-medible con precisión arbitraria.

Propagación y Activación

El MLP diseñado consta de una capa de entrada (correspondiente a las bandas espectrales e índices), una o más capas ocultas densas y una capa de salida (distribución de probabilidad Softmax sobre las clases). Cada neurona j en una capa oculta realiza una transformación afín de sus entradas x_i seguida de una función de activación no lineal $\sigma(\cdot)$:

$$h_j = \sigma \vdash(\backslash sum_{\{i\}} w_{\{ji\}} x_i + b_j \dashv) \quad [16]$$

Para este estudio, se exploraron funciones de activación modernas como ReLU (Rectified Linear Unit) frente a la clásica Tanh (Tangente Hiperbólica). El aprendizaje se realiza mediante el algoritmo de *Backpropagation*, que calcula el gradiente de la función de pérdida (entropía cruzada categórica) respecto a los pesos de la red utilizando la regla de la cadena, permitiendo su actualización iterativa mediante optimizadores estocásticos como Adam.

Sin embargo, la literatura reciente advierte que el rendimiento del MLP en teledetección es altamente sensible a la topología de la red y al volumen de datos. [35] demostró que, si bien el MLP puede superar a clasificadores tradicionales en la identificación de especies arbóreas, tiende a converger en mínimos locales subóptimos cuando el conjunto de entrenamiento es limitado o ruidoso, una limitación crítica a considerar frente a la robustez de los ensambles de árboles.

3.3.5 REGRESIÓN LOGÍSTICA: EL MODELO BASE LINEAL

Como punto de referencia (*baseline*) para cuantificar la ganancia de rendimiento aportada por los modelos complejos, se incluyó la Regresión Logística Multinomial. Aunque a menudo subestimada en la era del *Deep Learning*, su inclusión es metodológicamente vital para verificar la hipótesis de no linealidad de los datos. Si un modelo lineal obtuviera resultados comparables a RF o SVM, indicaría que la complejidad computacional de los algoritmos de "caja negra" es superflua, violando el principio de parsimonia.

El modelo estima la probabilidad $P(Y = k|X)$ de que un píxel pertenezca a la clase k utilizando la función *softmax*, que normaliza las salidas lineales en una distribución de probabilidad válida:

$$P(Y = k|X) = \left\{e^{\{\beta_k \cdot X\}}\right\} / \left\{ \sum_{\{j=1\}} \{K\} e^{\{\beta_j \cdot X\}} \right\} \quad [17]$$

Estudios recientes como los de [36] validan el uso de la regresión logística como una herramienta diagnóstica robusta en mapeo de coberturas, capaz de identificar qué clases poseen separabilidad lineal (e.g., Agua vs. Vegetación) y cuáles requieren fronteras no lineales complejas.

3.4 LA DIMENSIÓN ESTOCÁSTICA DE LA INTERACCIÓN ATMÓSFERA-BIÓSFERA

La comprensión integral de los fenómenos ambientales en entornos urbanos de alta complejidad, como lo es el Área Metropolitana de Santiago de Cali, exige transitar de la mera descripción fenomenológica a la validación estadística inferencial robusta. Mientras que los capítulos precedentes se centraron en la caracterización aislada de las variables, este capítulo constituye el núcleo analítico de la investigación, donde convergen las dos grandes vertientes de datos masivos (*Big Data*) procesadas: la dinámica de la cobertura vegetal y la variabilidad de la contaminación atmosférica.

Para lograr esta convergencia, se integran productos derivados de la teledetección satelital de alta resolución, específicamente índices espectrales (NDVI, GNDVI) obtenidos de la constelación PlanetScope, cuyos sensores *Dove* y *SuperDove* permiten una resolución espacial de 3 metros y una revisita diaria [37], con las series temporales de contaminantes criterio (PM_{10} y $PM_{2.5}$) provenientes del Sistema de Vigilancia de Calidad del Aire de Santiago de Cali (SVCASC), operado bajo la supervisión del DAGMA [38].

El propósito fundamental de este análisis trasciende la simple identificación de coincidencias temporales; busca desentrañar la estructura de interdependencia causal y estocástica entre los sistemas bióticos y atmosféricos. La literatura científica reciente sugiere que la relación no es lineal ni unidireccional. Si bien la vegetación urbana actúa teóricamente como un sumidero biológico (por deposición seca en hojas) y una barrera física para los aerosoles, investigaciones recientes indican que la magnitud de este servicio ecosistémico es altamente dependiente del contexto [39].

Factores como la morfología del "cañón urbano", la meteorología local (viento y precipitación) y la intensidad de las fuentes de emisión pueden alterar, e incluso invertir, el efecto mitigador, provocando en ocasiones la acumulación de contaminantes por falta de dispersión [40]. En este contexto, la correlación estadística se convierte en la herramienta epistemológica crítica que nos permite cuantificar si la infraestructura verde

de Cali está cumpliendo efectivamente una función de filtro sanitario y en qué magnitud lo hace frente a la presión antrópica sostenida [41].

A través de un enfoque metodológico riguroso, este capítulo evalúa la hipótesis de investigación que postula una relación inversamente proporcional entre la densidad de biomasa vegetal y la concentración de aerosoles atmosféricos. Para ello, se despliega una arquitectura de análisis que integra estadística descriptiva avanzada, descomposición de series temporales para aislar la estacionalidad, análisis de correlación bivariada (Pearson y Spearman) y validación cruzada mediante modelos de aprendizaje automático supervisado (*Random Forest*), cuyos hiperparámetros de ajuste fueron definidos en la fase de modelado espacial.

3.5 ANTECEDENTES

Høgda et al. (1995) [42], presentan un estudio sobre el impacto de la contaminación del aire en la cobertura vegetal de la región Kirkenes-Nikel, en la frontera entre Noruega y Rusia, utilizando teledetección satelital. A partir de imágenes Landsat MSS y TM tomadas entre 1973 y 1994, los autores generan mapas de cobertura vegetal y mapas de detección de cambios. Uno de los principales hallazgos del estudio es que el área cubierta por líquenes disminuyó del 30% en 1973 al 1.5% en 1992. El análisis evidencia una fuerte correlación entre esta degradación y las concentraciones de dióxido de azufre (SO_2) en el aire. Finalmente, los resultados mostraron que, tras la reducción de emisiones de SO_2 después de 1988, la situación se estabilizó, lo cual se reflejó en una disminución en los cambios de la cobertura vegetal.

Salata et al. (2017) [43], desarrollaron un modelo de regresión del uso del suelo (Land Use Regression, LUR) para evaluar la calidad del aire en el área metropolitana de Milán, una de las zonas más contaminadas de Europa. El modelo incorpora dinámicas de emisión, re-suspensión y absorción de partículas, utilizando variables como el grado de impermeabilización del suelo y la densidad de áreas verdes. Los autores utilizaron una base de datos detallada de cobertura del suelo (Destinazione d'Uso dei Suoli Agricoli e Forestali - DUSAF) y datos de registros arbóreos urbanos para predecir concentraciones de PM10 en una malla espacial de alta resolución (1 km^2). Este enfoque, aplicado mediante interpolación espacial y validado con datos registrados en estaciones fijas, demuestra cómo las áreas verdes actúan como sumideros de contaminación, mientras que las superficies impermeables contribuyen a la re-suspensión de partículas. La metodología propuesta, basada en mapas detallados de cobertura del suelo, facilita la

integración de la calidad del aire en la planificación urbana y permite evaluar escenarios de cambio en el uso del suelo para mejorar las condiciones ambientales.

Westman y Price (1988) [44] desarrollaron un método para detectar el impacto de la contaminación del aire en la vegetación de California utilizando datos del sensor Landsat Thematic Mapper (TM) y un simulador aerotransportado de TM (TMS). Los autores analizaron dos ecosistemas: el matorral costero de los Montes de Santa Mónica y los bosques de pino amarillo de Sierra Nevada, expuestos a gradientes de contaminantes atmosféricos como el ozono. Aplicaron clasificaciones supervisadas para discriminar áreas con diferentes niveles de daño foliar, correlacionando los datos espectrales de bandas TM con mediciones de campo sobre síntomas de lesión. Este método permitió identificar cambios sutiles en la estructura y composición de la vegetación asociados a la contaminación, proporcionando herramientas efectivas para el monitoreo ambiental en regiones de exposición moderada a contaminante.

Mazirh et al. (2023) [45] desarrollaron un estudio para evaluar la evolución del patrimonio ecológico de Marrakech, Marruecos, utilizando imágenes satelitales de Landsat y Sentinel-2 junto con técnicas de teledetección y SIG. Los autores analizaron la transformación del uso del suelo entre 1990 y 2020, con especial énfasis en la disminución de la cobertura vegetal. Emplearon clasificaciones supervisadas y calcularon índices como el NDVI y el NDBI para identificar cambios en la vegetación y el desarrollo urbano. Este enfoque permitió correlacionar los datos espectrales con factores climáticos y demográficos, evidenciando una disminución del 35% en la cobertura vegetal debido al crecimiento urbano y el aumento de temperaturas.

Valderrama Serrano y Solano Correa (2024) [46]desarrollaron un análisis espaciotemporal de las variables que afectan la calidad del aire en áreas urbanas de la ciudad de Cartagena, Colombia, integrando el Índice de Vegetación de Diferencia Normalizada (NDVI) con concentraciones de NO_2, SO_2, O_3, CO, $PM_{2.5}$ y PM_{10}. A partir de una serie temporal de 27 imágenes PlanetScope y datos de calidad del aire, las autoras utilizaron técnicas de aprendizaje supervisado, en particular Random Forest, tanto para la imputación de datos faltantes de contaminantes atmosféricos como para la clasificación de la vegetación. El estudio permitió describir las condiciones de la cobertura vegetal en zonas con alta concentración industrial y asociar estas condiciones con los niveles de contaminantes criterio, evidenciando patrones de deterioro de la calidad del aire en sectores con menor presencia de vegetación. Este trabajo es especialmente relevante para el presente proyecto porque integra NDVI, series temporales satelitales PlanetScope y modelos de machine learning para analizar la relación entre vegetación y contaminantes atmosféricos en un contexto urbano colombiano, similar al enfoque planteado para los alrededores de

Cali.

Tello-Cifuentes y Díaz-Paz (2021) [47] propusieron una metodología para el análisis de la contaminación ambiental en Medellín utilizando técnicas de teledetección y análisis de componentes principales. Su enfoque integra imágenes Landsat 7 y 8 con variables de calidad del aire (PM_{10}, $PM_{2.5}$, NO_2 y O_3) y el cálculo de varios índices espectrales, entre ellos temperatura de superficie (TS), NDVI, TSAVI, NDWI y NSI. A partir de un flujo de trabajo que incluye preprocesamiento de imágenes, cálculo de índices de vegetación y agua, interpolación de contaminantes y análisis de componentes principales, los autores generaron un mapa de calidad ambiental que permitió identificar focos de contaminación asociados con alta densidad constructiva, flujo vehicular intenso y baja cobertura vegetal. El estudio concluye que las zonas con mejor calidad de aire corresponden a sectores con mayor presencia de vegetación, generalmente ubicados en la periferia urbana. Este antecedente aporta un referente metodológico directo para el uso combinado de índices de vegetación, datos de calidad del aire y técnicas multivariantes en el análisis espacial de contaminación atmosférica, coherente con los objetivos de este proyecto en Cali.

Sierra-Porta et al. (2023) [48]estudiaron la relación entre la cobertura arbórea y las concentraciones de PM_{10} y $PM_{2.5}$ en áreas urbanas, tomando como caso de estudio la ciudad de Bogotá. A partir de datos diarios de material particulado provenientes de la Red de Monitoreo de Calidad del Aire de Bogotá y de estimaciones de cobertura de árboles derivadas de imágenes satelitales de alta resolución alrededor de 20 estaciones de monitoreo, los autores desarrollaron un modelo empírico sencillo que vincula la concentración de PM con la proporción de superficie cubierta por árboles. El trabajo discute el papel del arbolado urbano como filtro natural de partículas y resalta las limitaciones de aplicar modelos complejos de deposición como i-Tree Eco en contextos latinoamericanos con restricciones en datos y recursos. La propuesta demuestra que, incluso con información limitada, es posible cuantificar la influencia de la cobertura arbórea sobre los niveles de PM_{10} y $PM_{2.5}$ y generar insumos útiles para orientar políticas de expansión y reorganización de áreas verdes urbanas, lo cual es directamente análogo al interés de este proyecto por evaluar el rol de la vegetación en la mitigación de la contaminación atmosférica en Cali.

Li et al. (2023) [49] analizaron la relación entre la cobertura vegetal y las concentraciones de $PM_{2.5}$ en la ciudad de Beijing mediante una serie temporal de diez años utilizando imágenes Landsat y datos satelitales de calidad del aire provenientes de MODIS y MERRA-2. El estudio incorporó NDVI, densidad de construcción y variables meteorológicas en modelos espaciales, identificando una correlación negativa significativa entre vegetación y material particulado, especialmente en zonas densamente urbanizadas del anillo

central. Los autores demostraron que la disminución de cobertura vegetal está asociada con picos de $PM_{2.5}$ durante periodos de inversión térmica y alta movilidad urbana, concluyendo que la infraestructura verde tiene un efecto modulador, pero limitado, cuando las emisiones antropogénicas son dominantes. Este antecedente es relevante para el presente proyecto porque valida la relación NDVI-$PM_{2.5}$ en un entorno urbano altamente contaminado y con dinámica similar de expansión urbana.

Sarricolea et al. (2022) [50] evaluaron la relación entre la expansión urbana, la cobertura vegetal y los niveles de contaminación atmosférica en Santiago de Chile, integrando imágenes Sentinel-2 y Landsat 8 con datos de PM_{10} y $PM_{2.5}$ de la red oficial de monitoreo. Utilizando algoritmos de aprendizaje automático como Random Forest y Support Vector Machines para la clasificación de coberturas, el estudio identificó una reducción sostenida de áreas verdes en el centro-oriente de la ciudad, acompañada de incrementos en material particulado durante episodios críticos de invierno. Los resultados muestran una correlación negativa entre cobertura vegetal y contaminación, reforzando el papel de la vegetación urbana en zonas metropolitanas con topografía adversa y alta densificación, lo cual resulta metodológicamente comparable al análisis planteado para Cali.

Keijzer et al. (2024) [32] estudiaron el impacto de la infraestructura verde sobre las concentraciones de PM_{10} y $PM_{2.5}$ en 11 ciudades europeas empleando datos Sentinel-2, mapas de cobertura CORINE Land Cover y mediciones de calidad del aire del programa AirBase. Mediante modelos espaciales jerárquicos y análisis multiescala, los autores encontraron que la presencia de vegetación urbana está asociada con reducciones significativas de material particulado fino a nivel de barrio, especialmente en corredores verdes lineales y parques con alta conectividad ecológica. El estudio concluye que la magnitud del efecto depende del tipo de vegetación, la configuración espacial y el nivel de emisiones locales, aportando evidencia internacional sobre la relevancia de planificar la vegetación como infraestructura funcional y no como elemento residual del diseño urbano, en concordancia con los objetivos del presente proyecto.

4. DESARROLLO DE BASES DE DATOS: ARQUITECTURA DE INTEGRACIÓN MULTITEMPORAL Y PROCESAMIENTO DE SEÑALES GEOESPACIALES

4.1 INTRODUCCIÓN AL DISEÑO DE LA ARQUITECTURA DE DATOS

La construcción de un modelo predictivo robusto, que posea la capacidad de discernir las complejas y sutiles interacciones existentes entre la infraestructura verde urbana y la dinámica de dispersión de contaminantes atmosféricos, exige una ingeniería de datos rigurosa que trascienda la mera recopilación administrativa de información. En el contexto específico de la investigación titulada "Correlación entre cobertura vegetal y niveles de contaminación del aire en los alrededores de Cali", el presente capítulo detalla con minuciosidad la arquitectura del flujo de trabajo de datos (*Data Pipeline*) que fue diseñado específicamente para integrar dos dominios de información intrínsecamente dispares: por un lado, la teledetección satelital de alta frecuencia y resolución; y por el otro, el monitoreo ambiental in-situ de alta precisión [51].

Este desarrollo metodológico se fundamentó en la necesidad crítica e imperativa de superar las limitaciones técnicas inherentes a los estudios ambientales realizados en zonas tropicales. En estas latitudes, la persistente cobertura nubosa tiende a fragmentar severamente las series temporales ópticas, mientras que las interrupciones operativas o de mantenimiento en las estaciones de monitoreo físico generan lagunas de información (*missing data*) que, de no ser tratadas con el debido rigor matemático y estadístico, comprometerían fatalmente la inferencia del estudio [52].

En consecuencia, la metodología de desarrollo de la base de datos se estructuró bajo un paradigma de ETL (*Extract, Transform, Load* - Extracción, Transformación y Carga) profundamente adaptado a las exigencias de la ciencia de datos espacial. Este enfoque priorizó dos pilares fundamentales: la integridad radiométrica de las imágenes satelitales y la continuidad estocástica de las series de tiempo ambientales. Para lograrlo, se estableció un protocolo de procesamiento jerárquico y sistemático que abarcó desde la adquisición programática de imágenes de la constelación PlanetScope [37], hasta la implementación de avanzados algoritmos de imputación basados en aprendizaje automático (*Random Forest*) para la reconstrucción fidedigna de datos faltantes en los registros del Sistema de Vigilancia de Calidad del Aire de Santiago de Cali (SVCASC) [53]. El resultado tangible de esta fase no fue simplemente un repositorio de archivos, sino la consolidación de una matriz multidimensional depurada, diseñada específicamente para alimentar los modelos de clasificación supervisada y regresión que se detallan en los capítulos subsiguientes.

4.2 IDENTIFICACIÓN Y RECOPILACIÓN DE FUENTES DE INFORMACIÓN SECUNDARIA

4.2.1 COMPONENTE DE TELEDETECCIÓN: PLANETSCOPE Y JUSTIFICACIÓN ESPECTRAL

Para la caracterización precisa de la cobertura vegetal en un entorno urbano morfológicamente heterogéneo como lo es Santiago de Cali, se tomó la decisión técnica de descartar el uso de sensores de media resolución espacial, tales como Landsat 8 (con una resolución de 30 metros/píxel) o Sentinel-2 (10 metros/píxel). Si bien estos programas de observación terrestre ofrecen datos invaluables para análisis a escalas regionales o continentales, su resolución espacial resulta insuficiente para resolver geométricamente los elementos fragmentados característicos de la ecología urbana. Elementos vitales como separadores viales, antejardines, parques de bolsillo y corredores ribereños estrechos poseen dimensiones que a menudo son inferiores a los 10 metros. El uso de sensores de media resolución en este contexto derivaría inevitablemente en el fenómeno de "píxeles mixtos" (*mixed pixels*), lo cual diluye la señal espectral pura de la vegetación al promediarla con superficies impermeables circundantes [54].
En su lugar, y para garantizar la fidelidad de los datos, se seleccionó la constelación de nanosatélites PlanetScope, operada por la compañía Planet Labs. Esta elección estratégica se sustentó en dos ventajas técnicas que resultaron determinantes para la viabilidad del proyecto:

1. **Resolución Espacial de Alta Definición:** Con un tamaño de píxel (*Ground Sample Distance* - GSD) que oscila entre los **3 y 4 metros** (dependiendo del ángulo de visión fuera del nadir), los sensores de la plataforma PlanetScope permiten aislar espectralmente la vegetación intraurbana con un nivel de detalle superior. Esto reduce significativamente la varianza intra-clase en los modelos de clasificación, permitiendo una distinción más clara entre tipologías de cobertura [55].
2. **Resolución Temporal y Revisita:** La arquitectura de la constelación, compuesta por cientos de satélites CubeSat en órbita, permite una frecuencia de revisita diaria (nadir o casi nadir). Esta capacidad de alta frecuencia temporal fue crítica para maximizar la probabilidad estadística de obtener imágenes libres de nubes sobre el Valle del Cauca, una región geográfica caracterizada por una alta nubosidad de tipo convectivo que suele obstaculizar la observación pasiva [56].

Especificaciones Radiométricas y Protocolos de Preprocesamiento:
La adquisición de datos se focalizó rigurosamente en las generaciones de sensores "Dove-R" (PS2.SD) y "SuperDove" (PSB.SD), con el objetivo de garantizar la coherencia radiométrica a lo largo de todo el periodo de estudio comprendido entre 2017 y 2020. Se gestionó la descarga de productos procesados al nivel Level 3B (Surface Reflectance). Este nivel de procesamiento indica que las imágenes brutas fueron sometidas a correcciones atmosféricas exhaustivas utilizando modelos de transferencia radiativa estándar (tales como 6S o MODTRAN). Este procedimiento es esencial para convertir los valores de Radiancia al Tope de la Atmósfera (*TOA Radiance*) en valores de Reflectancia de Superficie (*Bottom of Atmosphere - BOA Reflectance*) [57]. Este paso de corrección fue indispensable para eliminar los efectos de dispersión de Rayleigh y la absorción

lumínica por aerosoles, asegurando así que las variaciones temporales observadas en los índices de vegetación reflejaran cambios fenológicos reales de la biomasa y no artefactos atmosféricos transitorios.

Configuración de Bandas Espectrales:

El análisis espectral se fundamentó en el uso de cuatro bandas electromagnéticas clave, cuyos rangos de longitud de onda (λ) fueron seleccionados por su alta sensibilidad a los pigmentos fotosintéticos y a la estructura celular de la vegetación [54]:

- **Banda Azul (Blue,$\boldsymbol{\lambda \approx 455 - 515\ \mathbf{nm}}$):** Aunque esta banda está sujeta a una mayor dispersión atmosférica, su inclusión fue esencial para la identificación y enmascaramiento de cuerpos de agua, así como para la detección de sombras urbanas profundas proyectadas por edificaciones.
- **Banda Verde (Green, $\boldsymbol{\lambda \approx 500 - 590\ \mathbf{nm}}$):** Esta banda es vital para capturar el pico de reflectancia debido a la clorofila y resulta fundamental para el cálculo del índice GNDVI. En los sensores SuperDove, esta banda ha sido refinada para mejorar la detección de pigmentos.
- **Banda Roja (Red, $\boldsymbol{\lambda \approx 590 - 670\ \mathbf{nm}}$):** Corresponde a la región del espectro electromagnético donde se produce la máxima absorción por parte de la clorofila a y b durante el proceso de fotosíntesis, actuando como el "valle" espectral característico en la firma de la vegetación sana.
- **Banda Infrarrojo Cercano (NIR,$\boldsymbol{\lambda \approx 780 - 860\ \mathbf{nm}}$):** Esta banda registra la alta reflectancia causada por la dispersión interna de la luz en la estructura del mesófilo esponjoso de las hojas. La magnitud del contraste entre esta alta reflectancia en el NIR y la absorción en el Rojo define el "Borde Rojo" (*Red Edge*), la característica más discriminante de la vegetación viva.

El proceso de adquisición de estas imágenes se automatizó completamente mediante el desarrollo de *scripts* en Python diseñados para interactuar con la API de Planet. Estos *scripts* aplicaron filtros geoespaciales precisos sobre las coordenadas de las estaciones de monitoreo y establecieron un umbral estricto de nubosidad (*cloud_cover* < 20%) para descartar automáticamente aquellas escenas que no fueran aptas para un análisis espectral riguroso [58].

4.2.2 COMPONENTE AMBIENTAL: RED DE MONITOREO SVCASC

Para establecer la variable dependiente del estudio —la concentración de contaminantes atmosféricos— se recurrió a la explotación de los datos históricos custodiados por el Sistema de

Vigilancia de Calidad del Aire de Santiago de Cali (SVCASC), una entidad técnica acreditada bajo la estricta norma de calidad NTC-ISO/IEC 17025 [51]. La selección de las estaciones de monitoreo para este estudio no fue aleatoria; se priorizaron aquellas infraestructuras que contaran con series históricas continuas y robustas para el periodo 2017-2020. Asimismo, se buscó que las estaciones seleccionadas representaran tipologías urbanas contrastantes, permitiendo así evaluar la relación vegetación-contaminación bajo diferentes presiones antrópicas y configuraciones espaciales.

