Pontificia Universidad
JAVERIANA
Cali

DETERMINACIÓN DE ISLAS DE CALOR EN LAS CIUDADES DE BARRANQUILLA, CARTAGENA Y SANTA MARTA A PARTIR DE IMÁGENES SATELITALES Y ALGORITMOS DE MACHINE LEARNING

Cristian Camilo Rodriguez Ortiz
Jonny Carlos Sánchez González

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

Director(a)
Yady Tatiana Solano Correa

Codirector(a)
Mario Milver Patiño Velasco

FACULTAD DE INGENIERÍA Y CIENCIAS

MAESTRÍA EN CIENCIA DE DATOS

SANTIAGO DE CALI, DICIEMBRE DEL 2025

Tabla de Contenido

Listado de Figuras.

Listado de Tablas.

Introducción

El fenómeno de las *islas de calor urbanas* constituye uno de los desafíos ambientales más significativos derivados de la urbanización acelerada y el cambio en los usos del suelo. Este fenómeno se manifiesta en un incremento en la temperatura en las zonas urbanas respecto de las áreas rurales, como resultado de la sustitución de superficies naturales por materiales que absorben y retienen mayor cantidad de calor. Dicho efecto impacta directamente la sostenibilidad ambiental, incrementa el consumo energético y agrava la contaminación atmosférica, con consecuencias en la salud pública y la calidad de vida de los habitantes.

En la región Caribe colombiana, las ciudades de Barranquilla, Cartagena y Santa Marta han experimentado en la última década un proceso sostenido de expansión urbana, impulsado por el crecimiento demográfico, la dinámica económica y el desarrollo del sector turístico. Este proceso ha favorecido la concentración de áreas densamente construidas, reduciendo progresivamente la cobertura vegetal. Ante este escenario, se hace necesario identificar las zonas afectadas por el incremento térmico.

El presente proyecto tuvo como propósito analizar el comportamiento y la evolución espacial de las islas de calor en las tres ciudades mencionadas durante el periodo 2015–2024, integrando tecnologías de *percepción remota*, *sistemas de información geográfica* y *algoritmos de aprendizaje automático*. A partir de imágenes satelitales multiespectrales de la misión *Landsat*, se estimó la *temperatura de la superficie terrestre* e índices espectrales como el índice de vegetación de diferencia normalizada y el índice de áreas construidas normalizadas, variables fundamentales para caracterizar las coberturas superficiales y su relación con la temperatura urbana.

El enfoque metodológico implementado se sustenta en el uso de herramientas de *código abierto* como *QGIS* y su módulo *PyQGIS*, junto con *PostgreSQL/PostGIS*, que permitieron automatizar las etapas de procesamiento, análisis y modelamiento de datos geoespaciales. Esta integración tecnológica favorece la reproducibilidad del estudio y la escalabilidad de los resultados hacia otros contextos urbanos. Asimismo, los modelos de aprendizaje supervisado empleados posibilitaron la clasificación de coberturas y la estimación de la temperatura del aire con base en variables derivadas de la teledetección.Este trabajo no plantea acciones de mitigación directa, sino que busca generar información geoespacial y analítica que sirva como insumo técnico para las entidades territoriales, en el marco de la planificación urbana sostenible y la gestión climática local.

1. Definición del problema.

1.1. Planteamiento del problema.

Las islas de calor son un fenómeno que incrementa la temperatura en áreas urbanas y centros poblados, debido principalmente a la intervención humana. El principal factor que desencadena este fenómeno es la urbanización, que incluye la construcción de viviendas, edificios y la pavimentación de vías. Estudios basados en datos satelitales han demostrado que la expansión urbana y el aumento de la densidad constructiva, acompañados de la disminución de áreas naturales ecosistémicas y el cambio climático, contribuyen al incremento de las islas de calor. Este fenómeno se caracteriza por un aumento significativo en la temperatura de la superficie terrestre (LST) en las zonas urbanas en comparación con las áreas rurales circundantes [2].

Las islas de calor urbanas generan múltiples efectos adversos en el bienestar humano y el medio ambiente. Este fenómeno está asociado con un incremento en el consumo energético debido a la mayor demanda de refrigeración en infraestructuras urbanas, como lo señala el estudio de Zhou et al. (2018), que analiza las implicaciones del UHI en el aumento de las necesidades energéticas en las ciudades [3]. Además, las temperaturas elevadas, junto con la contaminación atmosférica, incrementan la concentración de ozono y otros contaminantes, agravando la calidad del aire y contribuyendo al aumento de enfermedades respiratorias y cardiovasculares en la población urbana [3].