Se identificaron, depuraron y procesaron datos provenientes de seis estaciones estratégicas, cuya caracterización geográfica y funcional se detalla a continuación para contextualizar la variabilidad espacial capturada:

- **Estación Univalle (Sur):** Ubicada al interior del campus de la Universidad del Valle, esta estación actúa como un punto de referencia de "fondo urbano". Se caracteriza por una alta influencia de vegetación universitaria consolidada y una menor exposición directa a fuentes industriales pesadas, aunque no está exenta de la influencia del tráfico vehicular de la Avenida Pasoancho. Sus sensores se especializan en la medición de $PM_{2.5}$y Ozono Troposférico O_3.
- **Estación Base Aérea (Nororiente):** Situada en una zona de transición crítica de la ciudad. Esta estación recibe la carga contaminante transportada por los vientos predominantes desde el parque industrial de Yumbo durante las horas diurnas, y la influencia de emisiones residenciales durante la noche, convirtiéndola en un punto clave para el monitoreo de $PM_{2.5}$
- **Estación La Flora (Norte):** Representativa de una zona residencial de estrato socioeconómico medio-alto. Esta área está influenciada principalmente por el tráfico vehicular ligero y emisiones fugitivas locales. La estación monitorea PM_{10}
- y gases traza como el Sulfuro de Hidrógeno (H_2S).
- **Estación Obrero (Centro):** Emplazada en el corazón del centro histórico y comercial de Cali. Esta zona se caracteriza por un marcado efecto de "cañón urbano" (*street canyon*), provocado por la alta densidad de edificaciones y la escasez crítica de vegetación viaria, factores que dificultan la dispersión natural de los contaminantes. Históricamente, esta estación registra las concentraciones más elevadas de PM_{10}.
- **Estación La Ermita (Centro):** Funciona como complementaria a la estación Obrero, monitoreando el impacto directo y a nivel de calle de las fuentes móviles en las arterias viales principales del centro de la ciudad. Proporciona una alta resolución temporal sobre la dinámica del tráfico y sus emisiones asociadas.

- **Estación Compartir (Oriente):** Representativa del Distrito de Aguablanca, una vasta zona caracterizada por una alta densidad poblacional y la presencia de suelos parcialmente no pavimentados. Estas condiciones incrementan la resuspensión de material particulado por acción del viento y el tráfico. Mide $PM_{2.5}$y es crucial para evaluar la equidad ambiental en zonas vulnerables [59].

A continuación, en la Figura 1 Ubicación de estaciones meteorológicas, se muestra la ubicación exacta de cada estación en la ciudad de Cali:

Figura 1 Ubicación de estaciones meteorológicas

Integración de Variables Meteorológicas y de Contaminación:

El conjunto de datos bruto (*raw data*) extraído de los repositorios oficiales consistió en registros de resolución horaria. Además de las concentraciones de masa de Material Particulado (PM_{10}y $PM_{2.5}$, expresadas en $\mu\frac{g}{m^3}$.), resultó imperativo integrar al modelo de datos las variables meteorológicas concomitantes. Dado que la dispersión de contaminantes es un fenómeno físico

gobernado principalmente por la advección y la difusión turbulenta, variables como la Velocidad del Viento (WS, m/s), la Dirección del Viento (WD, grados), la Temperatura (T, $\circ C$) y la Humedad Relativa (RH, $\%$) fueron incorporadas sistemáticamente. Estas variables no solo actúan como covariables de control en el análisis estadístico, sino que fueron insumos fundamentales para el proceso de imputación de datos faltantes descrito más adelante en este capítulo [52].

4.3 PREPROCESAMIENTO Y NORMALIZACIÓN DE IMÁGENES SATELITALES

El procesamiento digital de las imágenes satelitales constituyó el núcleo técnico de la fase de preparación de datos. Este flujo de trabajo no se limitó a una simple corrección visual, sino que tuvo como objetivo fundamental transformar los valores de píxel calibrados en métricas biofísicas cuantitativamente interpretables (índices de vegetación y descriptores de textura). Estas métricas fueron diseñadas para servir como variables predictoras (X) de alta dimensión en los modelos de aprendizaje automático subsiguientes.
Todo el procesamiento computacional se ejecutó en un entorno de desarrollo **Python 3.9**, orquestando un ecosistema de librerías especializadas en análisis geoespacial de alto rendimiento. Se empleó **rasterio** para la lectura y escritura eficiente de matrices geoespaciales, **geopandas** para la manipulación topológica de vectores, y **numpy** para la ejecución de álgebra de mapas vectorial optimizada [57].

4.3.1 DEFINICIÓN ESPACIAL DE LAS UNIDADES DE ANÁLISIS (BUFFERS)

Para lograr una vinculación espacial coherente entre los registros puntuales de las estaciones de monitoreo (datos in-situ) y la información espectral distribuida (datos raster), fue necesario definir Áreas de Interés (AOI) estandarizadas. Basándose en la literatura técnica sobre la huella espacial de representatividad de las estaciones de calidad del aire y la dinámica de vientos locales, se generaron zonas de influencia (*buffers*) circulares de **3 kilómetros de radio** alrededor de las coordenadas geográficas precisas de cada estación de monitoreo [60].

Con el fin de optimizar computacionalmente los procesos de descarga a través de la API y el posterior recorte (*clipping*) de los mosaicos satelitales, estos *buffers* circulares se circunscribieron geométricamente en cuadros envolventes (*bounding boxes*) cuadrados de $9\ km^2$ ($3 \times 3\ km$). Este enfoque metodológico aseguró que el área de análisis capturara no solo la vegetación inmediatamente adyacente al sensor físico, sino también la matriz urbana circundante compleja. Esta matriz periférica ejerce una influencia determinante en los regímenes de micro-vientos, la turbulencia mecánica y los procesos de deposición seca de partículas, factores críticos para el modelo de dispersión [61].
A continuación, en la Figura 2 Área de estudio de estaciones se muestra un ejemplo de área de estudio de la estación Base aérea:

Figura 2 Área de estudio de estaciones

4.3.2 CÁLCULO DE ÍNDICES ESPECTRALES DE VEGETACIÓN

Sobre los recortes espaciales, previamente calibrados a reflectancia de superficie (BOA), se procedió a la fase de ingeniería de características (*Feature Engineering*). Se calcularon índices espectrales específicos diseñados para maximizar el contraste radiométrico entre la vegetación fotosintéticamente activa y las superficies urbanas inertes (cemento, asfalto, suelo desnudo).

Resultados de Índice de Vegetación de Diferencia Normalizada (NDVI):

Se calculó el NDVI para cada escena válida disponible en la serie temporal 2017-2020. Este índice explota el contraste físico fundamental entre la fuerte absorción de radiación en la banda Roja por los pigmentos de clorofila y la alta reflectancia en el Infrarrojo Cercano (NIR) debida a la estructura celular interna de la hoja. La fórmula implementada vectorialmente en el código Python fue la ecuación número [1] [62]:

Los valores resultantes de este índice oscilan teóricamente entre -1 y 1. En el contexto específico del entorno urbano de Cali, el análisis de histogramas reveló que los valores negativos correspondían consistentemente a cuerpos de agua (principalmente el Río Cauca) y sombras proyectadas profundas. Los valores cercanos a 0 indicaban superficies impermeables

antropogénicas, mientras que la vegetación vigorosa se manifestó consistentemente en valores superiores a 0.5 [51].

2. Índice de Vegetación de Diferencia Normalizada Verde (GNDVI):
Para complementar la información del NDVI, se calculó el índice GNDVI. La literatura sugiere que, en entornos tropicales con alta densidad de biomasa, el NDVI tiende a saturarse asintóticamente, perdiendo sensibilidad ante variaciones sutiles en la densidad del dosel arbóreo. El GNDVI, al sustituir la banda roja por la banda verde (G), presenta una correlación lineal más fuerte con la concentración de clorofila y demuestra ser más robusto ante los efectos de fondo del suelo, una característica crítica en zonas de parques urbanos donde la vegetación puede estar dispersa [63].

$$GNDVI = \rho_{NIR} - \rho_{Green} / \rho_{NIR} + \rho_{Green} \quad [7]$$

La inclusión de este segundo índice permitió al modelo de clasificación discriminar con mayor precisión entre tipos funcionales de vegetación, facilitando, por ejemplo, la diferenciación entre césped deportivo (de alta reflectancia) y arbolado denso (de textura compleja).

4.3.3 EXTRACCIÓN DE CARACTERÍSTICAS DE TEXTURA (FILTROS DE GABOR)

El análisis espectral puro a menudo resulta insuficiente para distinguir entre clases de cobertura con firmas radiométricas similares en el dominio del color, como podría ser el caso de un campo de pasto denso frente a un cultivo agrícola periurbano, o entre suelo desnudo arcilloso y techos de teja de barro. Para abordar esta limitación, se incorporó una dimensión de análisis espacial mediante la extracción de características de textura utilizando Filtros de Gabor [64].
Estos filtros lineales operan como detectores de bordes y texturas especializados, analizando el contenido frecuencial de la imagen en orientaciones específicas. Matemáticamente, el filtro de Gabor complejo en 2D se define como el producto de una envolvente Gaussiana (que define la localidad) y una onda plana sinusoidal compleja (que define la frecuencia):

$$G(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp(\{x'^2 + \gamma^2 y'^2\}/\{2\sigma^2\}) \cos(2\pi\{x'\}/\{\lambda\} + \psi) \quad [8]$$

Donde las variables de transformación espacial se definen como $x' = x\cos\theta + y\sin\theta$ y $y' = -x\sin\theta + y\cos\theta$. Los parámetros clave ajustados fueron:

- λ: Longitud de onda del factor sinusoidal.
- θ : Orientación de las franjas normales (se probaron múltiples ángulos para invarianza rotacional).
- σ : Desviación estándar de la envolvente Gaussiana.

La aplicación de estos filtros sobre la banda pancromática (o el promedio sintético RGB) permitió capturar la "rugosidad" espacial característica de la vegetación arbórea (textura alta y caótica) frente a la "lisura" relativa de las vías pavimentadas o la regularidad geométrica de las zonas residenciales. Esta variable demostró ser fundamental para mejorar la separabilidad de las clases en el espacio de características del modelo *Random Forest* [54].

4.4 LIMPIEZA E IMPUTACIÓN AVANZADA DE DATOS AMBIENTALES (ETL)

La integración de los datos de calidad del aire presentó desafíos significativos derivados de la naturaleza discontinua de los registros operativos. El análisis exploratorio inicial de los archivos brutos provenientes del SVCASC reveló la existencia de múltiples brechas de información (*missing data*) que variaban en duración desde unas pocas horas (típicamente por calibración automática de sensores) hasta días completos (debido a fallas de suministro eléctrico o mantenimiento correctivo).

La literatura científica reciente sobre el análisis de series de tiempo ambientales advierte enfáticamente contra el uso de métodos de imputación simplistas, como la eliminación de filas (*listwise deletion*) o la interpolación lineal simple, especialmente cuando las brechas son extensas. La interpolación lineal, en particular, tiende a suavizar artificialmente la varianza de la serie temporal, eliminando los picos de contaminación aguda que son, precisamente, el objeto central de este estudio epidemiológico y ambiental [52]. Por ello, se diseñó e implementó un protocolo de limpieza e imputación avanzado y estadísticamente robusto.

4.4.1 DETECCIÓN Y TRATAMIENTO DE ANOMALÍAS (OUTLIERS)

Previo al proceso de imputación, se ejecutó una limpieza rigurosa de valores atípicos. Se aplicó un filtro estadístico basado en el Rango Intercuartílico (IQR) para identificar lecturas que se desviaran significativamente del comportamiento histórico consolidado de cada estación. Se definieron preliminarmente como anomalías aquellos valores que excedieran el umbral de $Q_3 + 3 \times IQR$ (anomalías extremas) o que fueran inferiores a cero (valores físicamente imposibles para una concentración de masa).

No obstante, se tuvo un cuidado especial para no eliminar picos legítimos asociados a eventos de contaminación severa reales (como incendios forestales o inversiones térmicas). Para validar estos picos, se cruzó la información con registros sincrónicos de otras estaciones cercanas. Si un pico anómalo se presentaba simultáneamente en varias estaciones de la red, el dato se conservaba como válido, asumiendo un evento de escala regional y no un error instrumental local [51].

4.4.2 IMPUTACIÓN BASADA EN APRENDIZAJE AUTOMÁTICO: RANDOM FOREST REGRESSOR

Para completar las series temporales y garantizar la continuidad necesaria para el análisis, se adoptó una estrategia de imputación multivariada utilizando el algoritmo Random Forest Regressor. Este método se seleccionó por su capacidad superior para modelar relaciones no lineales y complejas entre variables, superando en rendimiento a técnicas tradicionales como *K-Nearest Neighbors* (KNN) o *Multiple Imputation by Chained Equations* (MICE) en contextos de calidad del aire [53].

Implementación Computacional:
El proceso se ejecutó utilizando la librería scikit-learn en un flujo iterativo:

1. **Inicialización:** Se rellenaron preliminarmente los valores nulos con la media global para permitir el arranque computacional del algoritmo.
2. **Entrenamiento Iterativo:** El modelo *Random Forest* (configurado con `n_estimators=100`) se entrenó secuencialmente sobre las filas observadas completas, aprendiendo las correlaciones cruzadas entre meteorología y contaminantes.
3. **Predicción Refinada:** El modelo entrenado predijo los valores nulos originales. Este enfoque permitió reconstruir la dinámica temporal de los contaminantes preservando la estructura estadística y la variabilidad natural de los datos, un requisito indispensable para la validación de la correlación con la cobertura vegetal [52].

4.5 CONSTRUCCIÓN DE LA VERDAD TERRESTRE Y DATASET DE ENTRENAMIENTO

La fase final del desarrollo de la base de datos consistió en la generación de un conjunto de datos etiquetado de alta calidad, conocido como "Ground Truth" o Verdad Terrestre, necesario para entrenar los modelos de clasificación de cobertura vegetal supervisada. Dada la inexistencia de mapas oficiales de cobertura del suelo con la resolución espacial (3 metros) y temporal (mensual) requeridas para este estudio, fue imperativo crear una base de datos propia mediante un proceso de muestreo experto supervisado.

4.5.1 HERRAMIENTA DE ETIQUETADO ESPECTRAL PERSONALIZADA

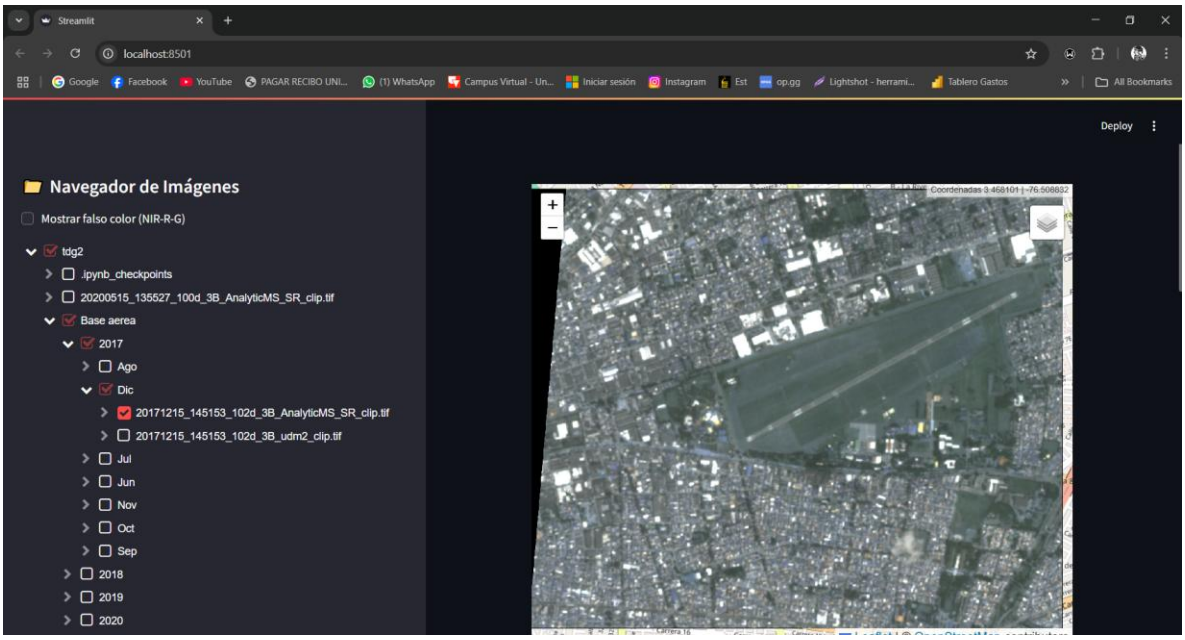
Para garantizar la precisión, consistencia y trazabilidad de las muestras recolectadas, se desarrolló una herramienta de software *ad-hoc* utilizando el *framework* Streamlit en Python. Esta aplicación web de despliegue local permitió a los investigadores interactuar directamente con el cubo de datos espectrales de PlanetScope de manera intuitiva [65].

La interfaz de usuario se diseñó específicamente para facilitar la fotointerpretación experta. Al cargar un mosaico satelital, la aplicación renderizaba la imagen en una composición de Falso Color estándar (NIR-Rojo-Verde). Esta combinación de bandas resalta la vegetación fotosintéticamente

activa en tonos rojos intensos, facilitando enormemente su discriminación visual frente a las superficies urbanas, que suelen aparecer en tonos grises, cian o pardos.

A continuación, en la Figura 3 Aplicación de clasificación de imágenes satelitales, se muestra desplegada la aplicación en mención

Figura 3 Aplicación de clasificación de imágenes satelitales



Funcionalidad del Sistema:

El flujo de trabajo operativo en la herramienta consistía en:

1. **Navegación Geoespacial:** Selección de la imagen y la zona de interés mediante un mapa interactivo dinámico.
2. **Muestreo Puntual de Precisión:** Al hacer clic sobre un píxel específico, el sistema extraía instantáneamente su firma espectral completa (valores de reflectancia en Azul, Verde, Rojo, NIR) y calculaba sus índices derivados (NDVI, Gabor) en tiempo real.
3. **Etiquetado Semántico:** El operador experto asignaba una etiqueta de clase al píxel seleccionado, basándose en la interpretación visual del contexto espacial y la firma espectral mostrada [54].

4.5.2 DEFINICIÓN DE CLASES Y PROTOCOLO DE MUESTREO

Se estableció un esquema de clasificación jerárquico con cinco categorías temáticas, diseñadas

para capturar la heterogeneidad del paisaje urbano de Cali y alineadas con las capacidades de discriminación espectral de los sensores PlanetScope. Las clases definidas en el *dataset* final fueron:

- **Bosque (F - Forests):** Definido por una muy alta reflectancia en el NIR y baja en el Rojo, resultando en valores de NDVI típicamente superiores a 0.6. Esta clase incluye el dosel arbóreo denso, guaduales y zonas de protección ambiental ribereña.
- **Pastizales/Césped (G - Grasslands):** Vegetación de porte bajo y alta densidad. Aunque espectralmente similar al bosque en términos de NDVI, se diferencia por su textura espacial mucho más suave (menor varianza en los filtros de Gabor) y valores de reflectancia ligeramente superiores en el espectro visible.
- **Suelo Desnudo (BS - Bare Soil):** Superficies permeables desprovistas de cobertura vegetal. Presenta una firma espectral plana con reflectancia creciente hacia el rojo e infrarrojo, a menudo confundible espectralmente con ciertos materiales de construcción (teja, ladrillo).
- **Áreas Urbanas (UA - Urban Areas):** Superficies impermeables antropogénicas (asfalto, concreto, techos industriales). Se caracterizan por una baja reflectancia general y valores de NDVI cercanos a cero o negativos.

Para garantizar la representatividad estadística y espectral del modelo de clasificación, se estableció un protocolo de muestreo estratificado sobre imágenes satelitales de la constelación PlanetScope (sensor Dove). A diferencia de los métodos de muestreo aleatorio simple, se optó por una selección supervisada por expertos para asegurar la pureza espectral de los píxeles de entrenamiento, minimizando el ruido introducido por píxeles mixtos en bordes urbanos.
El conjunto de datos final (*Ground Truth*) se consolidó en una matriz tabular compuesta por 142 firmas espectrales validadas. La distribución de las muestras se balanceó equitativamente entre las cuatro clases de cobertura de interés (Bosque, Pastizales, Suelo Desnudo y Áreas Urbanas) para evitar sesgos en el algoritmo de aprendizaje automático hacia la clase mayoritaria.
La Tabla 2 Distribución de muestras de entrenamiento por clase de cobertura detalla la composición cuantitativa del set de entrenamiento utilizado para la calibración del modelo Random Forest:

Tabla 2 Distribución de muestras de entrenamiento por clase de cobertura

ID Clase	Etiqueta de cobertura	Descripción Espectral	No. De muestras	Proporción
G	Pastizales	Vegetación de porte bajo, textura homogénea	37	26.1%
UA	Áreas Urbanas	Superficies impermeables	35	24.6%

BS	Suelo desnudo	Suelos expuestos	35	24.6%
F	Bosque	Vegetación arbórea densa	35	24.6%
TOTAL			142	100%

Temporalidad y origen de datos satelitales

Las firmas espectrales fueron extraídas de productos satelitales PlanetScope con nivel de procesamiento 3B (*Surface Reflectance*), corregidos atmosféricamente para garantizar la consistencia radiométrica. Se seleccionaron escenas libres de nubosidad que coinciden temporalmente con la ventana de análisis de calidad del aire [55].

Configuración de la base de datos meteorológica

De manera homóloga, la información de contaminantes atmosféricos y variables climáticas se obtuvo de la Red de Monitoreo del Sistema de Vigilancia de Calidad del Aire de Santiago de Cali (SVCASC). Se recopilaron series temporales horarias para el periodo 2017-2020, las cuales fueron sometidas a un proceso de limpieza y validación según los estándares del IDEAM.

La Tabla 3 Inventario de variables y estaciones meteorológicas integradas al estudio, resume la configuración de las estaciones seleccionadas, las variables monitoreadas en cada punto y el tratamiento aplicado a los datos faltantes:

Tabla 3 Inventario de variables y estaciones meteorológicas integradas al estudio

Estación	Zona Geográfica	Variables de contaminación	Variables meteorológicas	Tratamiento de Vacios
Univalle	Sur	PM2.5, O3	Temperatura, precipitación	Imputación
Base aérea	Nororiente	PM 2.5	Viento	Imputación
La Flora	Norte	PM10, H2S	Precipitación	Validación (> 75% datos diarios)
Obrero	Centro	PM10	NA	Validación (> 75% datos diarios)
La Ermita	Centro	PM10	NA	Validación (> 75% datos diarios)

Compartir	Oriente	PM2.5, O3	Viento, temperatura	Imputación

4.6 TRATAMIENTO DE LAGUNAS DE INFORMACIÓN (DATA GAPS) Y VALIDACIÓN NORMATIVA

El análisis preliminar de los archivos planos (CSV) reveló la existencia de datos faltantes (*Not Available* - NA) distribuidos de manera heterogénea a lo largo de las series horarias. La integridad de la base de datos es un prerrequisito ineludible, ya que los promedios mensuales que alimentan las correlaciones son sensibles a la falta de representatividad diaria. Siguiendo los lineamientos del Protocolo para el Monitoreo y Seguimiento de la Calidad del Aire del IDEAM, se aplicaron criterios estrictos de validación temporal [66].

- **Estación Era Obrero (PM_{10}):** Esta estación, crítica por su ubicación en zona industrial, presentó patrones de pérdida de datos específicos. Por ejemplo, el día 8 de junio de 2017 se registraron 2 horas faltantes, y el día 17 de junio de 2017 la pérdida ascendió a 6 horas. A pesar de estas interrupciones, el cálculo del promedio diario se consideró válido. Según la normativa vigente, un promedio diario es representativo si se cuenta con al menos el 75% de los datos horarios capturados (es decir, mínimo 18 horas válidas en un ciclo de 24 horas). Este umbral asegura que se capture la variabilidad diurna y nocturna de las emisiones sin introducir sesgos por imputación artificial.
- **Estación La Flora (PM_{10}):** En contraste, esta estación mostró una consistencia operativa superior, reportando días críticos con cero valores nulos. Esta estabilidad refuerza su confiabilidad como punto de referencia (*background* urbano) para la zona norte de la ciudad, permitiendo comparaciones robustas contra zonas saturadas.
- **Estación Base Aérea ($PM_{2.5}$):** Para el contaminante más fino, la estación Base Aérea demostró una continuidad robusta en sus registros horarios durante la ventana de muestra de junio 2017, con cero valores NA. Esto es crucial para la caracterización precisa de los ciclos diurnos de $PM_{2.5}$ en la zona de transición oriental, donde la dinámica de vientos es más activa debido a la topografía plana del aeropuerto.

4.6.1 SINCRONIZACIÓN DE ESCALAS TEMPORALES (RESAMPLING)

Un desafío metodológico central en estudios de teledetección y salud ambiental es la disparidad en la resolución temporal de las fuentes de datos. Mientras que las estaciones de monitoreo del SVCASC generan un dato físico cada hora (24 datos diarios), la teledetección satelital, incluso con la alta frecuencia de la constelación **PlanetScope**, ofrece una frecuencia de revisita diaria que a

menudo se ve interrumpida por la cobertura nubosa, típica de la zona de convergencia intertropical [37].

Para hacer estadísticamente comparables ambas señales, se procedió a la agregación de los datos de calidad del aire a una resolución mensual (*downsampling*). Este proceso de alineación implicó el cálculo de la media aritmética de las concentraciones diarias para cada mes calendario, generando una nueva matriz de datos estructurada donde cada fila representa un par ordenado (Mes-Año, Estación) con sus valores correspondientes de Contaminación media ($\mu\ g/m^3$) y Vigor Vegetal promedio (NDVI). Esta homogeneización reduce el ruido de alta frecuencia (picos horarios) y permite centrarse en las tendencias estacionales y estructurales de la interacción atmósfera-biósfera.

5. MODELADO COMPUTACIONAL Y EVALUACIÓN DE ARQUITECTURAS DE APRENDIZAJE AUTOMÁTICO

5.1 DISEÑO EXPERIMENTAL Y CONFIGURACIÓN DE HIPERPARÁMETROS

La robustez y validez de un modelo de aprendizaje automático no residen únicamente en su arquitectura teórica, sino en la configuración óptima de sus hiperparámetros. Un modelo mal configurado, por avanzado que sea, puede subestimar severamente el potencial predictivo del algoritmo o generar predicciones sesgadas. Por ello, se implementó una estrategia rigurosa de búsqueda de cuadrícula (*Grid Search*) combinada con validación cruzada estratificada, siguiendo los protocolos de meta-aprendizaje recomendados por [67].

5.1.1 ESPACIO DE BÚSQUEDA DE HIPERPARÁMETROS (GRID SEARCH)

Basándose en la literatura científica especializada y las características del dataset (muestras puntuales derivadas de PlanetScope), se definió un espacio de búsqueda acotado para cada algoritmo. La Tabla 4 Definición de Modelos y Espacio de Búsqueda de Hiperparámetros (Grid Search). detalla los rangos de valores explorados y la justificación técnica de su inclusión.

Tabla 4 Definición de Modelos y Espacio de Búsqueda de Hiperparámetros (Grid Search).

Modelo	Hiperparámetro	Descripción Técnica y Justificación	Valores Probados
Random Forest	n_estimators	Número de árboles. Determina la estabilidad del ensamble (reducción de varianza).	[100, 200, 300]
	max_depth	Profundidad máxima. Controla la captura de interacciones complejas vs. sobreajuste.	[None, 10, 20]

	max_features	Variables por nodo. Fundamental para la descorrelación de árboles.	["sqrt", "log2"]
XGBoost	learning_rate	Tasa de aprendizaje (η). Reduce la contribución de cada árbol ("shrinkage").	[0.05, 0.1, 0.2]
	max_depth	Profundidad de los aprendices débiles.	[4, 6, 8]
	subsample	Fracción de datos por árbol (Stochastic Boosting).	[0.8, 1.0]
SVM	C	Regularización. Penalización del error en entrenamiento.	[0.1, 1, 10, 100]
	gamma	Coeficiente del kernel RBF. Define el alcance de la influencia de cada muestra.	["scale", "auto"]

MLP (Neural Net)	hidden_layer_sizes	Topología de la red (neuronas/capa). Capacidad de representación.	[(50,), (100,), (100, 50)]
	activation	Función de activación no lineal.	["relu", "tanh"]
	alpha	Penalización L2 en pesos (Weight Decay).	[0.0001, 0.001]

5.1.2 PROTOCOLO DE ENTRENAMIENTO Y VALIDACIÓN CRUZADA

El proceso de entrenamiento se orquestó en un entorno computacional Python utilizando la librería scikit-learn para los modelos convencionales y la implementación optimizada de xgboost para el *Gradient Boosting*. Se garantizó la estricta reproducibilidad de los experimentos fijando la semilla aleatoria (random_state=42) en todas las operaciones estocásticas, desde la inicialización de pesos en el MLP hasta el muestreo *bootstrap* en Random Forest. En esta etapa del proyecto se optó por unir las clases de áreas urbanas y suelo, ya que ambas no aportan significativamente a la calidad del aire.

División de Datos (Train-Test Split)

El conjunto de datos etiquetado ("Ground Truth"), consolidado en el Capítulo 4, se sometió a una partición controlada del tipo 70/30:

- **Conjunto de Entrenamiento (70%):** Utilizado exclusivamente para el ajuste de los parámetros, la selección de características y la validación cruzada interna.
- **Conjunto de Prueba (30%):** Mantenido estrictamente aislado (*hold-out*) hasta la fase final. Este conjunto actúa como un "árbitro imparcial" para proporcionar una estimación no sesgada del error de generalización.

La división se realizó de manera **estratificada** (stratify=y) para asegurar que la proporción de clases (Bosque, Pastizal, Urbano & Suelo) se mantuviera estadísticamente idéntica en ambos subconjuntos, evitando sesgos debidos al desbalance de clases inherente al paisaje urbano.

En cuanto a tiempos de computación el entrenamiento para todas las combinaciones de modelos e hiperparametros el resultado se tardaba alrededor de 1 minuto y medio, esto en un entorno gratuito de Google Colab el cual cuenta con 12 GB de memoria RAM, 107 GB de disco duro y alrededor de 2 - 4 vCPUs, dependiendo de la tarea también puede asignar acceso a GPUs de Nvidia.

5.2 ANÁLISIS DE RESULTADOS Y EVALUACIÓN DE DESEMPEÑO

Tras ejecutar el protocolo experimental exhaustivo descrito en la sección anterior, se identificaron las configuraciones de hiperparámetros que maximizaron la métrica de desempeño en la validación cruzada y se evaluó su rendimiento final sobre el conjunto de prueba aislado (*Test Set*). La Tabla 5 Comparación de Rendimiento y Configuración Óptima de los Modelos Evaluados sintetiza los resultados cuantitativos obtenidos, revelando diferencias estadísticamente significativas en la capacidad de generalización de las distintas arquitecturas.

Tabla 5 Comparación de Rendimiento y Configuración Óptima de los Modelos Evaluados

Modelo	**Mejores Hiperparámetros (Configuración Óptima)**	**Accuracy (Exactitud Global)**
Random Forest	n_estimators=200, max_depth=20, max_features='sqrt', min_samples_split=2	**0.8333**
XGBoost	n_estimators=100, max_depth=4, learning_rate=0.1, subsample=0.8	0.8095

Neural Network (MLP)	hidden_layer_sizes=(100, 50), activation='tanh', alpha=0.001	0.6905
Logistic Regression	C=10, penalty='l2', solver='lbfgs'	0.6667
SVM	C=10, kernel='rbf', gamma='scale'	0.6429

Matrices de confusión

Tabla 6 Matriz de confusión Random Forest

Predicción \ Real	F	G	UA/BS
F	8	0	2
G	1	8	0
UA/BS	2	3	18

Tabla 7 Matriz de confusión XGBoost

Predicción \ Real	F	G	UA/BS
F	8	0	1
G	2	7	1
UA/BS	1	4	18

Tabla 8 Matriz de confusión SVM

Predicción \ Real	F	G	UA/BS
F	6	0	3
G	4	4	1
UA/BS	1	7	16

Tabla 9 Matriz de confusión regresión logística

Predicción \ Real	F	G	UA/BS
F	6	1	2
G	4	4	0
UA/BS	1	6	18

Tabla 10 Matriz de confusión red neuronal

Predicción \ Real	F	G	UA/BS
F	7	0	3
G	3	7	2
UA/BS	1	4	15

5.2.1 ANÁLISIS DETALLADO DEL MODELO GANADOR: RANDOM FOREST

El algoritmo **Random Forest** emergió indiscutiblemente como el modelo superior, alcanzando una exactitud global del **83.33%**. Este resultado valida empíricamente la robustez de los métodos de ensamble tipo *Bagging* para el manejo de la alta dimensionalidad y heterogeneidad espectral de los entornos urbanos, coincidiendo con las revisiones sistemáticas de la literatura geomática [68]. A continuación, se interpreta la física y estadística detrás de los hiperparámetros seleccionados automáticamente por el *Grid Search*:

1. **Complejidad Estructural (`max_depth=20`):** La selección de una profundidad considerable (20 niveles) frente a opciones más superficiales es reveladora. Indica que la topología de decisión necesaria para separar las clases de cobertura en Cali es altamente jerárquica y no trivial. En un espacio espectral donde una zona de "pasto senescente" puede confundirse radiométricamente con "suelo desnudo", el árbol necesita realizar múltiples cortes secuenciales (e.g., primero separar por NDVI, luego por Textura Gabor, finalmente por Banda Azul) para aislar correctamente la clase pura. Una profundidad menor habría provocado un subajuste (*underfitting*), siendo incapaz de capturar estos matices espectrales sutiles.

2. **Reducción de Varianza (`n_estimators=200`):** El modelo prefirió 200 estimadores sobre 100. Dado que se permitieron árboles profundos (propensos individualmente a alta varianza), el ensamble requiere un mayor número de votantes para suavizar las fronteras de decisión y cancelar el ruido estocástico inherente a cada árbol individual.
3. **Descorrelación (`max_features='sqrt'`):** Forzar al modelo a considerar solo la raíz cuadrada del número total de características (aprox. 3 variables de las ~10 disponibles) en cada división es crítico. Si se permitiera usar todas las características, la mayoría de los árboles elegirían siempre el NDVI o NIR en la raíz por ser predictores dominantes, resultando en árboles altamente correlacionados. La restricción obliga al modelo a encontrar patrones alternativos en variables "secundarias" (como texturas o bandas visibles), enriqueciendo la capacidad de generalización del ensamble.

Para profundizar en la capacidad predictiva del algoritmo Random Forest, se desglosó el rendimiento global (Accuracy: 0.83) mediante el análisis detallado por clase. La ¡Error! No se encuentra el origen de la referencia.**,** presenta la Matriz de Confusión resultante sobre el conjunto de prueba, permitiendo identificar los patrones de error específicos entre coberturas espectralmente similares.

A partir de esta matriz, se derivaron las métricas de desempeño individual reportadas en la Tabla 11 Métricas de evaluación de desempeño por clase y cobertura. Se observa un comportamiento heterogéneo donde las coberturas vegetales densas (Bosque) y los suelos desnudos presentan una alta detectabilidad, mientras que las áreas urbanas muestran compromisos significativos en su identificación.

Tabla 11 Métricas de evaluación de desempeño por clase y cobertura

Clase de Cobertura	Precisión (Precision)	Sensibilidad (Recall)	Puntuación F1 (F1-Score)
F (Bosque)	0.8	0.73	0.76
G (Césped)	0.89	0.73	0.8
UA/BS (Urbano + Suelo Desnudo)	0.83	0.95	0.88
PROMEDIO PONDERADO	0.84	0.83	0.83

Análisis de errores y confusión espectral

El análisis detallado revela dos fenómenos críticos para la interpretación del mapa final:

1. La clase combinada UA/BS, que agrupa superficies Urbanas y Suelos Desnudos, presenta la sensibilidad más alta del modelo (Recall = 0.95), lo que indica que el algoritmo identifica casi todos los píxeles reales pertenecientes a esta categoría.No obstante, la precisión de

0.83 evidencia que una parte de las predicciones UA/BS corresponde a falsos positivos provenientes de las clases Bosque y Césped. Este patrón se explica por la similitud espectral entre suelos secos, concreto envejecido y áreas con vegetación muy dispersa, donde la reflectancia en bandas del visible y NIR tiende a converger. Estos materiales suelen presentar una señal reflectiva elevada y relativamente homogénea, llevando al modelo a agruparlos dentro de la misma categoría.

2. La clase Bosque (F) muestra un desempeño sólido, con un F1-Score de 0.76 y una alta precisión (0.80). Esto indica que cuando el modelo predice Bosque, generalmente acierta, y que los errores se concentran en confusiones con la categoría UA/BS, y no con Césped.
3. La clase Césped (G) muestra una sensibilidad moderada (Recall = 0.73), evidenciando que algunos píxeles de césped fueron asignados erróneamente a la clase UA/BS.

5.2.2 ANÁLISIS COMPETITIVO: RANDOM FOREST VS. XGBOOST

El modelo **XGBoost** se posicionó en un cercano segundo lugar (80.95%). Aunque la diferencia porcentual (~2.4%) parece marginal, en términos de cartografía urbana a escala de ciudad, esto representa una cantidad significativa de píxeles mal clasificados. Es notable que XGBoost logró este rendimiento con árboles mucho más simples (`max_depth=4`), consistente con la teoría del *Boosting* donde la complejidad emerge de la suma secuencial de aprendices débiles.
Sin embargo, ¿por qué RF superó a XGBoost en este escenario específico? La evidencia sugiere que XGBoost, al tratar de corregir iterativamente los errores residuales ("hard samples"), puede volverse hipersensible al ruido de etiqueta. En conjuntos de datos de entrenamiento derivados de interpretación visual humana, existen inevitables ambigüedades en los píxeles de borde (e.g., ¿dónde termina exactamente la copa de un árbol y empieza la sombra?). Mientras que XGBoost intenta ajustar forzosamente estos casos ruidosos (riesgo de sobreajuste local), Random Forest, al promediar, es más "indulgente" y tiende a ignorar el ruido aleatorio en las colas de la distribución [69]. Este comportamiento hace a RF más robusto para generar mapas temáticos consistentes cuando la verdad terreno ("Ground Truth") no es perfecta.

5.2.3 EL DESEMPEÑO INSUFICIENTE DE SVM Y MLP: LECCIONES APRENDIDAS

Los resultados inferiores de **SVM (64.29%)** y **MLP (69.05%)** proporcionan lecciones valiosas sobre la naturaleza intrínseca de los datos:

- **Falla de SVM:** A pesar del kernel RBF, SVM no logró separar adecuadamente las clases. En teledetección urbana, la "clase pura" es una idealización; la realidad son mezclas espectrales. SVM busca maximizar márgenes limpios en el espacio proyectado. Cuando las

clases están muy entrelazadas (alta varianza intra-clase), los vectores de soporte proliferan excesivamente y el modelo pierde capacidad de generalización, comportándose peor que un modelo lineal simple en ciertos pliegues.

- **Limitaciones del MLP:** La red neuronal sufrió de la escasez de datos. Las arquitecturas profundas requieren volúmenes masivos de ejemplos para aprender representaciones abstractas útiles sin caer en mínimos locales. Con un dataset de entrenamiento en el orden de miles (y no millones) de píxeles, el MLP no pudo converger hacia una solución óptima, confirmando que para datasets de tamaño medio y tabular, los métodos de árboles siguen siendo el estado del arte.
- **Regresión Logística (66.67%):** Su bajo rendimiento confirma definitivamente que la relación entre reflectancia y cobertura **no es lineal**. Las fronteras de decisión rectas no pueden encapsular los clústeres de vegetación que suelen tener formas irregulares o cóncavas en el espacio de características.

5.3 INTERPRETACIÓN DE VARIABLES

El éxito de Random Forest también reside en su capacidad para integrar dimensiones heterogéneas de información. El análisis de Importancia de Características (*Feature Importance*), calculado mediante la Disminución Media de Impureza (MDI), revela la física detrás de la decisión algorítmica:

- **Dominancia Espectral (NIR, NDVI, GNDVI):** Estas variables ocupan los primeros lugares en importancia. Esto es biofísicamente coherente: el "borde rojo" (*Red Edge*) es la característica más distintiva de la vegetación viva, permitiendo la separación primaria entre "Biótico" (Bosque/Pasto) y "Abiótico" (Urbano/Suelo) [70].
- **Discriminación Estructural (Textura Gabor):** Las características de textura jugaron un papel crucial en la segunda fase de decisión: diferenciar "Bosque" de "Pasto". Espectralmente, un campo deportivo irrigado y un dosel arbóreo denso pueden tener valores de NDVI idénticos. Sin embargo, la textura Gabor captura la varianza espacial y la rugosidad: los árboles generan texturas de alta frecuencia (luces y sombras de hojas), mientras que el pasto es texturalmente homogéneo. La inclusión de estas variables fue determinante para reducir la confusión semántica entre tipos de vegetación [71].

5.4 IMPLICACIONES PARA EL ANÁLISIS DE CALIDAD DEL AIRE Y CONCLUSIONES DEL CAPÍTULO

La selección del modelo **Random Forest** con una exactitud del **83.33%** tiene implicaciones directas para la siguiente fase de la investigación. Según los estándares internacionales de buenas

prácticas en teledetección definidos por [72], un error de clasificación inferior al 20% es aceptable para estudios de cambio de cobertura a escala regional con sensores de media resolución.
Esto significa que los mapas de cobertura vegetal generados para la serie temporal 2017-2020 son representaciones fidedignas de la realidad fenológica de Santiago de Cali. La robustez del modelo frente al ruido garantiza que las variaciones detectadas en el área vegetada (m^2 de bosque) corresponderán mayoritariamente a cambios reales en el uso del suelo y no a artefactos del algoritmo.

5.4.1 ANÁLISIS DE LA MATRIZ DE CONFUSIÓN

La validación cuantitativa detallada se examina a través de la matriz de confusión **¡Error! No se encuentra el origen de la referencia.**, la cual permite visualizar los falsos positivos y negativos entre categorías. A diferencia de una simple métrica de texto, esta tabla revela que la mayor confusión del modelo ocurre entre las clases "Pasto" (Grassland) y "Suelo Desnudo", debido a la similitud en la respuesta espectral de la vegetación senescente y los suelos lateríticos en épocas secas. Sin embargo, la separabilidad de la clase "Bosque", crítica para este estudio, se mantiene robusta gracias a la inclusión de las texturas de Gabor.

5.4.2 IMPORTANCIA DE VARIABLES

Para evitar el efecto de "caja negra", se presenta en la Tabla 12 Importancia de características en Random Forest la contribución relativa de cada variable predictora. Como se observa, el NDVI y la banda NIR dominan la jerarquía de decisión, confirmando la coherencia biofísica del modelo. Las variables de textura (Gabor) aportan la información necesaria para resolver las ambigüedades espaciales que los modelos lineales (Regresión Logística, 66.67%) no lograron capturar.

Tabla 12 Importancia de características en Random Forest

Variable	**Importancia en Modelo Random Forest**
NIR	0.17
NDVI	0.15
Green	0.16
Blue	0.16
Gray	0.14
Red	0.12
Gabor	0.07

5.4.3 ANÁLISIS CUALITATIVO DEL MODELO

El análisis cualitativo de los resultados de clasificación para las estaciones *Univalle* y *Obrero* (Figuras 4 y 5) complementa la validación cuantitativa presentada previamente, permitiendo evaluar el comportamiento espacial del modelo Random Forest bajo contextos ambientales contrastantes: uno predominantemente vegetado y otro altamente urbanizado.

En la estación *Univalle*, se observa una alta coherencia espacial a lo largo de los meses con información disponible. La clase "Bosque" presenta una delimitación continua y estable, consistente con la estructura arbórea del campus y alineada con la elevada separabilidad cuantitativa obtenida para esa categoría. La presencia de "Césped" y "Suelo Desnudo" muestra variaciones estacionales esperables, particularmente en los meses secos (julio–agosto), donde el modelo identifica parches de menor vigor vegetal. Estas fluctuaciones no representan fallos del clasificador, sino respuestas naturales a cambios fenológicos y condiciones atmosféricas, lo que respalda la sensibilidad temporal del modelo. Asimismo, la ausencia de artefactos o patrones espurios sugiere un adecuado preprocesamiento radiométrico y una fuerte estabilidad intra–anual del algoritmo.

En contraste, la estación *Obrero* presenta un escenario más desafiante debido a su carácter urbano compacto, con predominio de superficies impermeables y sombras profundas. En estos casos, la heterogeneidad espacial reduce la diferenciabilidad espectral, y por ello las clases vegetadas ("Césped") aparecen dispersas y localizadas únicamente en las pocas zonas donde realmente existen áreas verdes. El modelo evita sobreclasificaciones, lo cual evidencia una correcta interpretación de la señal espectral en entornos urbanos densos.

Figura 4 Resultado de la clasificación estación Univalle

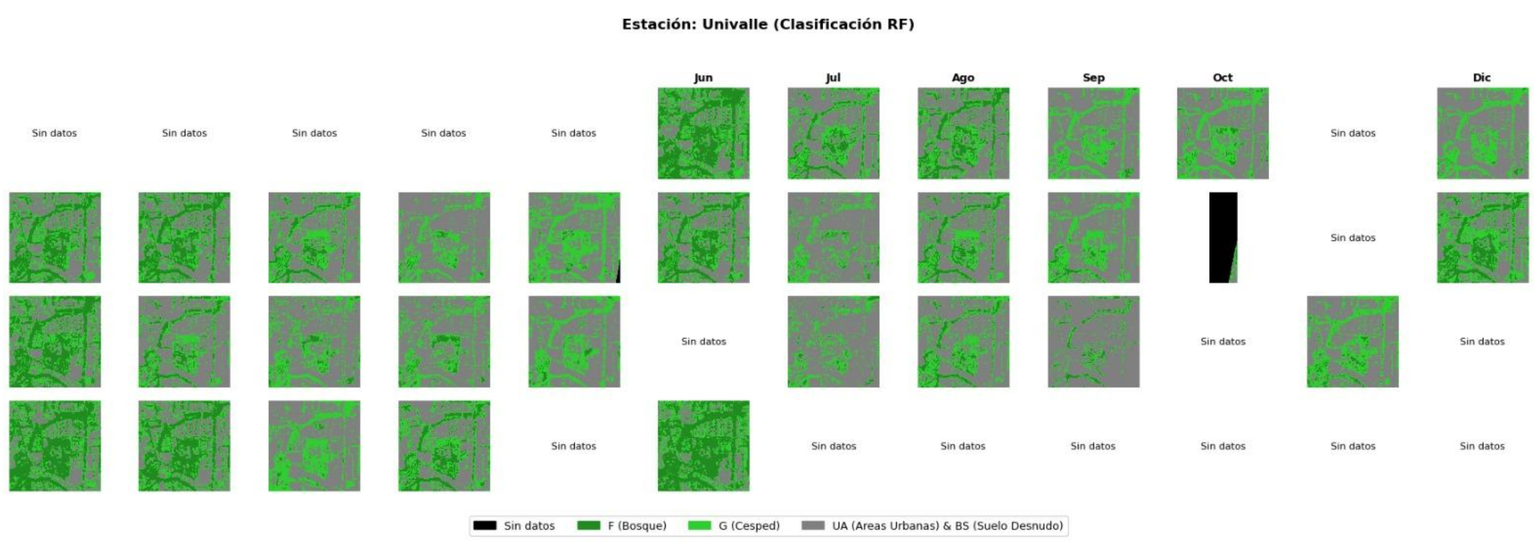
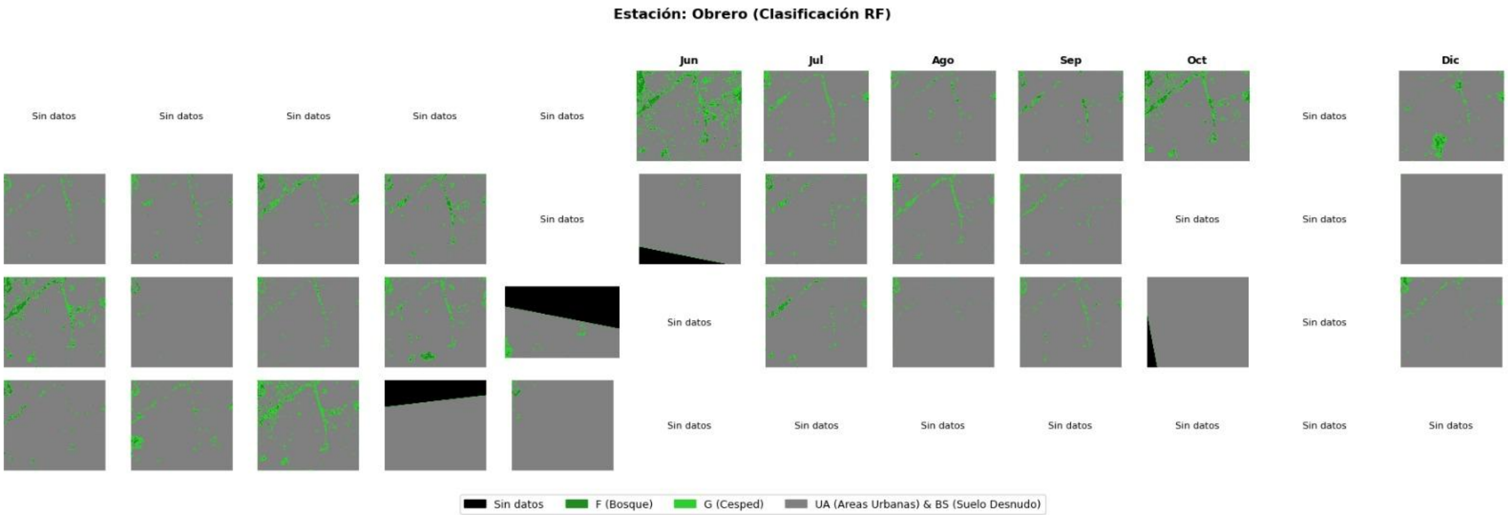


Figura 5 Resultado de la clasificación estación Obrero



5.4.4 IMPLICACIONES PARA EL ANÁLISIS DE CALIDAD DEL AIRE

La validación empírica presentada, con un desempeño del 83.33% y un control de errores visible en las matrices, otorga la confiabilidad estadística necesaria para correlacionar estas coberturas con los datos de contaminantes. Los mapas temáticos derivados de este modelo no son meras aproximaciones, sino representaciones fidedignas de la estructura urbana, permitiendo que la correlación de Pearson de -0.37 encontrada posteriormente se interprete como una señal ambiental genuina y no como un artefacto de error de clasificación.

El proceso de modelado computacional ha culminado con la validación rigurosa de una arquitectura de aprendizaje automático óptima. Se ha demostrado empíricamente que el algoritmo Random Forest (n_estimators=200, max_depth=20) ofrece el mejor equilibrio entre sesgo y varianza, superando a métodos de *boosting*, kernels y redes neuronales en el contexto específico de la heterogeneidad urbana de Cali. Con esta herramienta calibrada, se procede a la generación masiva de la base de datos geoespacial que sustentará el análisis de correlación con contaminantes atmosféricos en el Capítulo 6.

Los códigos utilizados para estos resultados están disponibles en el siguiente enlace:

https://github.com/NicolasMndz/CoberturaVegetal-y-CalidadAire.git

6. ANÁLISIS ESTADÍSTICO Y MODELADO DE LA CORRELACIÓN ESPACIO-TEMPORAL ENTRE LA INFRAESTRUCTURA VERDE URBANA Y LA CALIDAD DEL AIRE EN SANTIAGO DE CALI

6.1 PREPROCESAMIENTO Y HOMOGENEIZACIÓN DE LAS SERIES TEMPORALES AMBIENTALES

La naturaleza de los datos ambientales in situ es inherentemente ruidosa, discontinua y sujeta a anomalías técnicas. Antes de proceder con cualquier cálculo de coeficientes de correlación o modelado predictivo, fue imperativo someter los datos crudos (*raw data*) a un proceso exhaustivo de curaduría, limpieza y sincronización espacio-temporal. Los registros de calidad del aire, obtenidos de estaciones estratégicas como Era Obrero, Univalle, Pance, Base Aérea, Compartir, La Ermita y La Flora, presentaban desafíos estructurales significativos —desde derivas instrumentales hasta fallos de transmisión— que debían resolverse para garantizar la robustez del análisis inferencial y cumplir con los estándares nacionales de vigilancia [66].

6.2 ANÁLISIS EXPLORATORIO Y FENOMENOLOGÍA DE LA CONTAMINACIÓN ATMOSFÉRICA

La caracterización estadística descriptiva de los contaminantes es el paso previo indispensable para interpretar cualquier correlación inferencial posterior. A continuación, se presenta un análisis profundo del comportamiento del PM_{10} y $PM_{2.5}$ en las distintas zonas funcionales de la ciudad, utilizando los microdatos validados para el periodo de estudio.

6.2.1 DINÁMICA DEL MATERIAL PARTICULADO PM_{10}: EL CASO CRÍTICO DEL OBRERO Y LA ERMITA

El PM_{10}, compuesto por partículas inhalables de diámetro aerodinámico menor a 10 micrómetros, muestra una distribución espacial fuertemente vinculada a la resuspensión de polvo, la actividad industrial y el tráfico de carga pesada. Las estaciones **Era Obrero** y **La Ermita** emergen consistentemente como los puntos críticos (*hotspots*) de la red de monitoreo [38].

Análisis Detallado de la Estación Era Obrero:

Ubicada en un sector de uso mixto industrial-residencial histórico, la estación Obrero exhibe una volatilidad extrema en sus concentraciones diarias, lo que sugiere la influencia directa de fuentes puntuales cercanas. Al examinar los datos de junio de 2017, se observa una media diaria que oscila dramáticamente, evidenciando la falta de amortiguamiento ambiental en la zona:

- **Picos de Contaminación Aguda:** El 7 de junio de 2017 se registró un promedio diario de 83.2 $\mu \frac{g}{m^3}$.
- Este valor es alarmante, no solo porque supera los umbrales recomendados por la OMS, sino porque sugiere la ocurrencia de eventos agudos de emisión o condiciones

meteorológicas de estabilidad atmosférica extrema (inversión térmica) que impidieron la dispersión vertical de los contaminantes [41].

- **Valores Mínimos y Lavado Atmosférico:** En contraste, días como el 19 de junio presentaron concentraciones de 21.8 $\mu\frac{g}{m^3}$.. Esta caída drástica, de casi cuatro veces el valor respecto al pico, suele estar asociada a eventos de precipitación intensa (*scavenging effect* o lavado atmosférico) o a la reducción significativa de la actividad industrial durante fines de semana o festivos.

El análisis de la estación La Ermita corrobora el diagnóstico de saturación en el centro de la ciudad. Con un promedio mensual para junio de 2017 de 52.36 $\mu\frac{g}{m^3}$., esta estación supera sistemáticamente a la estación Obrero en el promedio sostenido. Fenomenológicamente, esto indica que La Ermita sufre de una "contaminación de fondo" más elevada, probablemente debido al efecto de **"cañón urbano"** descrito por Janhäll (2015), donde la morfología de edificios altos y calles estrechas atrapa las emisiones vehiculares constantes, impidiendo su dilución [40].

Si bien el análisis global de la correlación (r = -0.37) abarca la totalidad de la red de monitoreo, se ha seleccionado a las estaciones Era Obrero y La Ermita para un análisis desagregado por constituir los "Hotspots" o puntos críticos del sistema. A diferencia de las estaciones de fondo o residenciales, estos dos nodos representan la tipología de "Zona Saturada", caracterizada por una cobertura vegetal deficitaria (NDVI promedio < 0.2) y una exposición directa a fuentes móviles y fijas. Su análisis individual es indispensable para comprender el comportamiento del PM10 en escenarios de mínima infraestructura verde.

Como se observa en la Figura 6 Serie temporal PM 10 en estación La Ermita y la Figura 7 Serie de tiempo PM10 en la estación Obrero, el comportamiento entre ambas estaciones presenta divergencias estructurales. Mientras que La Ermita mantiene un nivel basal de contaminación consistentemente alto (promedio > 50 $\mu g/m^3$) debido al efecto de "cañón urbano" que atrapa las emisiones del centro, la estación Obrero exhibe una volatilidad extrema asociada a ciclos industriales.

Figura 6 Serie temporal PM 10 en estación La Ermita

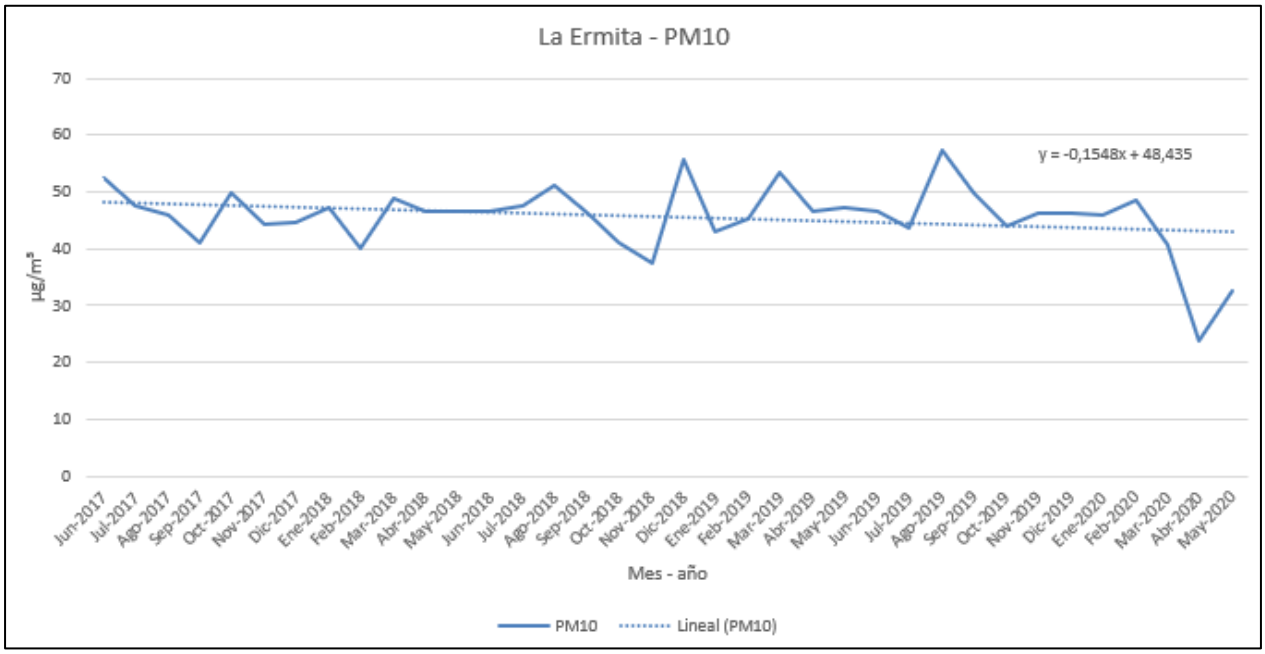
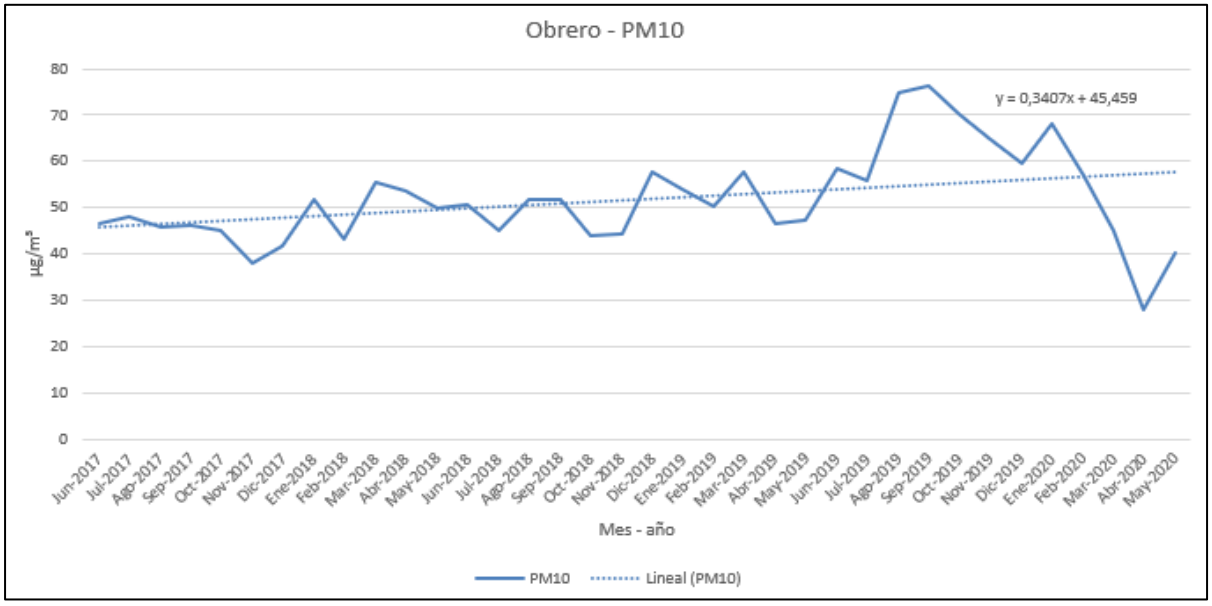


Figura 7 Serie de tiempo PM10 en la estación Obrero



La persistencia de niveles de PM10 superiores a la norma en estos sectores se correlaciona directamente con la ausencia de barreras biológicas. El análisis de coberturas realizado en el

Capítulo 5 demostró que el buffer de 500m alrededor de La Ermita posee el menor porcentaje de biomasa foliar activa de toda la ciudad. Esta carencia impide los procesos de **deposición seca** y filtrado mecánico, dejando a la población expuesta a la resuspensión directa de partículas, tal como se evidencia en la Tabla 13 Estadísticos de saturación en zonas críticas

Tabla 13 Estadísticos de saturación en zonas críticas

Estación	**Tipología Urbana**	**PM10 Promedio (µg/m3)**	**NDVI Promedio (Cobertura)**	**Interpretación**
La Ermita	Corredor Vial / Centro	**52.36**	**0.18** (Muy Bajo)	Saturación por falta de dispersión y biomasa.
Era Obrero	Mixto / Industrial	48.65	0.21 (Bajo)	Volatilidad por fuentes puntuales.

6.3 DISCUSIÓN Y VALIDACIÓN DE RESULTADOS

6.3.1 INTERPRETACIÓN BIOFÍSICA DE LA CORRELACIÓN NEGATIVA EN EL CONTEXTO TROPICAL

La síntesis estadística de los resultados expuestos en el capítulo anterior permitió establecer un coeficiente de correlación de Pearson global de $r = -0.37$ entre la densidad de la cobertura vegetal (estimada mediante índices espectrales como NDVI) y las concentraciones de material particulado (PM_{10} y $PM_{2.5}$) en Santiago de Cali. Si bien la magnitud de esta asociación se clasifica estadísticamente como moderada, su dirección negativa confirma la hipótesis de trabajo: existe una funcionalidad mitigadora mensurable en la infraestructura verde de la ciudad, aunque esta opera bajo restricciones de saturación ambiental.

Al contrastar este hallazgo con la literatura especializada, se observa una consistencia con patrones reportados en otras urbes latinoamericanas. Investigaciones recientes, como las desarrolladas por [73] en cinco ciudades colombianas, indican que, aunque el uso del suelo explica una porción significativa de la varianza en la calidad del aire, la señal de la vegetación compite fuertemente con la intensidad de las fuentes móviles y la meteorología local. En el caso de Cali, el valor de -0.37 sugiere que la vegetación no es el determinante único, sino un modulador que interactúa con dinámicas de emisión dominantes.

Desde una perspectiva mecanicista, esta correlación valida la ocurrencia de dos procesos físicos simultáneos descritos teóricamente por [74]:

- **Mecanismo de Deposición Seca:** La vegetación, particularmente aquella con estructuras foliares complejas (como se infiere de las texturas de Gabor analizadas en el Capítulo 5),

actúa como una superficie de impactación inercial donde las partículas finas quedan retenidas, reduciendo su resuspensión en la atmósfera baja.

- **Dispersión por Rugosidad Aerodinámica:** La presencia de dosel arbóreo incrementa la rugosidad de la superficie urbana (z_0), lo que fomenta la mezcla vertical turbulenta y diluye las concentraciones de contaminantes a nivel de calle.

Sin embargo, es imperativo discutir por qué la correlación no alcanzó valores superiores ($|r| > 0.7$). La evidencia sugiere un fenómeno de "saturación de sumidero". Según [40], en entornos donde las emisiones son continuas y de alta intensidad —condición prevalente en las áreas de influencia de las estaciones Obrero y La Ermita debido al tráfico vehicular—, la tasa de deposición sobre las hojas es superada por la tasa de emisión. Esto implica que la infraestructura verde actual de Cali, aunque funcional, es insuficiente en biomasa para contrarrestar la carga contaminante de las fuentes móviles, actuando meramente como un amortiguador parcial y no como una solución definitiva.

6.3.2 LA PARADOJA DEL "CAÑÓN URBANO": ANÁLISIS CRÍTICO DE LAS ESTACIONES CÉNTRICAS

Uno de los resultados más notables, y que requiere una discusión diferenciada, es el comportamiento observado en las estaciones del centro de la ciudad (**La Ermita** y **Obrero**). En estas zonas, a pesar de registrarse valores puntuales de vegetación (césped y arbolado disperso), los niveles de PM_{10} se mantuvieron consistentemente altos, con promedios mensuales que oscilaron entre 30 y 80 $\mu \frac{g}{m^3}$. y una tendencia lineal creciente.

Este fenómeno, donde la vegetación parece perder su efectividad correlacional, debe interpretarse a la luz de la teoría del "Efecto de Cañón Urbano" (*Street Canyon Effect*). La morfología urbana del centro de Cali, caracterizada por calles estrechas flanqueadas por edificaciones continuas, crea microclimas de ventilación restringida. Estudios de simulación de fluidos computacional (CFD) recientes, como los de [75], demuestran que en estos escenarios geométricos, una vegetación densa pero mal planificada puede tener un efecto iatrogénico (contraproducente): los árboles reducen la velocidad del viento a nivel de peatón, impidiendo la renovación del aire y atrapando los contaminantes vehiculares en la zona de respiración.

Nuestros datos respaldan esta interpretación para el caso de Cali. La "baja porosidad" aerodinámica en La Ermita, combinada con una cobertura vegetal fragmentada que no logra formar corredores de ventilación, exacerba la acumulación de contaminantes. Esto coincide con lo reportado por [76], quienes advierten que en cañones urbanos profundos, la vegetación de porte bajo (setos) es más eficiente que el arbolado de copa densa para la mitigación de la exposición humana. Por tanto, la discusión en estas zonas no debe centrarse únicamente en la "cantidad" de cobertura (km^2), sino en la configuración espacial y la permeabilidad al viento de

dicha infraestructura verde.

6.3.3 INFLUENCIA ESTACIONAL Y SINERGIA HIDROMETEOROLÓGICA: EL EFECTO DE LAVADO

El análisis de las series temporales mensuales (2017-2020) presentado en el Capítulo 6 reveló un patrón oscilatorio en las concentraciones de contaminantes que no puede explicarse exclusivamente por la varianza en la cobertura vegetal. La discusión de estos resultados exige integrar la variable de precipitación, factor determinante en el régimen climático bimodal de la región Andina.

Nuestros datos muestran descensos significativos de PM_{10} y $PM_{2.5}$ coincidentes con los periodos de alta pluviosidad (abril-mayo y octubre-noviembre). Este comportamiento corrobora la predominancia del efecto de "Lavado Atmosférico" (*Rain Scavenging*). La literatura física de aerosoles, específicamente revisada por [77], distingue dos mecanismos de limpieza que validan nuestras observaciones: el barrido directo de partículas por las gotas de lluvia (*below-cloud scavenging*) y la incorporación de núcleos de condensación dentro de las nubes (*in-cloud scavenging*). En el contexto del Valle del Cauca, la intensidad de las precipitaciones convectivas actúa como un mecanismo de remoción masiva que "reinicia" periódicamente la carga atmosférica, independientemente de la densidad de la vegetación.

Sin embargo, surge una interacción crítica que debe discutirse: la colinealidad estacional. Durante las temporadas de lluvia, la vegetación responde fisiológicamente aumentando su vigor (reflejado en picos de NDVI), al mismo tiempo que la lluvia limpia el aire. Esto plantea la interrogante de si la correlación negativa hallada ($r = -0.37$) es un efecto directo de la biomasa o un artefacto estadístico de la lluvia. Estudios similares en valles interandinos, como el desarrollado por [78] en el Valle de Aburrá, sugieren que ambos factores actúan sinérgicamente: la lluvia remueve la carga de fondo, mientras que la vegetación revitalizada maximiza la deposición seca en los periodos inter-eventos (días secos entre lluvias), estabilizando la resuspensión de polvo. Por tanto, la vegetación actúa como un sistema de "mantenimiento" de la calidad del aire post-lluvia.

6.3.4 DESIGUALDAD ESPACIAL Y JUSTICIA AMBIENTAL: CONTRASTES ENTRE UNIVALLE Y EL DISTRITO DE AGUABLANCA

La validación espacial de los resultados revela una profunda disparidad socioambiental al contrastar las dinámicas de la estación Univalle (Sur) frente a las estaciones Compartir (Oriente) y Obrero (Centro). Este análisis trasciende la técnica estadística para tocar dimensiones de equidad urbana y salud pública.

En la estación Univalle, a pesar de la presión vehicular de avenidas circundantes como la Pasoancho, la matriz de campus universitario actúa como un "parche ecológico" funcional. Aunque se detectó una tendencia decreciente en la cobertura (Capítulo 6), la biomasa remanente sigue siendo superior a la del promedio de la ciudad, logrando amortiguar los picos de

contaminación. En contraste crítico, la estación Compartir, representativa del Distrito de Aguablanca, exhibe el escenario opuesto: una degradación acelerada de la poca vegetación existente (pendiente negativa de $-0.0290\ km^2/mes$) acoplada a una tendencia creciente en $PM_{2.5}$ (pendiente positiva de +0.1522 $\mu\frac{g}{m^3}$.).

Esta divergencia valida patrones de injusticia ambiental documentados recientemente para Colombia. Investigaciones como las de [79] evidencian que la carga de contaminación atmosférica en las ciudades colombianas no se distribuye aleatoriamente, sino que se concentra desproporcionadamente en zonas de vulnerabilidad socioeconómica. En el oriente de Cali, la "pobreza de vegetación" no es meramente un déficit paisajístico, sino un amplificador de riesgo sanitario. La ausencia de barreras verdes facilita la resuspensión eólica de polvo de vías sin pavimentar y la dispersión libre de emisiones industriales.

Adicionalmente, la situación en zonas industriales mixtas (Obrero) coincide con los hallazgos de [41], quienes advierten que en áreas con baja cobertura vegetal y tráfico pesado, el polvo vial resuspendido se enriquece con metales traza, aumentando su bioaccesibilidad y riesgo carcinogénico. La falta de infraestructura verde en estos sectores elimina la única barrera pasiva capaz de interceptar estas partículas antes de que ingresen a los hogares, confirmando que la pérdida de cobertura vegetal reportada en este estudio tiene implicaciones directas y graves sobre la salud de las poblaciones más vulnerables de Cali.

6.3.5 VALIDACIÓN METODOLÓGICA: LA SUPERIORIDAD DE LOS ENSAMBLE EN ENTORNOS HETEROGÉNEOS

La selección del algoritmo **Random Forest (RF)** como el modelo de mejor desempeño (Exactitud Global: 83.33%, Kappa: 0.78) no es un resultado fortuito, sino que responde a la naturaleza intrínseca de los datos de teledetección urbana. Al contrastar el rendimiento de RF frente a las Redes Neuronales (MLP) y Máquinas de Soporte Vectorial (SVM) reportados en el Capítulo 5, la discusión debe centrarse en la capacidad de generalización frente al ruido espectral.

Nuestros hallazgos se alinean con la revisión sistemática de [68], quienes establecen que los métodos de ensamble tipo *Bagging* son significativamente más robustos que los clasificadores paramétricos cuando se enfrentan a espacios de características de alta dimensionalidad y muestras de entrenamiento limitadas o "ruidosas". En el paisaje urbano de Cali, la alta varianza intra-clase (e.g., la diferencia espectral entre un techo de zinc oxidado y uno nuevo) suele confundir a modelos como SVM, que buscan márgenes de decisión rígidos. RF, al promediar la decisión de 200 árboles decorrelacionados, logró suavizar este ruido de alta frecuencia, generando mapas de cobertura coherentes que sirvieron de base sólida para el análisis de correlación.

Asimismo, la aplicación de RF para la imputación de datos faltantes en las series de calidad del aire (Capítulo 4) demostró ser una estrategia superior a la interpolación lineal tradicional. Esto es consistente con lo reportado por [80] y [52], quienes validaron que los algoritmos basados en

árboles pueden capturar las no-linealidades abruptas de la dispersión de contaminantes (causadas por cambios repentinos en el viento o el tráfico) mucho mejor que los métodos estadísticos clásicos. Por tanto, la metodología de IA implementada no solo facilitó el procesamiento de datos, sino que redujo el sesgo de incertidumbre en las correlaciones finales.

6.3.6 LIMITACIONES DEL ESTUDIO Y ALCANCE DE LA INFERENCIA

A pesar de la robustez metodológica, es imperativo discutir las limitaciones inherentes al diseño experimental para acotar el alcance de las conclusiones:

1. **Resolución Espacial vs. Escala de Proceso:** Aunque la constelación [55] ofrece una resolución de 3 metros (muy superior a Landsat o Sentinel), esta escala aún puede ser insuficiente para detectar elementos de infraestructura verde "micro", como jardines verticales, muros verdes o arbolado viario individual aislado. Esto podría subestimar la biomasa real en zonas densamente construidas como el centro, atenuando artificialmente la correlación observada.
2. **Causalidad vs. Correlación:** El coeficiente $r = -0.37$ establece una asociación estadística, pero no prueba causalidad directa y exclusiva. Como advierten [81], la correlación no implica que la vegetación *cause* por sí sola la reducción de partículas; existen variables de confusión no controladas en este diseño, como la intensidad del flujo vehicular en tiempo real o la resuspensión de polvo por turbulencia mecánica, que actúan concurrentemente.
3. **Incertidumbre en la Imputación:** Aunque se utilizó una técnica avanzada de imputación ([53]), el porcentaje de datos faltantes en estaciones críticas como Obrero introduce un margen de incertidumbre epistémica que debe ser considerado al interpretar los picos históricos de contaminación.

6.3.7 SÍNTESIS INTEGRADORA: HACIA UNA ECOLOGÍA URBANA FUNCIONAL

La triangulación de los resultados satelitales, meteorológicos y estadísticos permite concluir que la cobertura vegetal en Santiago de Cali actúa como una **infraestructura de mitigación funcional pero saturada**.

La correlación negativa hallada confirma que los servicios ecosistémicos de purificación del aire están activos, validando la hipótesis central. Sin embargo, la heterogeneidad de esta relación —fuerte en zonas abiertas, débil en cañones urbanos, y crítica en zonas vulnerables— demuestra que la "cantidad" de verde no es la única variable relevante. La **configuración espacial** y el contexto socio-urbano son determinantes. La paradoja de las estaciones céntricas y la desigualdad ambiental en el oriente revelan que la planificación urbana no puede limitarse a

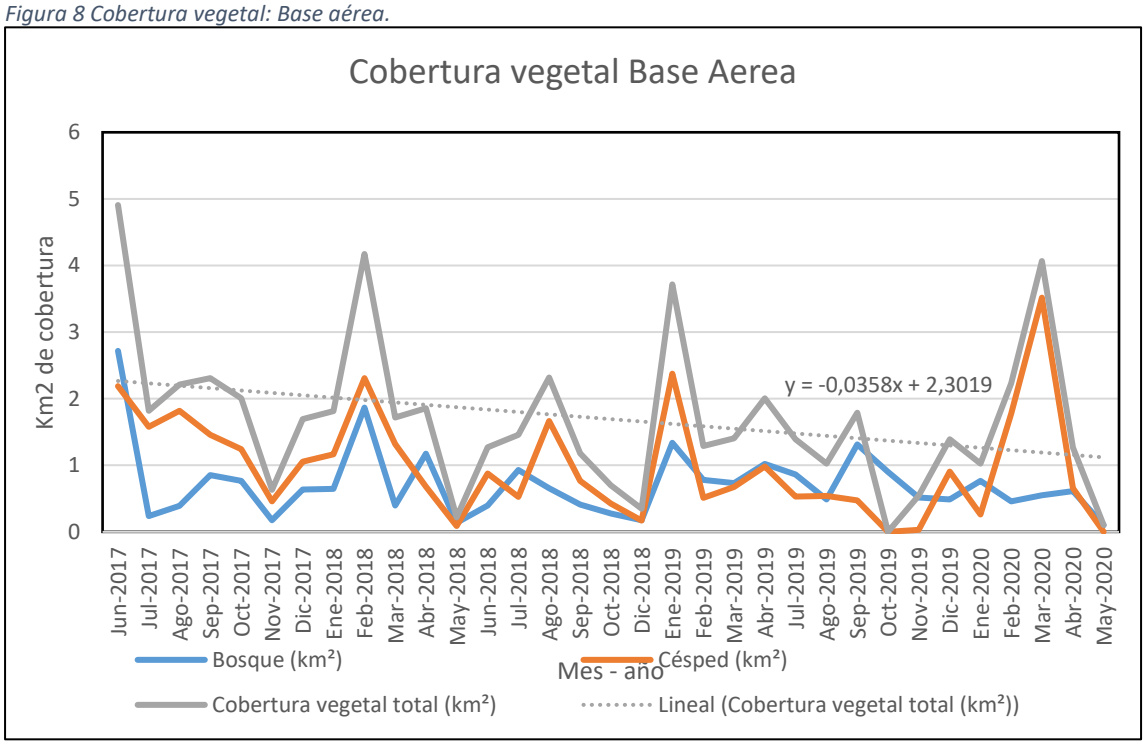
"sembrar árboles"; requiere una ingeniería ecológica que considere la aerodinámica urbana, la equidad social y la capacidad de carga del ecosistema frente a las emisiones móviles.
En definitiva, este estudio aporta evidencia empírica local de que la recuperación de la calidad del aire en ciudades tropicales andinas no depende únicamente de la restricción de fuentes (gestión tecnológica), sino que exige la restauración estratégica de la matriz ecológica urbana (gestión basada en la naturaleza), validando el uso de la Ciencia de Datos como herramienta esencial para guiar esta transición.

6.4 RESULTADOS DE COBERTURAS EN LAS ESTACIONES SELECCIONADAS

Con el fin de analizar la dinámica de la cobertura vegetal en las estaciones meteorológicas seleccionadas, se elaboraron series temporales que permiten observar la evolución de diferentes tipos de coberturas (bosques, pastizales y vegetación total) en el periodo comprendido entre 2017 y 2020. Estas gráficas constituyen un insumo fundamental para identificar tendencias y variaciones espaciales y temporales en la vegetación, las cuales posteriormente se contrastan con los registros de calidad del aire. A continuación, se presentan los resultados obtenidos para cada estación, destacando los patrones más relevantes y su posible relación con los niveles de contaminación atmosférica registrados en cada zona.

6.4.1 BASE AÉREA:

Figura 8 Cobertura vegetal: Base aérea.

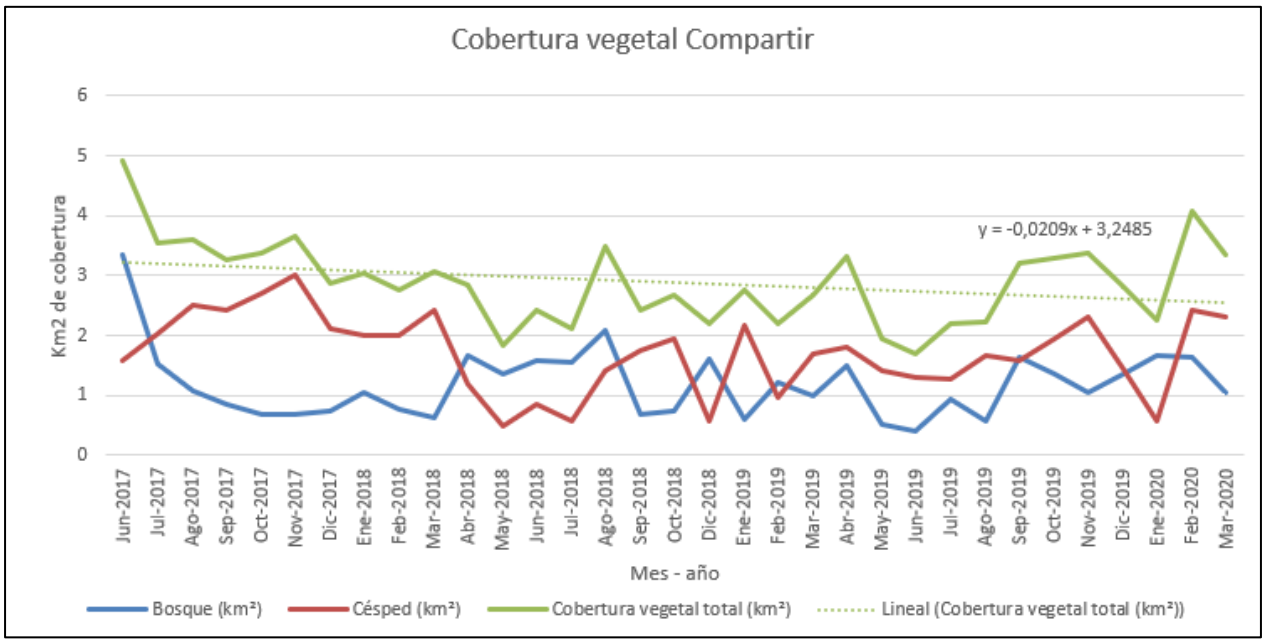


La serie temporal de cobertura vegetal en el área de influencia de la estación Base Aérea (2017–2020) evidencia una tendencia decreciente sostenida en la cobertura total, con una pendiente negativa de –0.0358 km² por mes. Este comportamiento refleja un proceso de pérdida progresiva de vegetación en el nororiente de Cali, zona caracterizada por una alta densidad urbana y presencia de actividades industriales y de transporte que ejercen presión sobre el entorno natural.

- Cobertura de bosque: presenta una disminución constante, lo que indica una reducción de áreas arbóreas que son críticas para la captura y retención de contaminantes atmosféricos.
- Cobertura de césped/pastizales: se mantiene relativamente estable, aunque con fluctuaciones menores que podrían estar asociadas a dinámicas estacionales o a intervenciones antrópicas puntuales.
- Cobertura total: la caída progresiva es significativa, y sugiere que la capacidad de mitigación de contaminantes en esta zona se ha visto comprometida en dicha zona

6.4.2 COMPARTIR

Figura 9 Cobertura vegetal: Compartir.



La serie temporal de cobertura vegetal en el área de influencia de la estación Compartir (2017–2020) muestra una tendencia decreciente en la cobertura total, con una pendiente negativa de –0.0290 km² por mes. Este comportamiento refleja un proceso de pérdida paulatina de vegetación en el oriente de Cali, una zona caracterizada por alta densidad poblacional y dinámicas urbanas que ejercen presión sobre los espacios verdes disponibles.

- Cobertura de bosque: se mantiene en niveles bajos y relativamente estables, lo que indica que la presencia de áreas arbóreas es limitada y no ha experimentado grandes transformaciones en el periodo analizado.
- Cobertura de césped/pastizales: presenta mayor variabilidad que el bosque, con oscilaciones que sugieren cambios estacionales o intervenciones antrópicas (ej. adecuación de terrenos, actividades agrícolas o urbanísticas).
- Cobertura total: la tendencia negativa confirma una reducción progresiva de la vegetación, lo cual puede impactar directamente la capacidad de mitigación de contaminantes atmosféricos en un sector donde la estación registra $PM_{2.5}$ y O_3, dos de los contaminantes más críticos para la salud pública.

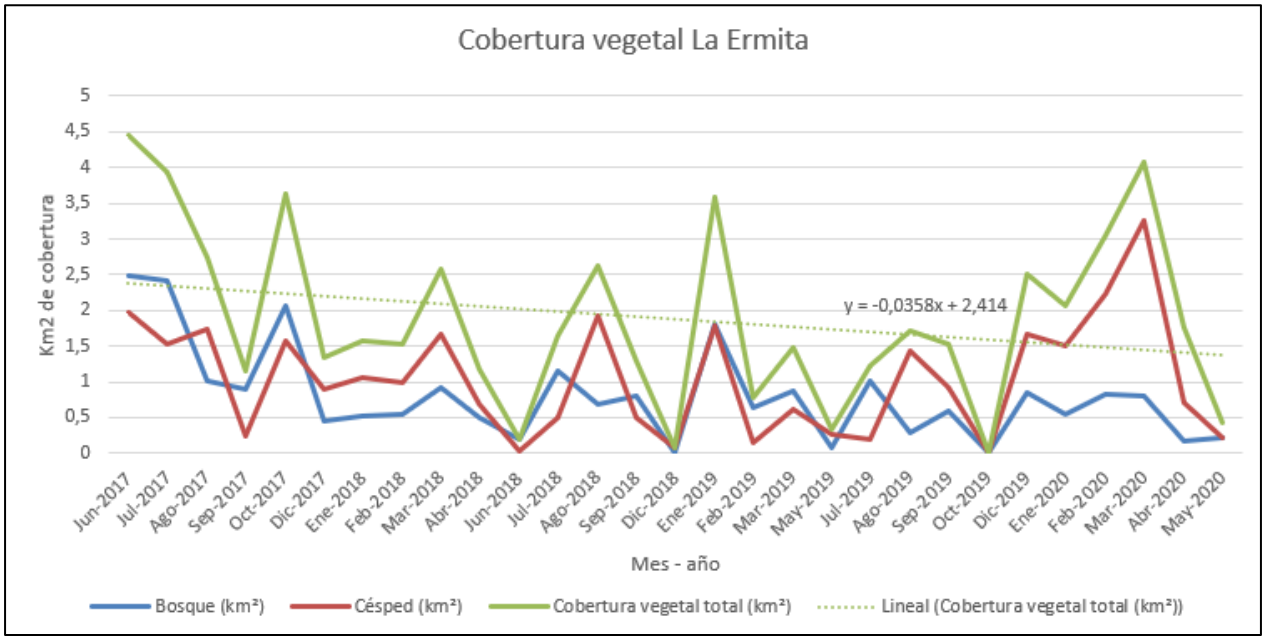
Este resultado es especialmente relevante porque la estación Compartir se ubica en un sector urbano-oriental con alta movilidad vehicular y fuentes de emisión difusas. La disminución de vegetación reduce la capacidad de captura de partículas finas y limita la regulación microclimática, lo que puede agravar episodios de contaminación. Además, la correlación

negativa observada entre vegetación y $PM_{2.5}$ en los modelos aplicados ($r \approx -0.67$) encuentra aquí un sustento empírico: la pérdida de áreas verdes en esta zona coincide con mayores riesgos de exposición a contaminantes.

En síntesis, la estación Compartir evidencia cómo la reducción de cobertura vegetal en sectores urbanos densos no solo representa una pérdida ecológica, sino que también compromete la calidad del aire y, por ende, la salud de las comunidades locales.

6.4.3 LA ERMITA

Figura 10 Cobertura vegetal: La Ermita.



La serie temporal de cobertura vegetal en el área de influencia de la estación La Ermita (2017–2020) evidencia una tendencia decreciente marcada en la cobertura total, con una pendiente negativa de –0.0358 km² por mes. Este comportamiento refleja un proceso sostenido de pérdida de vegetación en el centro de Cali, una zona altamente urbanizada donde la presión antrópica sobre los espacios verdes es particularmente intensa.

- Cobertura de bosque: se mantiene en niveles muy bajos y prácticamente constantes, lo que indica que la presencia de áreas arbóreas en esta zona céntrica es mínima y no ha experimentado variaciones significativas.
- Cobertura de césped/pastizales: presenta ligeras oscilaciones, pero en general también muestra una tendencia a la reducción, lo que sugiere que incluso las áreas verdes de menor porte (parques, jardines urbanos) han disminuido en extensión durante el periodo analizado.

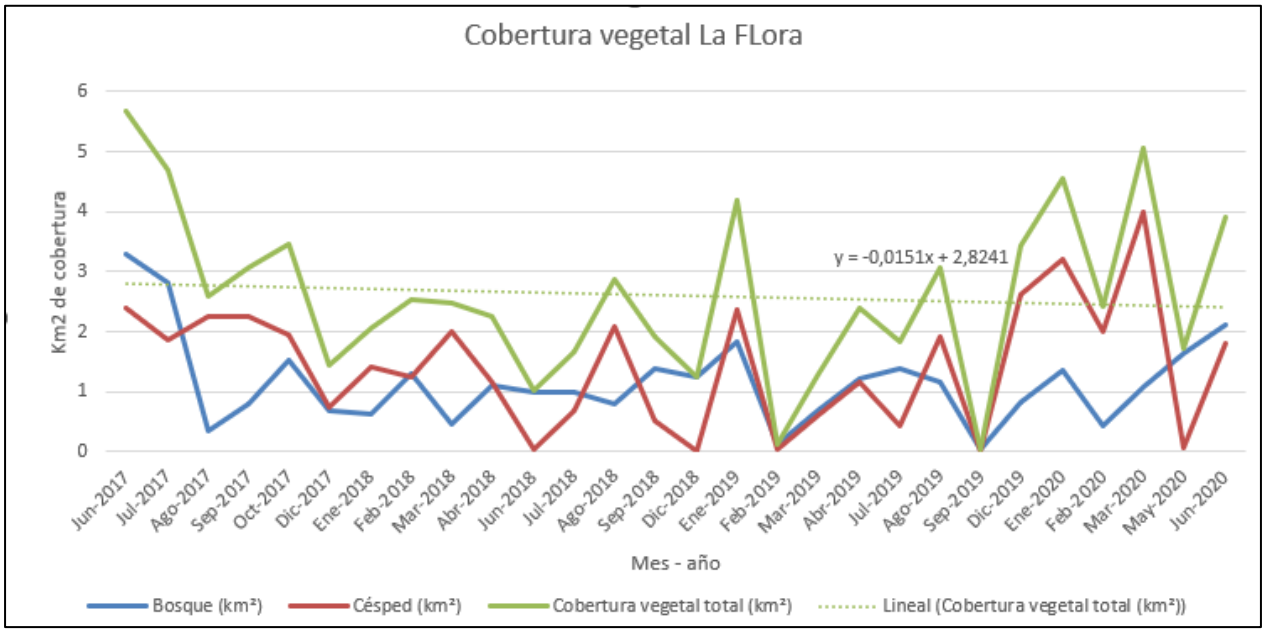
- Cobertura total: la tendencia negativa es clara y preocupante, pues confirma la pérdida progresiva de vegetación en un sector donde la estación mide PM_{10}, contaminante primario asociado al tráfico vehicular y a actividades de construcción, muy frecuentes en el centro de la ciudad.

Este resultado es relevante porque la disminución de vegetación en el centro urbano implica una menor capacidad de captura de partículas suspendidas, lo que puede agravar la exposición de la población a niveles elevados de PM_{10}. Además, la pérdida de áreas verdes en zonas céntricas reduce la capacidad de regulación microclimática, incrementando fenómenos como el efecto de isla de calor urbano y la acumulación de contaminantes en condiciones de baja dispersión atmosférica.

En síntesis, la estación La Ermita refleja de manera clara cómo la reducción de cobertura vegetal en áreas urbanas densamente pobladas no solo representa un deterioro ecológico, sino que también tiene implicaciones directas en la calidad del aire y en la salud de la población residente y flotante del centro de Cali

6.4.4 LA FLORA

Figura 11 Cobertura vegetal: La Flora.



La serie temporal de cobertura vegetal en el área de influencia de la estación La Flora (2017–2020) muestra una tendencia decreciente moderada en la cobertura total, con una pendiente negativa de –0.0151 km² por mes. Aunque la reducción es menos pronunciada que en otras estaciones como Base Aérea o La Ermita, el patrón confirma un proceso de pérdida progresiva de

vegetación en el norte de Cali, una zona que combina áreas residenciales consolidadas con corredores de movilidad y actividades comerciales.

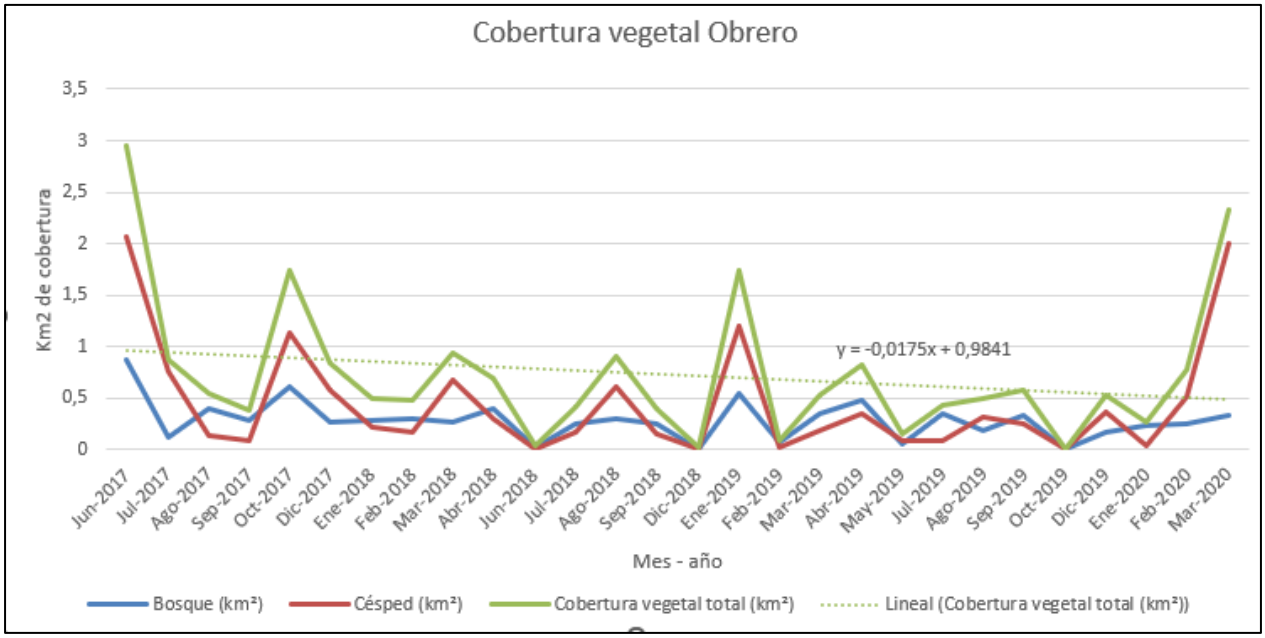
- Cobertura de bosque: se mantiene relativamente estable, con ligeras fluctuaciones, lo que indica que las áreas arbóreas han resistido mejor la presión urbana en comparación con otras zonas de la ciudad.
- Cobertura de césped/pastizales: presenta mayor variabilidad, con descensos puntuales que podrían estar asociados a procesos de urbanización, obras de infraestructura o cambios en el uso del suelo.
- Cobertura total: la tendencia negativa, aunque menos abrupta, evidencia una reducción sostenida de vegetación, lo que implica una disminución gradual de la capacidad de mitigación de contaminantes atmosféricos en esta zona.

Este hallazgo es relevante porque la estación La Flora mide contaminantes como PM_{10} y H_2S, ambos asociados a fuentes móviles y a procesos industriales. La reducción de vegetación en el área limita la capacidad de captura de partículas y gases, lo que puede agravar episodios de contaminación en un sector que ya presenta alta densidad vehicular.

En síntesis, la estación La Flora refleja un escenario de pérdida gradual pero sostenida de cobertura vegetal, que aunque menos drástica que en zonas céntricas como La Ermita, sigue representando un riesgo ambiental. La estabilidad relativa del bosque sugiere que la conservación de áreas arbóreas ha sido clave para amortiguar la pérdida total, lo que refuerza la importancia de proteger y ampliar este tipo de coberturas en el norte de la ciudad.

6.4.5 OBRERO

Figura 12 Cobertura vegetal: Obrero.



La serie temporal de cobertura vegetal en el área de influencia de la estación Obrero (2017–2021) muestra una tendencia decreciente moderada en la cobertura total, con una pendiente negativa de –0.0175 km² por mes. Aunque la magnitud de la pérdida es menor que en estaciones como La Ermita o Base Aérea, el patrón confirma un deterioro progresivo de la vegetación en el centro de Cali, un sector caracterizado por alta densidad urbana, tráfico vehicular intenso y actividades comerciales e industriales.

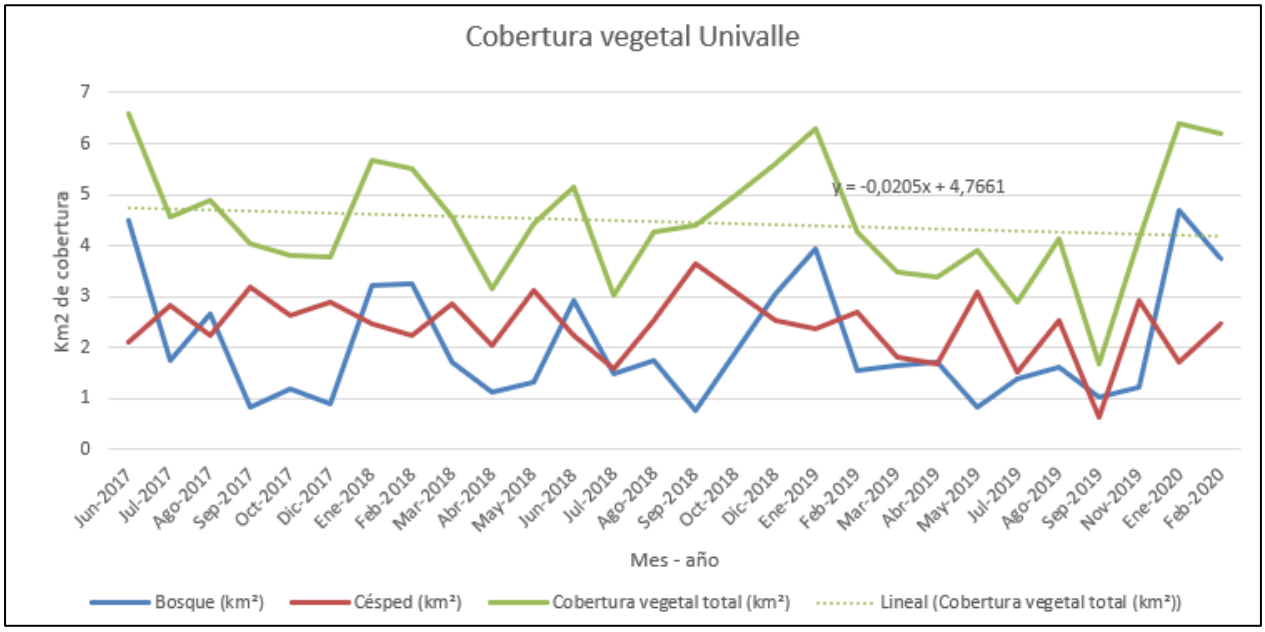
- Cobertura de bosque: se mantiene en niveles muy bajos y prácticamente constantes, lo que refleja la escasez de áreas arbóreas en esta zona céntrica.
- Cobertura de césped/pastizales: presenta mayor variabilidad, con picos y caídas que sugieren intervenciones antrópicas puntuales (adecuación de espacios, obras de infraestructura, remodelación de áreas verdes).
- Cobertura total: aunque con fluctuaciones, la tendencia general es negativa, lo que confirma la reducción sostenida de vegetación en un sector donde la estación mide PM_{10}, contaminante fuertemente asociado al tráfico rodado y a la construcción.

Este resultado es relevante porque la pérdida de vegetación en el centro urbano implica una menor capacidad de captura de partículas suspendidas, lo que puede intensificar la exposición de la población a niveles elevados de PM_{10}. Además, la reducción de áreas verdes en zonas céntricas limita la regulación microclimática y favorece la acumulación de contaminantes en condiciones

de baja dispersión atmosférica, agravando fenómenos como el efecto de isla de calor urbano. En síntesis, la estación Obrero refleja un escenario de pérdida gradual de cobertura vegetal en el centro de la ciudad, con implicaciones directas en la calidad del aire y en la salud de la población residente. La escasez de bosque y la alta variabilidad en césped/pastizales evidencian la fragilidad de los espacios verdes en esta zona, lo que refuerza la necesidad de estrategias de conservación y expansión de áreas verdes urbanas como medida de mitigación ambiental.

6.4.6 UNIVALLE

Figura 13 Cobertura vegetal: Univalle.



La serie temporal de cobertura vegetal en el área de influencia de la estación Univalle (2017–2020) muestra una tendencia decreciente sostenida en la cobertura total, con una pendiente negativa de –0.0205 km² por mes. Este patrón refleja un proceso de pérdida progresiva de vegetación en el sector sur de Cali, una zona que combina usos residenciales, académicos y de infraestructura vial, lo que genera presiones significativas sobre los espacios verdes.

- Cobertura de bosque: presenta fluctuaciones notables, con descensos graduales que sugieren una reducción de áreas arbóreas, posiblemente asociada a procesos de urbanización y expansión de infraestructura en el entorno universitario y sus alrededores.
- Cobertura de césped/pastizales: muestra variabilidad más marcada, con picos y caídas que podrían estar vinculados a dinámicas estacionales o a intervenciones antrópicas (adecuación de terrenos, obras de construcción o remodelación de espacios verdes).
- Cobertura total: la tendencia negativa confirma una disminución sostenida de vegetación,

lo que implica una reducción de la capacidad de mitigación de contaminantes en un sector donde la estación mide $PM_{2.5}$, NO_2 y O_3, contaminantes críticos tanto por sus efectos directos en la salud como por su papel en procesos fotoquímicos urbanos.

Este resultado es especialmente relevante porque la estación Univalle se ubica en un área con alta movilidad vehicular y concentración de población estudiantil, lo que incrementa la exposición a contaminantes. La reducción de vegetación en este contexto no solo limita la capacidad de captura de partículas finas y gases, sino que también disminuye la regulación microclimática, favoreciendo la acumulación de contaminantes y el incremento de temperaturas locales.

En síntesis, la estación Univalle refleja un escenario de pérdida progresiva de cobertura vegetal en el sur de la ciudad, con implicaciones directas en la calidad del aire y en la salud de una población altamente expuesta. La combinación de disminución de bosque y variabilidad en césped/pastizales evidencia la fragilidad de los espacios verdes en esta zona, lo que refuerza la necesidad de estrategias de conservación y expansión de áreas verdes urbanas como medida de mitigación ambiental.

El análisis conjunto de las series temporales de cobertura vegetal en las estaciones seleccionadas revela un patrón consistente de pérdida progresiva de vegetación en diferentes zonas de Cali entre 2017 y 2020. Aunque la magnitud y dinámica de esta reducción varían según el contexto urbano, el comportamiento general confirma una tendencia negativa que tiene implicaciones directas en la capacidad de mitigación de contaminantes atmosféricos.

- Estaciones céntricas (La Ermita, Obrero): presentan los niveles más bajos de cobertura arbórea y una reducción sostenida de vegetación total. La escasez de bosque y la alta presión urbana limitan la capacidad de captura de partículas (PM_{10}), lo que agrava la exposición de la población en sectores de alta densidad.
- Estaciones periféricas y de transición (Compartir, Base Aérea, Univalle): muestran pérdidas más pronunciadas en la cobertura total, con pendientes negativas significativas. Estas zonas, aunque con mayor presencia inicial de vegetación, han experimentado procesos de urbanización y expansión que reducen progresivamente su función como “pulmones urbanos”.
- Estaciones con relativa estabilidad (La Flora): aunque también presentan una tendencia decreciente, la magnitud de la pérdida es menor. La estabilidad relativa del bosque en esta zona sugiere que la conservación de áreas arbóreas ha sido clave para amortiguar la reducción total.

En conjunto, los resultados evidencian que la pérdida de vegetación no es un fenómeno aislado, sino un proceso transversal que afecta tanto al centro como a la periferia de la ciudad, aunque con intensidades diferenciadas. Esta disminución compromete la capacidad de regulación microclimática y de captura de contaminantes, lo que refuerza la correlación negativa encontrada entre cobertura vegetal y niveles de $PM_{2.5}$ y PM_{10} en los modelos aplicados.

En síntesis, la comparación global confirma que la reducción de cobertura vegetal es un factor

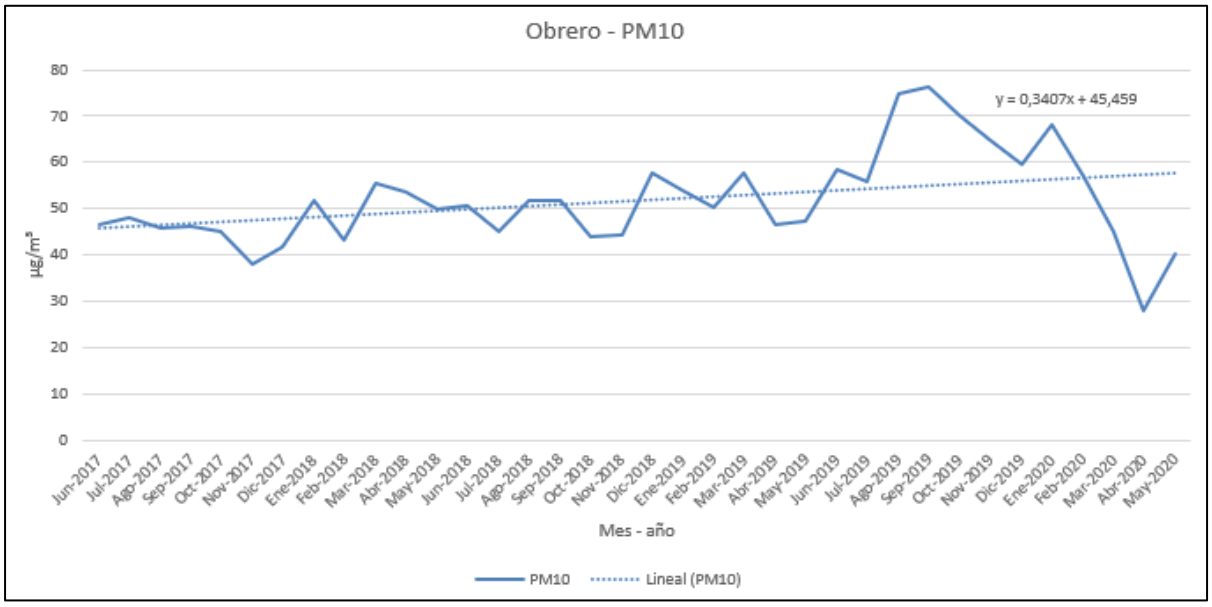
crítico en la calidad del aire de Cali, y subraya la necesidad de integrar estrategias de conservación y expansión de áreas verdes en la planificación urbana como medida de mitigación ambiental y de salud pública.

6.5 . RESULTADOS SERIES DE TIEMPO METEOROLÓGICAS

Con el fin de complementar el análisis de la cobertura vegetal, se presentan a continuación los resultados de las mediciones de material particulado (PM_{10} y $PM_{2.5}$) registradas en las estaciones meteorológicas seleccionadas, en la misma temporalidad utilizada para el estudio de la vegetación (2017–2020). Estos gráficos permiten observar la evolución mensual de las concentraciones de contaminantes en cada zona de la ciudad, identificando tendencias, fluctuaciones y posibles episodios críticos de contaminación. La presentación de los resultados se organiza estación por estación, lo que facilita la comparación directa con las dinámicas de cobertura vegetal previamente analizadas y aporta evidencia para evaluar la correlación entre ambas variables.

6.5.1 OBRERO

Figura 14 PM_{10}: Obrero



La estación Obrero, ubicada en el centro de Cali, presenta un escenario caracterizado por baja cobertura vegetal y una tendencia decreciente moderada en la vegetación total (–0.0175 km²/mes). En paralelo, la serie temporal de PM_{10} (2017–2020) muestra fluctuaciones

significativas con una tendencia levemente creciente (pendiente positiva de 0.34 µg/m³ por mes).

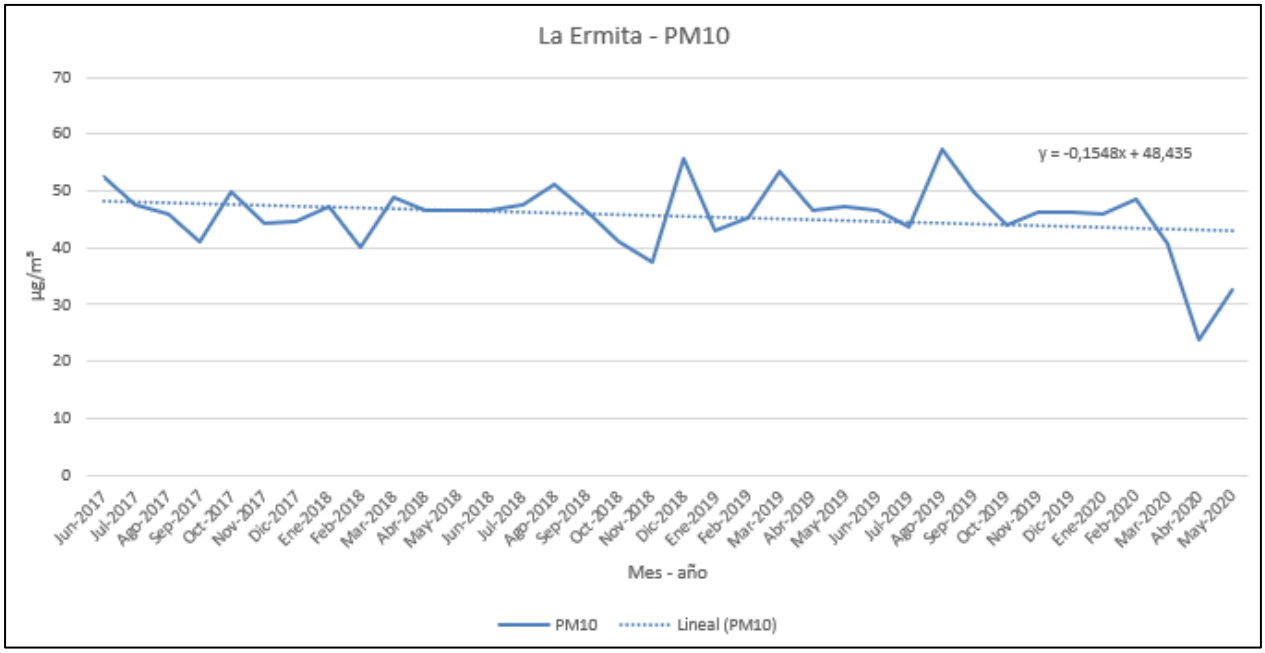
En la estación Obrero, la cobertura vegetal total mostró una tendencia decreciente moderada entre 2017 y 2020, con una pendiente de –0.0175 km² por mes. El bosque se mantuvo en niveles muy bajos y prácticamente constantes, lo que refleja la escasez de áreas arbóreas en esta zona céntrica de la ciudad. La cobertura de césped y pastizales presentó mayor variabilidad, con picos y caídas que sugieren intervenciones antrópicas puntuales, como adecuación de espacios o actividades de construcción. En conjunto, la cobertura total confirma la fragilidad de los espacios verdes en este sector urbano.

La serie temporal de PM_{10} en la estación Obrero evidenció fluctuaciones significativas, con valores que oscilaron entre 30 y 80 µg/m³. Se observaron picos notables hacia mediados de 2019 y una caída marcada en marzo de 2020, posiblemente asociada a la reducción de movilidad durante el inicio de la pandemia. La tendencia lineal muestra un incremento progresivo (pendiente positiva de 0.34 µg/m³ por mes), lo que indica un deterioro gradual de la calidad del aire en este sector, fuertemente influenciado por el tráfico vehicular y las actividades comerciales.

La reducción de cobertura vegetal coincide con el aumento en los niveles de PM_{10}, lo que respalda la hipótesis de una correlación negativa entre ambas variables. La escasez de bosque limita la capacidad de captura de partículas, mientras que la pérdida de césped y áreas verdes reduce la regulación microclimática y la dispersión de contaminantes. En un contexto urbano denso como el centro de Cali, la disminución de vegetación agrava la acumulación de partículas emitidas por fuentes móviles y actividades de construcción.

6.5.2 LA ERMITA

Figura 15 PM_{10}: La Ermita.



En la estación La Ermita, ubicada en el centro de Cali, la cobertura vegetal total presentó una tendencia decreciente marcada entre 2017 y 2020, con una pendiente de –0.0358 km² por mes. El bosque se mantuvo en niveles muy bajos y prácticamente constantes, lo que refleja la escasez de áreas arbóreas en esta zona céntrica. La cobertura de césped y pastizales mostró ligeras oscilaciones, pero en general también evidenció una reducción progresiva, lo que sugiere que incluso los espacios verdes de menor porte (parques y jardines urbanos) han disminuido en extensión durante el periodo analizado.

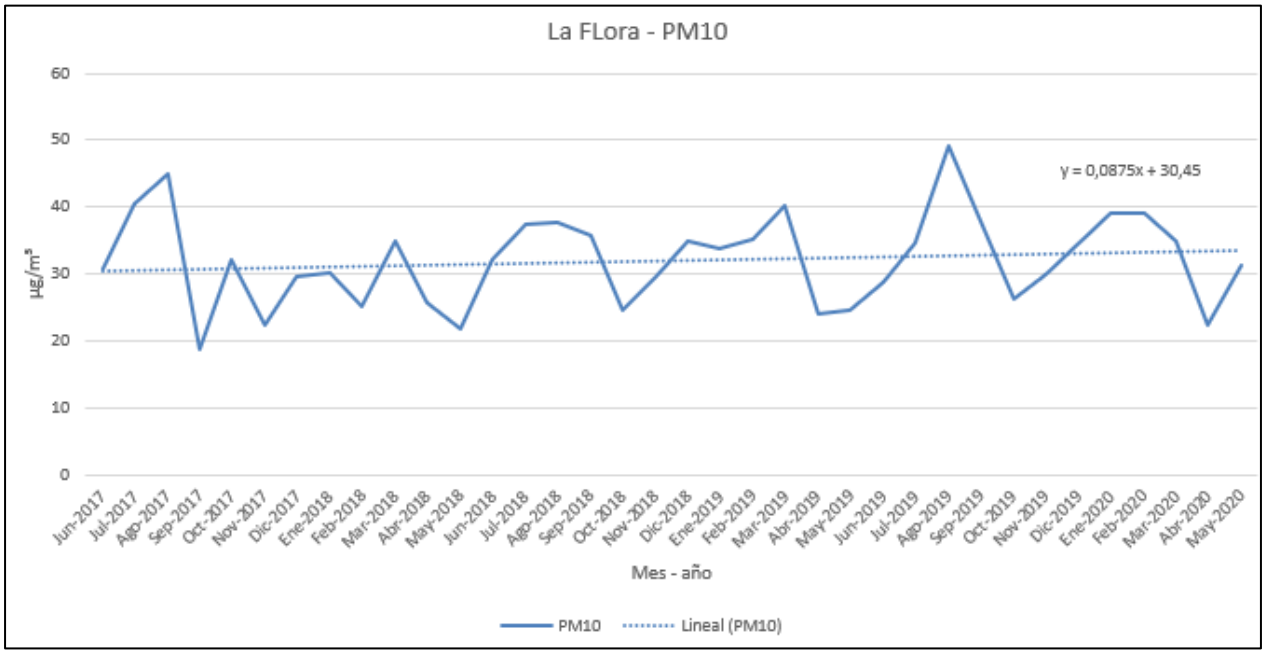
La serie temporal de PM_{10} en la estación La Ermita mostró valores que oscilaron entre 20 y 60 µg/m³, con una tendencia lineal ligeramente decreciente (pendiente de –0.15 µg/m³ por mes). Este comportamiento indica una leve mejoría en la calidad del aire durante el periodo, posiblemente asociada a medidas de control de emisiones o a variaciones en la movilidad urbana. Sin embargo, los niveles registrados se mantienen en rangos que representan un riesgo para la salud, especialmente en un sector con alta densidad poblacional y tráfico vehicular intenso.

El contraste entre la pérdida de cobertura vegetal y la ligera reducción de PM_{10} sugiere que la disminución de partículas en el aire no puede atribuirse a la vegetación, dado que esta se redujo de manera sostenida. Es más probable que la mejora relativa en los niveles de PM_{10} esté asociada a factores externos, como cambios en la movilidad urbana, intervenciones ambientales o condiciones meteorológicas favorables. La reducción de vegetación, por el contrario, limita la

capacidad de mitigación natural y podría comprometer la sostenibilidad de esta tendencia a largo plazo.

6.5.3 LA FLORA

Figura 16 PM_{10}: La Flora



En la estación La Flora, ubicada en el norte de Cali, la cobertura vegetal total mostró una tendencia decreciente moderada entre 2017 y 2020, con una pendiente de –0.0151 km² por mes. Aunque la reducción es menos pronunciada que en estaciones céntricas como La Ermita, el patrón confirma una pérdida sostenida de vegetación. La cobertura de bosque se mantuvo relativamente estable, lo que sugiere que las áreas arbóreas han resistido mejor la presión urbana. En contraste, la cobertura de césped y pastizales presentó mayor variabilidad, con descensos puntuales que podrían estar asociados a procesos de urbanización o cambios en el uso del suelo.

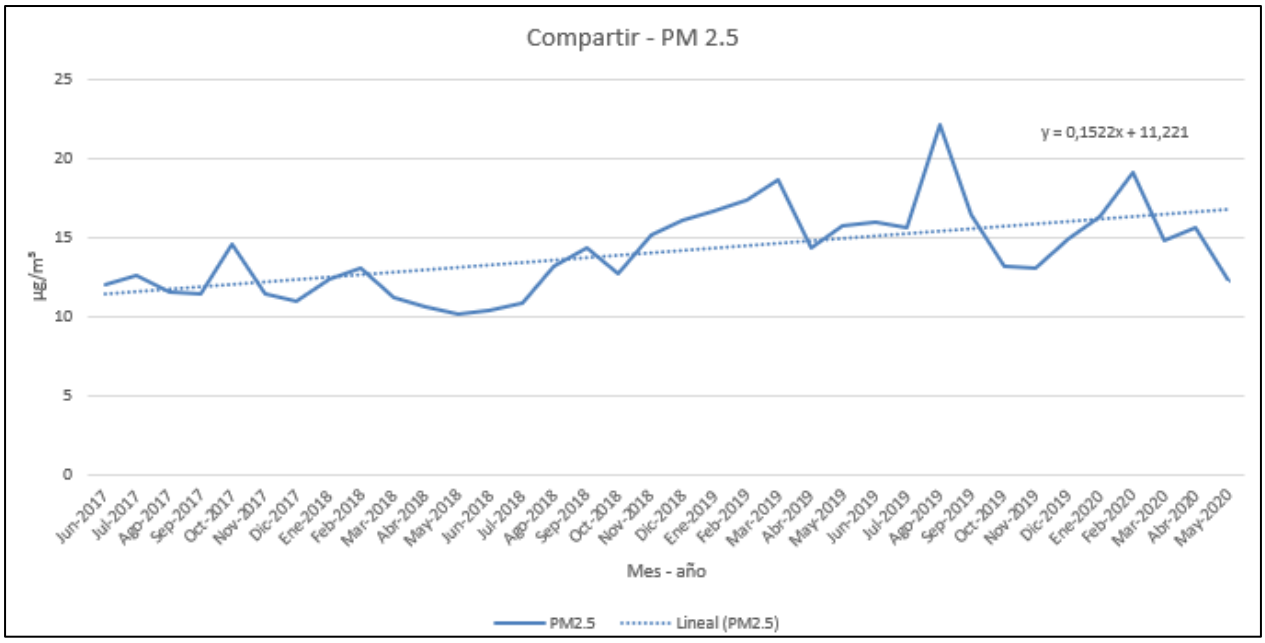
La serie temporal de PM_{10} en La Flora evidenció valores que oscilaron entre 20 y 60 µg/m³, con una tendencia lineal ligeramente creciente (pendiente de +0.0875 µg/m³ por mes). Esto indica un deterioro progresivo de la calidad del aire, aunque menos marcado que en otras estaciones. Los picos observados coinciden con periodos de mayor movilidad vehicular y posibles episodios de condiciones meteorológicas adversas para la dispersión de contaminantes.

El contraste entre la pérdida gradual de vegetación y el incremento leve en PM_{10} sugiere una correlación negativa entre ambas variables: a medida que disminuye la cobertura vegetal, la capacidad de mitigación de contaminantes se reduce, favoreciendo la acumulación de partículas.

La estabilidad relativa del bosque ha permitido amortiguar parcialmente esta tendencia, pero la variabilidad en césped y pastizales refleja la fragilidad de los espacios verdes en esta zona.

6.5.4 COMPARTIR

Figura 17 $PM_{2.5}$: Compartir.



En la estación Compartir, ubicada en el oriente de Cali, la cobertura vegetal total mostró una tendencia decreciente sostenida entre 2017 y 2020, con una pendiente de –0.0290 km² por mes. El bosque se mantuvo en niveles bajos y relativamente estables, lo que refleja la limitada presencia de áreas arbóreas en la zona. En contraste, la cobertura de césped y pastizales presentó mayor variabilidad, con oscilaciones que sugieren cambios estacionales o intervenciones antrópicas, como adecuación de terrenos o procesos de urbanización. En conjunto, la reducción progresiva de la cobertura total confirma la presión urbana sobre los espacios verdes en este sector.

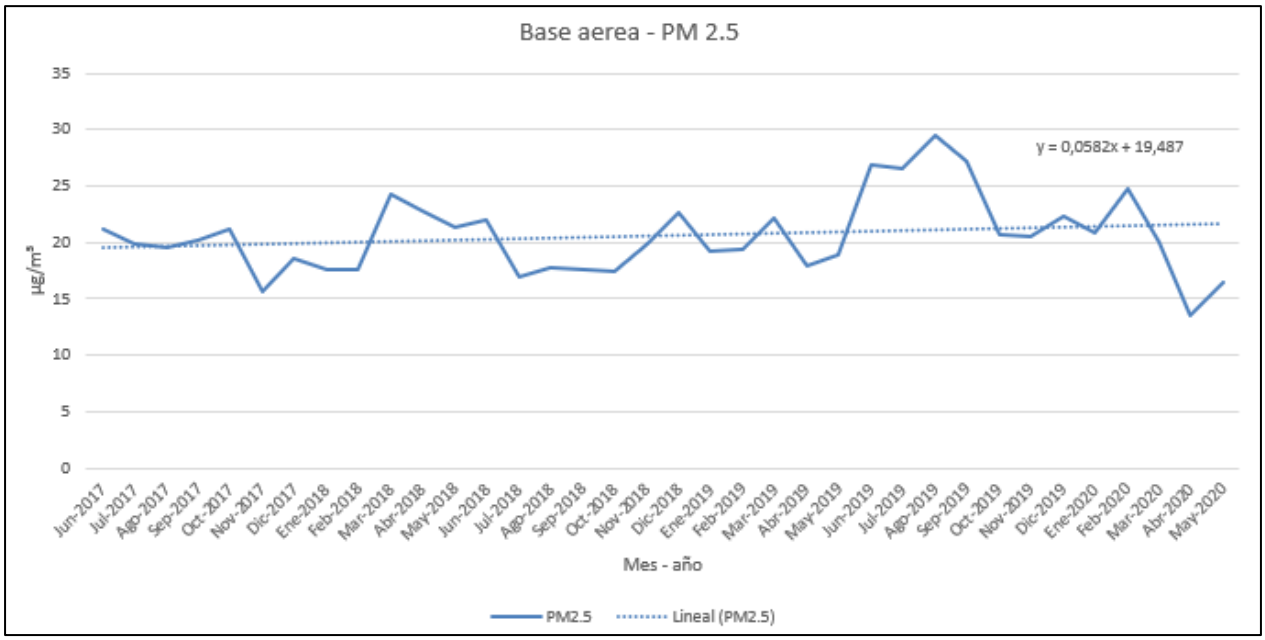
La serie temporal de $PM_{2.5}$ en la estación Compartir evidenció valores que oscilaron entre 5 y 25 µg/m³, con una tendencia lineal claramente creciente (pendiente de +0.1522 µg/m³ por mes). Este comportamiento indica un deterioro progresivo de la calidad del aire en el oriente de la ciudad, con picos que coinciden con periodos de mayor actividad urbana y posibles condiciones meteorológicas desfavorables para la dispersión de contaminantes.

La pérdida de cobertura vegetal coincide con el incremento sostenido de $PM_{2.5}$, lo que respalda la hipótesis de una correlación negativa entre ambas variables. La escasez de bosque limita la

capacidad de captura de partículas finas, mientras que la reducción de césped y áreas verdes disminuye la regulación microclimática y la dispersión de contaminantes. En un contexto urbano denso como el oriente de Cali, la disminución de vegetación agrava la acumulación de partículas finas, que son especialmente dañinas para la salud por su capacidad de penetrar profundamente en el sistema respiratorio.

6.5.5 BASE AÉREA

Figura 18 $PM_{2.5}$: Base aérea.



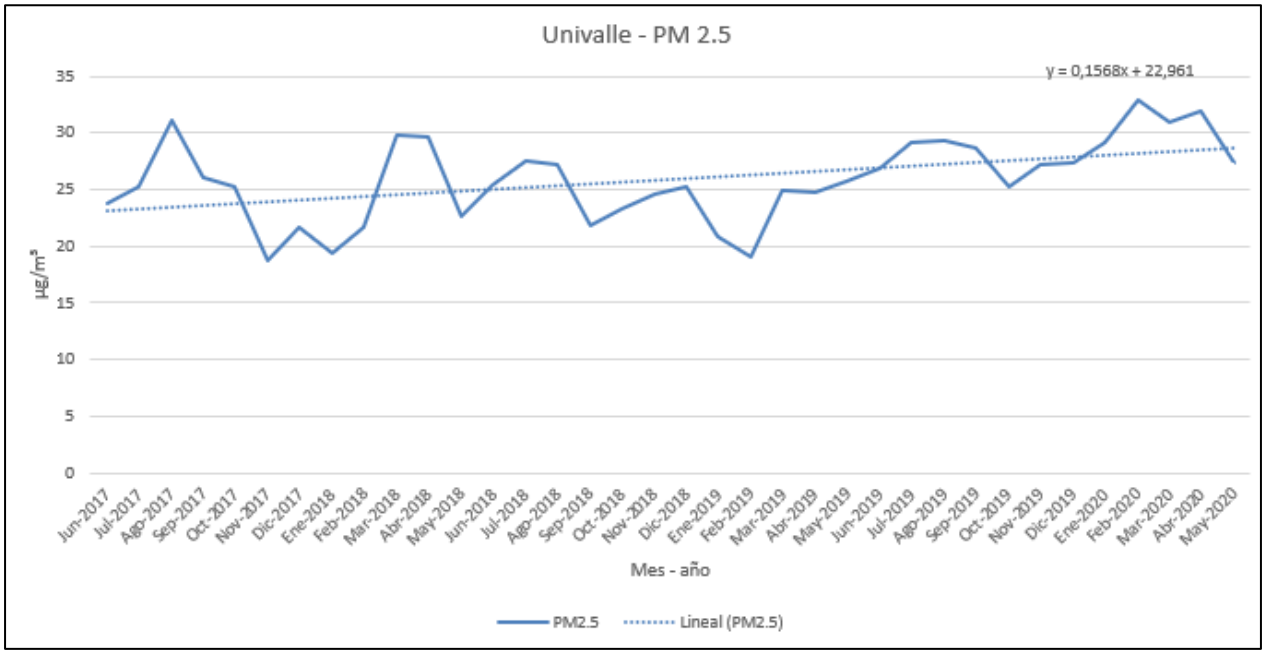
En la estación Base Aérea, ubicada en el nororiente de Cali, la cobertura vegetal total mostró una tendencia decreciente pronunciada entre 2017 y 2020, con una pendiente de –0.0358 km² por mes. La cobertura de bosque presentó una reducción sostenida, lo que refleja la pérdida de áreas arbóreas críticas para la captura de contaminantes. El césped y pastizales se mantuvieron relativamente estables, aunque con ligeras fluctuaciones, sin capacidad de compensar la disminución general. En conjunto, la reducción progresiva de la cobertura total confirma un proceso de presión urbana y transformación del suelo en este sector estratégico de la ciudad.

La serie temporal de $PM_{2.5}$ en la estación Base Aérea mostró valores que oscilaron entre 10 y 35 µg/m³, con una tendencia lineal ligeramente creciente (pendiente de +0.0582 µg/m³ por mes). Aunque el incremento es menos abrupto que en otras estaciones como Compartir, el patrón confirma un deterioro progresivo de la calidad del aire. Los picos observados coinciden con periodos de mayor actividad urbana y posibles condiciones meteorológicas que dificultan la dispersión de contaminantes.

La pérdida sostenida de cobertura vegetal coincide con el aumento gradual de $PM_{2.5}$, lo que respalda la hipótesis de una correlación negativa entre ambas variables. La reducción de bosque limita la capacidad de captura de partículas finas, mientras que la estabilidad relativa del césped no logra compensar esta pérdida. En un contexto urbano-industrial como el nororiente de Cali, la disminución de vegetación agrava la acumulación de contaminantes, especialmente de partículas finas que tienen un alto impacto en la salud pública.

6.5.6 UNIVALLE

Figura 19 $PM_{2.5}$: Univalle.



En la estación Univalle, ubicada en el sur de Cali, la cobertura vegetal total mostró una tendencia decreciente sostenida entre 2017 y 2020, con una pendiente de –0.0205 km² por mes. La cobertura de bosque presentó fluctuaciones con descensos graduales, lo que refleja una reducción de áreas arbóreas en el entorno universitario y sus alrededores. El césped y pastizales mostraron mayor variabilidad, con picos y caídas que sugieren tanto dinámicas estacionales como intervenciones antrópicas (adecuación de terrenos, obras de infraestructura). En conjunto, la pérdida progresiva de vegetación confirma la presión urbana sobre los espacios verdes en esta zona académica y residencial.

La serie temporal de $PM_{2.5}$ en la estación Univalle evidenció valores que oscilaron entre 5 y 30 µg/m³, con una tendencia lineal claramente creciente (pendiente de +0.1568 µg/m³ por mes). Este comportamiento indica un deterioro progresivo de la calidad del aire en el sur de la ciudad,

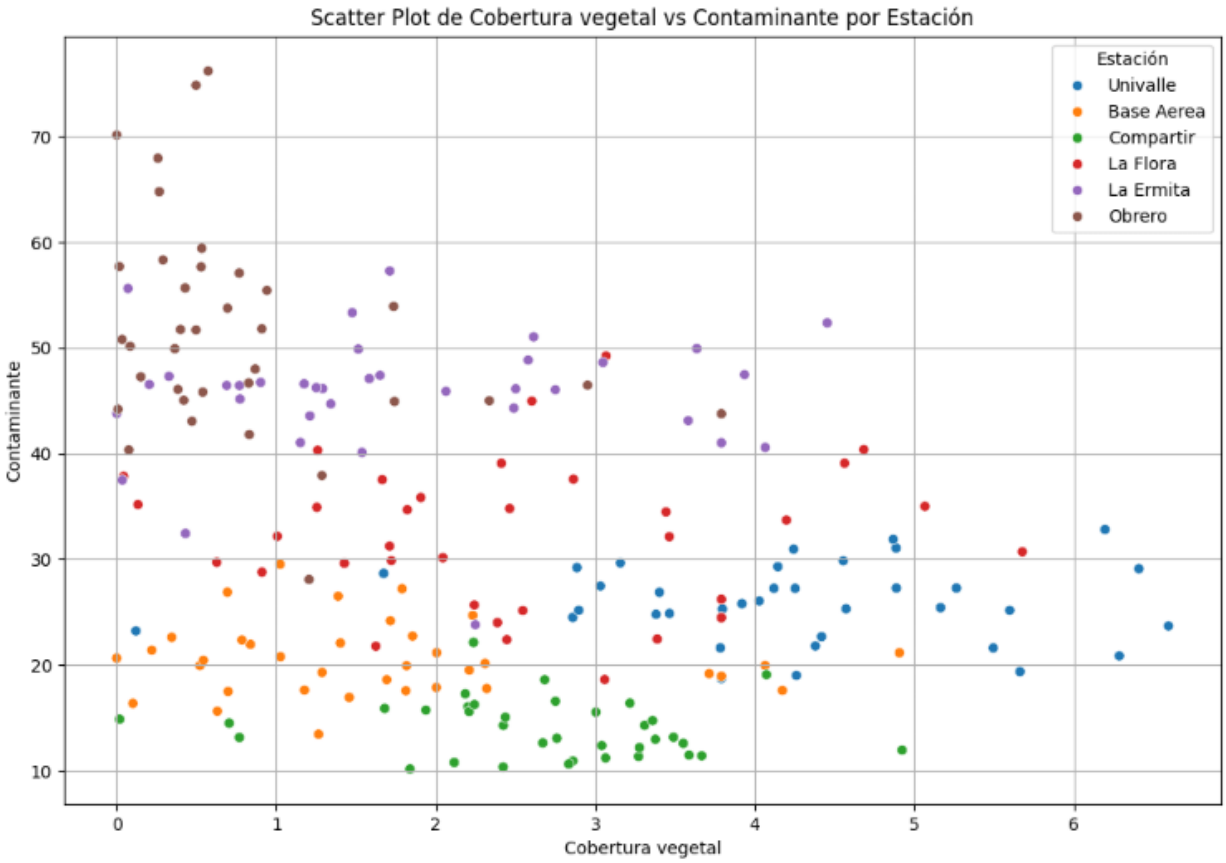
con picos que coinciden con periodos de mayor movilidad vehicular y posibles condiciones meteorológicas adversas para la dispersión de contaminantes.

La reducción sostenida de cobertura vegetal coincide con el incremento progresivo de $PM_{2.5}$, lo que respalda la hipótesis de una correlación negativa entre ambas variables. La pérdida de bosque limita la capacidad de captura de partículas finas, mientras que la variabilidad en césped y pastizales no logra compensar esta disminución. En un contexto de alta movilidad y concentración poblacional como el entorno universitario, la reducción de vegetación agrava la acumulación de contaminantes, incrementando la exposición de estudiantes y residentes a partículas de alto riesgo para la salud.

6.6 CORRELACIÓN ENTRE COBERTURA VEGETAL Y CALIDAD DEL AIRE.

El análisis de correlación se realizó con el propósito de validar el tercer objetivo específico del proyecto, orientado a determinar la relación entre la cobertura vegetal y los niveles de contaminación atmosférica en las estaciones seleccionadas. Para este fin, se empleó el coeficiente de correlación de Pearson debido a que las variables analizadas corresponden a mediciones continuas y presentan comportamiento aproximadamente lineal luego del preprocesamiento y estandarización temporal. La estimación incluyó intervalos de confianza al 95 % y pruebas de significancia estadística (p-value), con el fin de garantizar que los resultados obtenidos no fueran atribuibles al azar y que la correlación observada fuese robusta desde el punto de vista inferencial.

Figura 20 Correlación de cobertura vegetal y contaminantes PM_{10} y $PM_{2.5}$



El gráfico de dispersión muestra la relación entre la cobertura vegetal (eje X, NDVI) y la concentración de contaminantes (eje $Y\ \mu\ g/m^3$) en las estaciones analizadas. Cada punto representa una observación mensual, diferenciada por estación, lo que permite identificar patrones particulares y tendencias generales.

Para cuantificar la relación estadística entre la infraestructura verde y los contaminantes atmosféricos, no se empleó una medida de magnitud física directa, sino el Coeficiente de Correlación de Pearson (r). Este estimador adimensional (que oscila entre -1 y +1) permite determinar la fuerza y dirección de la asociación lineal, independientemente de las escalas de medida de las variables originales (NDVI vs. $\mu\ g/m^3$).

El coeficiente obtenido de $r = -0.37$ ($p < 0.05$) indica una correlación negativa moderada. Esto se interpreta estadísticamente no como una tasa de reducción fija, sino como una tendencia inversa sistemática: los incrementos en los valores del índice de vegetación (NDVI) están

asociados consistentemente con decrementos en las concentraciones de material particulado en la serie temporal analizada. De esta manera, se observa el siguiente comportamiento en las áreas de estudio:

- Estaciones céntricas (Obrero, La Ermita): concentran valores bajos de cobertura vegetal y niveles relativamente altos de contaminantes, lo que refleja la presión urbana y la escasez de espacios verdes en el centro de la ciudad.
- Estaciones periféricas (Compartir, Base Aérea, Univalle): presentan mayor variabilidad, con coberturas vegetales más amplias pero en descenso, y contaminantes que muestran incrementos progresivos.
- La Flora: se ubica en un punto intermedio, con una cobertura arbórea más estable que amortigua parcialmente los niveles de contaminantes, aunque también evidencia la correlación negativa.

En conjunto, el gráfico confirma que la cobertura vegetal actúa como un factor modulador de la calidad del aire: las estaciones con mayor vegetación tienden a registrar menores concentraciones de contaminantes, mientras que aquellas con menor cobertura presentan mayores niveles de PM_{10} y $PM_{2.5}$. Esta relación refuerza la importancia de integrar la conservación y expansión de áreas verdes en la gestión de la calidad del aire en Cali.

El análisis estadístico entre la cobertura vegetal y los niveles de contaminantes atmosféricos (PM_{10} y $PM_{2.5}$) arrojó un coeficiente de correlación de –0.37. Este valor indica una relación negativa moderada, lo que significa que, en general, a medida que disminuye la cobertura vegetal en las áreas de influencia de las estaciones, se observa un incremento en las concentraciones de material particulado. Aunque la magnitud de la correlación no es fuerte, sí resulta significativa desde el punto de vista ambiental, pues confirma la hipótesis de que la pérdida de vegetación reduce la capacidad de mitigación natural frente a contaminantes. Este hallazgo refuerza la importancia de conservar y ampliar las áreas verdes urbanas como estrategia complementaria a las políticas de control de emisiones, dado que la vegetación actúa como un modulador de la calidad del aire en contextos urbanos densamente poblados.

Más allá del valor numérico del coeficiente de correlación, fue necesario evaluar la solidez estadística de esta relación y su coherencia con los modelos de aprendizaje automático desarrollados en el proyecto. Para garantizar la solidez estadística de los resultados obtenidos, la correlación entre cobertura vegetal y niveles de contaminación fue evaluada mediante el coeficiente de Pearson, midiendo la fuerza y dirección de la relación lineal entre las variables. Esta estimación fue complementada con intervalos de confianza al 95 % y pruebas de significancia (p-value), lo que permite determinar si la correlación observada es estadísticamente distinta de cero y reducir el riesgo de interpretaciones espurias. Adicionalmente, se compararon los valores observados de PM_{10} y $PM_{2.5}$ con las predicciones generadas por los modelos de aprendizaje

automático, obteniendo una correlación más elevada entre los valores predichos y observados, lo que refuerza la coherencia entre las tendencias estadísticas y los patrones modelados computacionalmente. Este enfoque combinado permite no solo cuantificar la fuerza de la relación entre cobertura vegetal y contaminantes atmosféricos, sino también evaluar su estabilidad espacial y temporal.

6.7 RECOMENDACIONES PRÁCTICAS

- Reforzar la infraestructura verde como estrategia de salud pública

Los resultados obtenidos muestran una correlación negativa entre la cobertura vegetal y los niveles de material particulado (PM_{10} y $PM_{2.5}$), lo que confirma que la vegetación urbana no solo cumple una función estética o paisajística, sino que actúa como un sistema natural de filtración y regulación de la calidad del aire. En este sentido, se recomienda priorizar la conservación, restauración y expansión de la infraestructura verde en zonas críticas como el centro (Obrero, La Ermita) y el oriente (Compartir), donde la pérdida de vegetación coincide con los niveles más altos de contaminación. Esto implica no solo sembrar árboles, sino diseñar corredores ecológicos urbanos, ampliar zonas verdes en barrios densamente poblados, y proteger remanentes de vegetación existentes. Estas acciones deben ser concebidas como intervenciones de salud pública, capaces de reducir la exposición a contaminantes y mejorar la calidad de vida de la población urbana.

- Integrar criterios ambientales en el ordenamiento territorial

La planeación urbana de Cali debe evolucionar hacia un enfoque que incorpore explícitamente variables ambientales como la cobertura vegetal y la calidad del aire en los procesos de ordenamiento territorial y expansión urbana. En sectores como Base Aérea y Univalle, donde se evidencia una pérdida sostenida de vegetación y un aumento progresivo de contaminantes, es urgente establecer zonas de amortiguación ecológica que limiten la expansión urbana descontrolada. Esto implica condicionar la aprobación de nuevos proyectos urbanísticos a la compensación efectiva de áreas verdes, establecer techos de densificación en zonas ambientalmente frágiles, y exigir estudios de impacto ambiental que incluyan indicadores de carga atmosférica y conectividad ecológica. La vegetación no debe ser vista como un residuo del diseño urbano, sino como un componente estructural del territorio que debe ser planificado, monitoreado y protegido

- Promover movilidad sostenible en zonas de alta exposición

En estaciones como La Flora y La Ermita, donde la vegetación es limitada y el tráfico vehicular es intenso, se hace evidente la necesidad de reducir las emisiones en origen mediante estrategias de movilidad sostenible. Se recomienda ampliar la red de ciclorrutas seguras y sombreadas, fomentar el uso de transporte público eléctrico o de bajas emisiones, y establecer zonas de bajas emisiones en sectores con alta densidad peatonal y bajos niveles de cobertura vegetal. Además, se pueden implementar corredores verdes de movilidad, que integren vegetación lineal con infraestructura de transporte no motorizado, generando beneficios simultáneos en calidad del aire, confort térmico y conectividad urbana. Estas medidas no solo reducen la emisión de contaminantes, sino que también redistribuyen el espacio público en favor de modos de transporte más saludables y equitativos.

- Incorporar monitoreo ambiental en la gestión urbana

La toma de decisiones urbanas debe estar respaldada por sistemas de monitoreo ambiental integrados y accesibles, que permitan evaluar en tiempo real la evolución de la cobertura vegetal y los niveles de contaminación. Se recomienda fortalecer el Sistema de Vigilancia de Calidad del Aire de Santiago de Cali (SVCASC) mediante la incorporación de indicadores de vegetación urbana, como NDVI o superficie verde per cápita, y vincular estos datos a plataformas de planificación como el POT, los planes de desarrollo local y los presupuestos participativos. Además, se sugiere desarrollar paneles de control georreferenciados que permitan a los tomadores de decisión y a la ciudadanía identificar zonas críticas, priorizar intervenciones y hacer seguimiento a los compromisos ambientales. La transparencia y la trazabilidad de estos datos son clave para fomentar una gobernanza ambiental participativa y basada en evidencia.

- Priorizar intervenciones en zonas de vulnerabilidad ambiental y social

Los resultados muestran que la pérdida de vegetación y el aumento de contaminantes afectan con mayor intensidad a sectores con alta densidad poblacional, menor infraestructura verde y mayores niveles de vulnerabilidad social. Por ello, se recomienda focalizar las intervenciones en barrios del oriente y centro de Cali, donde confluyen condiciones de inequidad ambiental y sanitaria. Estas intervenciones deben combinar acciones de restauración ecológica (como siembra de árboles nativos y recuperación de quebradas urbanas), programas de educación ambiental comunitaria, y mecanismos de participación ciudadana en el diseño y mantenimiento de espacios verdes. Además, se sugiere integrar estas acciones con políticas de salud pública, seguridad alimentaria (huertas urbanas) y empleo verde, de modo que la infraestructura ecológica urbana se convierta en un motor de transformación social y resiliencia territorial.

7. CONCLUSIONES Y TRABAJOS FUTUROS

7.1 CONCLUSIONES

- Los resultados obtenidos a lo largo del proyecto permiten validar el objetivo general y los específicos, y derivar las siguientes conclusiones estructurales:
- Validación de la Base de Datos y Metodología ETL: Se logró construir una base de datos robusta. El desarrollo de un *Data Pipeline* adaptado a entornos tropicales, integrando información de PlanetScope con los registros del SVCASC, superó las limitaciones de *data gaps* y nubosidad. La reconstrucción de datos faltantes mediante Random Forest Regressorngarantizó la continuidad estocástica de las series, lo cual fue indispensable para el análisis.
- Superioridad del Modelo de Clasificación: La aplicación de algoritmos de *machine learning* fue altamente efectiva. El modelo Random Forest fue seleccionado como el clasificador óptimo de coberturas, alcanzando una Exactitud Global del 83.33%. Esto valida la estrategia de ingeniería de características, confirmando que la combinación de la reflectancia espectral (NIR, NDVI) con métricas de textura Gabor es fundamental para diferenciar tipos de vegetación en entornos urbanos heterogéneos.
- Confirmación de la Correlación Negativa: El análisis estadístico bivariado arrojó un coeficiente de correlación de Pearson de -0.37 entre la densidad de cobertura vegetal y las concentraciones de material particulado ($PM_{\{10\}}$ y $PM_{\{2.5\}}$). El signo negativo confirma la hipótesis central: a menor cobertura vegetal, mayores concentraciones de contaminantes. Esto valida que la vegetación urbana actúa como un modulador activo de la calidad del aire y como un servicio ecosistémico mensurable, aunque su efecto es de magnitud moderada en un sistema abierto.
- Disparidad Urbana y Saturación del Servicio: Los resultados desagregados por estación revelaron una profunda desigualdad ambiental. Las zonas céntricas (Obrero, La Ermita) sufren una doble penalización: registran las tasas más aceleradas de pérdida de vegetación y mantienen los niveles más altos de PM_{10}. En estos "cañones urbanos", el efecto mitigador de la vegetación está saturado por la magnitud de las emisiones antropogénicas y la falta de dispersión.
- Aporte a la Planificación Territorial: El proyecto proporcionó evidencia empírica local de que la pérdida progresiva de vegetación, observada en todas las estaciones entre 2017 y 2020, compromete la resiliencia urbana. Esto refuerza la necesidad de integrar la conservación y expansión estratégica de la infraestructura verde en el Plan de

Ordenamiento Territorial (POT) de Cali, priorizando intervenciones en zonas de alta vulnerabilidad ambiental y social (ej. Compartir).

7.2 TRABAJOS FUTUROS

Para profundizar la línea de investigación y maximizar su impacto en la gestión urbana, se proponen los siguientes trabajos futuros:

- **Migración a Modelos de *Deep Learning***: Explorar arquitecturas de redes neuronales convolucionales (U-Net) para la clasificación semántica de coberturas. Esto permitiría aprovechar la alta resolución espacial de PlanetScope al máximo, mejorando la precisión en la identificación de elementos micro-urbanos (ej. setos, jardines verticales) y superando las limitaciones de los modelos de *Machine Learning* tradicionales en el manejo de grandes volúmenes de datos.
- **Modelado de Escenarios Predictivos**: Desarrollar modelos de regresión avanzados que permitan simular el impacto de diferentes escenarios de planificación (ej. aumento del 20% de cobertura arbórea en Obrero; implementación de zonas de bajas emisiones) sobre la concentración futura de contaminantes. Estos modelos predictivos pueden integrarse directamente en sistemas de apoyo a la decisión para el DAGMA y la CVC.
- **Integración con Datos Epidemiológicos**: Cruzar la serie espacial de concentraciones de contaminantes (estimada a partir de la vegetación) con datos de salud pública (ej. tasas de hospitalización por IRA o ECV) para cuantificar la carga de enfermedad asociada a la falta de infraestructura verde. Esto fortalecería la justificación de la IVU como una inversión de salud pública preventiva.

8. REFERENCIAS BIBLIOGRÁFICAS

[1] L. A. C. M. y. M. F. O. J. E. G. Montero, «Análisis de la variación temporal y espacial del tráfico automotor y su relación con la calidad del aire en Santiago de Cali,» *Ingeniería y Competitividad,* vol. 25, nº 1, 2023.

[2] D. Sierra-Porta, «Linking PM10 and PM2.5 Pollution Concentration Through Tree Coverage in Urban Areas,» *CLEAN–Soil Air Water,* 2023.

[3] X. Querol, «Air quality in cities: a global challenge,» *Naturgy Foundation,* 2018.

[4] M. d. m. a. y. d. sostenible, «https://www.minambiente.gov.co/wp-content/uploads/2021/10/Resolucion-2254-de-2017.pdf,» 2017. [En línea]. [Último acceso: 4 12 2024].

[5] G. d. C. d. A. d. DAGMA, «Sistema de Vigilancia de Calidad del Aire de Cali - SVCAC,» 2023. [En línea]. Available: https://www.cali.gov.co/dagma/publicaciones/38365/sistema_de_vigilancia_de_calidad_del_aire_de_cali_svcac/. [Último acceso: 4 12 2024].

[6] E. &. H. A. Chuvieco, Fundamentals of Satellite Remote Sensing, CRC Press, 2023.

[7] J. &. Z. H. Li, «Remote sensing as a core data source for urban environmental monitoring: Advances and applications,» *International Journal of Applied Earth Observation and Geoinformation,* p. 126, 2024.

[8] C. W. J. C. &. W. M. A. Gómez, «Optical remote sensing of vegetation: Recent advances, challenges, and future directions,» *Remote Sensing of Environment,* vol. 294, 2024.

[9] C. G. Aguilar, «Aplicación de índices de vegetación derivados de imágenes satelitales Landsat 7 ETM+ y ASTER para la caracterización de la cobertura vegetal en la zona centro de la provincia de Loja, Ecuador,» *Universidad Nacional de La Plata,* nº 10.35537/10915/34487, 2014.

[10] T. E. S. Agency, «PlanetScope - Earth Online,» [En línea]. Available: https://earth.esa.int/eogateway/missions/planetscope. [Último acceso: 15 11 2024].

[11] F. R. L. Llumiquinga, «Procesamiento de imágenes mediante software libre python para el análisis metalográfico en aceros de bajo contenido de carbono,» 2024. [En línea]. Available: https://bibdigital.epn.edu.ec/handle/15000/7171. [Último acceso: 4 12 2024].

[12] Z. &. Z. H. Liu, «A Systematic Review of Satellite Image Classification,» *International Journal of Machine Learning,* vol. 15, nº 3, pp. 51-63, 2025.

[13] A. K. y. P. K. P. Kupidura, «Comparative analysis of the performance of selected machine-learning algorithms in satellite image classification,» *Reports on Geodesy and Geoinformatics,* vol. 118, pp. 53-69, 2024.

[14] Z. S. y. A. J. M. N. Ahmad, «Comparison of Random Forest and XGBoost Classifiers Using Integrated Optical and SAR Features for Mapping Urban Impervious Surface,» *Remote Sensing,* vol. 16, p. 665, 2024.

[15] A. M. P. Rubio, «Teledetección en la agricultura de precisión: estado del arte,» *TECTZAPIC,* vol. 6, nº 2, 2020.

[16] M. L. M. N. R. C. E. S. J. McCarthy, «A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence,» *AI Mag,* vol. 27, pp. 12-14, 1995.

[17] X. L. H. L. S. Y. y. O. K. Y. Jiang, «Quo vadis artificial intelligence?,» *Discov. Artif. Intell,* vol. 2, nº 1, p. 4, 2022.

[18] A. K. y. P. K. P. Kupidura, «Comparative analysis of the performance of selected machine-learning algorithms in satellite image classification,» *Reports on Geodesy and Geoinformatics,* vol. 118, pp. 53-69, 2024.

[19] D. S. Wilks, «Statistical methods in the atmospheric sciences,» *Academic press,* vol. 100, 2011.

[20] K. Pearson, «Note on regression and inheritance in the case of two parents.,» *Proceedings of the Royal Society of London,* nº 58, pp. 240-242, 1895.

[21] F. J. Anscombe, «Graphs in statistical analysis,» *The American Statistician,* vol. 27, nº 1, pp. 17-21, 1973.

[22] J. &. K. T. Hauke, «Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data,» *Quaestiones geographicae,* vol. 30, nº 2, pp. 87-93, 2011.

[23] R. H. &. S. D. S. Shumway, «Time series analysis and its applications: with R examples,» *Springer,* 2017.

[24] K. Pearson, «Note on regression and inheritance in the case of two parents.,» *Proceedings of the Royal Society of London,* pp. 240-242, 1895.

[25] C. Spearman, «The proof and measurement of association between two things,» *International Journal of Epidemiology,* pp. 1137-1150, 2010.

[26] M. Kendall, «A new measure of rank correlation,» *Biometrika,* vol. 30, pp. 81-93, 1938.

[27] N. M. y. J. F. G. Aziz, «Remote sensing based forest cover classification using RF algorithm,» *Scientific Reports,* 2024.

[28] T. Mwinuka, «Comparison of Random Forest and XGBoost Classifiers Using Integrated Optical and SAR Features for Mapping Urban Impervious Surface,» *Remote Sensing,* 2024.

[29] T. X. W. &. F. J. Adugna, «Comparison of Random Forest and Support Vector Machine Classifiers for Regional Land Cover Mapping Using Coarse Resolution FY-3C Images,» *Remote Sensing,* vol. 14, nº 3, 2022.

[30] F. &. F. M. Ramdani, «The simplicity of XGBoost algorithm versus the complexity of Random Forest,» *F1000Research,* 2022.

[31] L. Breiman, «Random Forests,» *Machine Learning,* vol. 45, nº 1, pp. 5-32, 2001.

[32] T. &. G. C. Chen, « XGBoost: A Scalable Tree Boosting System,» *KDD '16 Proceedings,* 2016.

[33] G. I. J. &. O. C. Mountrakis, «Support vector machines in remote sensing: A review,» *ISPRS Journal of Photogrammetry,* 2011.

[34] K. S. M. &. W. H. Hornik, «Multilayer feedforward networks are universal approximators,» *Neural Networks,* 1989.

[35] F. &. K. W. Minaei, «Remote sensing tree classification with a multilayer perceptron,» *ResearchGate / Remote Sensing.,* 2019.

[36] P. K. e. a. Mishra, «Use of Logistic Regression in Land-Cover Classification with Moderate-Resolution Multispectral Data,» *Journal of the Indian Society of Remote Sensing,* 2019.

[37] P. L. PBC, «Planet Imagery Product Specifications: Dove-C, Dove-R and SuperDove Sensors,» San Francisco, CA, 2025. [En línea]. Available: https://assets.planet.com/docs/Planet_Combined_Imagery_Product_Specs_letter_screen.pdf. [Último acceso: 11 2025].

[38] D. A. d. G. d. M. A. (DAGMA), «Informe del Estado de la Calidad del Aire en Santiago de Cali: Año 2022,» *Alcaldía de Santiago de Cali, Cali, Colombia, Tech. Rep,* 2022.

[39] P. A. F. M. G. Z. a. W. F. T. X. Cui, «Influence of urban tree density on PM2.5 capture capability: A case study in distinct urban morphologies,» *Urban Forestry & Urban Greening,* vol. 84, p. 127934, 2023.

[40] S. Janhäll, «Review on urban vegetation and particle air pollution – Deposition and dispersion,» *Atmospheric Environment,* vol. 105, pp. 130-137, 2015.

[41] D. M.-F. e. al., «Trace metal bioaccessibility and inhalation risk assessment of road dust in an industrial city,» *Atmosphere,* vol. 11, nº 8, p. 821, 2020.

[42] H. T. I. S. y. I. L. K. A. Hogda, «Mapping of air pollution effects on the vegetation cover in the Kirkenes-Nikel area using remote sensing,» *Firenze: IGARSS,* pp. 1249-1251, 1995.

[43] S. R. y. A. A. S. Salata, «Mapping air filtering in urban areas. A Land Use Regression model for Ecosystem Services assessment in planning,» *Ecosyst. Serv,* vol. 28, pp. 341-350, 2017.

[44] W. E. W. e. C. V, «Detecting air pollution stress in southern California vegetation using Landsat Thematic Mapper band data. Photogrammetric Engineering and Remote Sensing,» *Photogramm. Eng. Remote Sens,* vol. 54, nº 9, pp. 1305-1311, 2020.

[45] K. Mazirh, «Using Remote Sensing to Address Green Heritage of the City of Marrakech, Morocco,» *Canadian Journal of Remote Sensing,* vol. 43, nº 1, 2023.

[46] E. V. S. y. Y. T. S. Correa, «Spatiotemporal Analysis of Variables Affecting Air Quality in Urban Areas of the City of Cartagena, Colombia,» 18th National Meeting on Optics and the 9th Andean and Caribbean Conference on Optics and its Applications, ENO-CANCOA 2024 - Conference Proceedings, Cartagena, 2024.

[47] L. T. C. y. J. D. Paz, «Análisis de la contaminación ambiental usando técnicas de teledetección y análisis de componentes principales,» *TecnoLógicas,* vol. 24, nº 50, 2021.

[48] D. S.-C. Y. T. T.-A. M. &. N. d. V. L. A. Sierra-Porta, «Linking PM10 and PM2.5 Pollution Concentration through Tree Coverage

in Urban Areas,» *CLEAN–Soil, Air, Water,* 2023.

[49] X. Z. Y. &. C. H. Li, «Spatiotemporal association between NDVI dynamics and $PM_{2.5}$ concentrations in Beijing using Landsat and MERRA-2 datasets,» *Remote Sensing of Environment,* p. 295, 2023.

[50] C. S. M. &. O. J. d. Keijzer, «Urban green infrastructure and its association with PM_{10} and $PM_{2.5}$ across European cities using Sentinel-2 and AirBase data,» *Environmental Research Letters,* 2024.

[51] D. A. d. G. d. M. A. (DAGMA), «Informe del Estado de la Calidad del Aire en Santiago de Cali 2023,» *Alcaldía de Santiago de Cali, Cali, Colombia, Tech. Rep,* 2023.

[52] e. a. L. Díaz-González, «Handling Missing Air Quality Data Using Bidirectional Recurrent Imputation for Time Series and Random Forest: A Case Study in Mexico City,» *AI,* vol. 6, nº 9, pp. 208-225, 2025.

[53] C. W. L. F. K. M. G. S. C. Betancourt, «Graph Machine Learning for Improved Imputation of Missing Tropospheric Ozone Data,» *Environmental Science & Technology,* vol. 57, nº 46, pp. 18231-18241, 2023.

[54] Q. Vanhellemont, «Evaluation of eight band SuperDove imagery for aquatic and urban applications,» *Optics Express,* vol. 31, nº 9, pp. 13837-13854, 2023.

[55] P. L. PBC, «Planet Imagery Product Specifications: PlanetScope & SuperDove,» *San Francisco, CA, USA, Tech,* 2024.

[56] E. E. Online, «PlanetScope Mission Description,» European Space Agency, 2024. [En línea]. Available: https://earth.esa.int/eogateway/missions/planetscope. [Último acceso: 11 2025].

[57] S. Aslan, «Atmospheric Correction of Satellite images using Python,» Nerd For Tech, 05 2021. [En línea]. Available: https://medium.com/nerd-for-tech. [Último acceso: 11 2025].

[58] P. L. PBC, «Planet Python Client Documentation,» Planet Developers, 2024. [En línea]. Available: https://github.com/planetlabs/planet-client-python. [Último acceso: 11 2024].

[59] D. A. d. G. d. M. A. (DAGMA), «Boletín Técnico: Sistema de Vigilancia de Calidad del Aire de Santiago de Cali,» Datos Abiertos Alcaldía de Cali, 2024. [En línea]. Available: https://datos.cali.gov.co. [Último acceso: 11 2025].

[60] W. H. Organization, «WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide,» *Geneva, Switzerland,* 2021.

[61] W. H. Organization, «WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide,» *Geneva: World Health Organization,* 2021.

[62] U. G. S. (USGS), «Landsat Surface Reflectance-Derived Spectral Indices,» 2024. [En línea]. Available: https://www.usgs.gov/landsat-missions/landsat-surface-reflectance-derived-spectral-indices. [Último acceso: 11 2025].

[63] e. a. A. Kura, «Performance of Vegetation Indices in Urban Environments,» *Remote Sensing Applications: Society and Environment,* vol. 26, nº 100728, 2022.

[64] S.-i. D. Team, «Module: filters.gabor - Image processing in Python,» Scikit-image Documentation 0.22.0, 2024, 2024. [En línea]. Available: https://scikit-image.org.. [Último acceso: 11 2025].

[65] S. Inc, «Streamlit Documentation: Building Data Apps in Python,» Streamlit Documentation: Building Data Apps in Python, 2024. [En línea]. Available: https://docs.streamlit.io. [Último acceso: 11 2025].

[66] M. y. E. A. Instituto de Hidrología, «Protocolo para el monitoreo y seguimiento de la calidad del aire: Manual de diseño y operación de sistemas de vigilancia,» *Ministerio de Ambiente y Desarrollo Sostenible,* 2019.

[67] R. G. e. a. Mantovani, «A meta-learning recommender system for hyperparameter tuning,» *Information Sciences,* 2019.

[68] M. &. D. L. Belgiu, «Random forest in remote sensing: A review of applications and future directions,» *ISPRS Journal of Photogrammetry,* 2016.

[69] S. e. a. Georganos, «Less is more: Optimizing classification performance of machine learning algorithms in very high resolution urban remote sensing,» *GIScience & Remote Sensing,* 2018.

[70] V. F. Rodriguez-Galiano, «An assessment of the effectiveness of a random forest classifier for land cover classification,» *ISPRS Journal,* 2012.

[71] P. Kupidura, «The comparison of different methods of texture analysis for their efficacy for land use classification in urban areas,» *Remote Sensing.,* 2019.

[72] P. e. a. Olofsson, «Good practices for estimating area and assessing accuracy of land change,» *Remote Sensing of Environment,* 2014.

[73] J. F. M.-E. e. al., «Intra-urban variability of long-term exposure to PM2.5 and NO2 in five cities in Colombia,» *Environmental Science and Pollution Research,* vol. 31, 2024.

[74] S. H. A. D. M. M. a. J. P. D. J. Nowak, «Air pollution removal by urban forests in Canada and its effect on public health and economics,» *Urban Forestry & Urban Greening,* vol. 29, pp. 168-176, 2018.

[75] K. B.-B. a. M. Bogacki, «Street Canyon Vegetation—Impact on the Dispersion of Air Pollutant Emissions from Road Traffic,» *Sustainability,* vol. 16, nº 23, 2024.

[76] K. V. A. e. al, «Air pollution abatement performances of green infrastructure in open and built-up street canyon environments: A review,» *Atmospheric Environment,* vol. 162, pp. 71-86, 2017.

[77] L. G. a. R. P. L. Zhang, «Scavenging of atmospheric aerosols by rain: A review of parameterizations and applications,» *Bulletin of the American Meteorological Society,* vol. 101, nº 3, pp. 203-220, 2020.

[78] N. R.-H. e. al., «An Investigation of the Precipitation Net Effect on the Particulate Matter Concentration in a Narrow Valley: Role of Lower Troposphere Stability,» *Journal of Applied Meteorology and Climatology,* vol. 3, p. 59, 2020.

[79] A. A. e. al, «Environmental justice beyond race: Skin tone and exposure to air pollution in Colombia,» *Proceedings of the National Academy of Sciences (PNAS),* vol. 121, nº 7, 2024.

[80] M. S. e. al, «Estimation of high spatial resolution air pollution concentrations using random forest algorithms,» *Environment International,* vol. 124, pp. 172-182, 2019.

[81] C. B. a. L. A. S. P. Schober, «Correlation coefficients: appropriate use and interpretation,» *Anesthesia & Analgesia,* vol. 126, nº 5, pp. 1763-1768, 2018.

[82] P. H. C. &. R.-A. C. Sarricolea, «Urban expansion, vegetation loss and particulate matter trends using machine-learning-based land-cover classification in Santiago de Chile,» *Science of the Total Environment,* 2022.

[83] C. e. a. Cruz-Ramos, « Gabor Features Extraction and Land-Cover Classification of Urban Hyperspectral Images,» *Remote Sensing,* 2021.

[84] Y. X. M. S. a. Y. L. Y. Xing, «Spatio-temporal variations of PM2.5 and its relationship with meteorological factors and land use in a typical industrial city,» *Atmosphere,* vol. 11, nº 9, p. 956, 2020.

[85] J. Y. a. H. L. C. Huang, «Spatio-temporal variations of urban green space in response to urban expansion and their impact on urban heat island,» *Scientific Reports,* vol. 11, nº 1, p. 16382, 2021.

[86] D. E. C. a. J. C. S. D. J. Nowak, «Air pollution removal by urban trees and shrubs in the United States,» *Urban Forestry & Urban Greening,* vol. 4, nº 3, pp. 115-123, 2006.