En la práctica, la medición de las temperaturas urbanas se ha realizado tradicionalmente mediante estaciones meteorológicas locales. Aunque estas estaciones proporcionan datos puntuales y precisos de las condiciones climáticas en los sitios donde están instaladas, su distribución es limitada, ya que muchas ciudades cuentan con un número limitado de estaciones. Esta baja densidad de estaciones restringe significativamente su capacidad para capturar las variaciones térmicas en toda la ciudad, lo que resulta en una representación incompleta y sesgada de los patrones térmicos. Esto se convierte en una limitación crítica para analizar fenómenos como las islas de calor, que requieren una visión más amplia y detallada de las diferencias térmicas en áreas urbanas y rurales.

Los avances en sensores térmicos satelitales han permitido obtener datos más detallados y continuos sobre la temperatura de la superficie terrestre (LST) a nivel territorial. Este cálculo se realiza a partir de la radiancia emitida por la superficie terrestre, utilizando algoritmos como la Ley de Planck para derivar la temperatura de brillo y ajustando la emisividad según la cobertura terrestre. Sin embargo,

aunque los sensores satelitales proporcionan información más precisa, la LST por sí sola no es suficiente para identificar las islas de calor, ya que estas dependen no solo de los valores absolutos de temperatura, sino también del contraste térmico entre áreas urbanas y rurales, además de otros factores como índices espectrales (NDVI y NDBI), el uso del suelo y los patrones urbanos [4].

El análisis de este fenómeno requiere procesar grandes volúmenes de datos espaciales y espectrales que integren múltiples variables, lo cual representa un desafío significativo. En este contexto, el machine learning se presenta como una herramienta clave, ya que permite identificar patrones complejos en los datos y modelar relaciones entre la configuración urbana y las temperaturas superficiales. Estas técnicas facilitan el análisis multitemporal, ayudando a comprender cómo las dinámicas urbanas y climáticas han contribuido al fenómeno de las islas de calor a lo largo del tiempo [5].

Durante la última década, las ciudades de Barranquilla, Cartagena y Santa Marta han experimentado un rápido proceso de expansión urbana, impulsado por su crecimiento poblacional, desarrollo económico, atractivo turístico y ubicación estratégica en el mar Caribe [6]. Este proceso ha transformado significativamente el paisaje urbano de estas ciudades, reemplazando áreas naturales por infraestructuras asociadas a viviendas, complejos hoteleros, servicios y vías de comunicación. Como resultado, se ha intensificado el fenómeno de las islas de calor en estas ciudades. Esto hace necesario un análisis más profundo para comprender cómo se han propagado y cómo interactúan las dinámicas urbanas, climáticas y térmicas en estas regiones.

1.2. Formulación del problema.

¿Es posible determinar las islas de calor en las ciudades de Barranquilla, Cartagena y Santa Marta durante la última década mediante el uso de imágenes satelitales y algoritmos de machine learning?

1.2.1. Sistematización.

- ¿Existen imágenes satelitales disponibles de la última década para la identificación de islas de calor en las ciudades Barranquilla, Cartagena y Santa Marta?
- ¿Es posible identificar las islas de calor con algoritmos de machine learning e índices espectrales?
- ¿Cuál es el nivel de precisión de las islas de calor determinadas con algoritmos de machine learning respecto de datos de referencia como estaciones meteorológicas?
- ¿Es posible identificar cambios en la distribución de las islas de calor durante la última década?

2. Objetivos.

2.1. Objetivo general.

Determinar las islas de calor en las ciudades de Barranquilla, Cartagena y Santa Marta durante la última década, utilizando imágenes satelitales y algoritmos de machine learning supervisado.

2.2. Objetivos específicos.

- Crear una base de datos de imágenes satelitales de la última década para la identificación de las islas de calor en las ciudades de Barranquilla, Cartagena y Santa Marta.
- Identificar las islas de calor mediante algoritmos de machine learning supervisado e índices espectrales.
- Evaluar la precisión de los resultados obtenidos mediante la comparación con datos de estaciones meteorológicas.
- Realizar un análisis multitemporal para identificar patrones de cambio en la distribución de las islas de calor durante la última década.

3. Marco de referencia y Antecedentes.

3.1. Marco teórico.

3.1.1. Fenómeno térmico urbano y conceptos físicos.

Islas de calor urbanas - ICU (Urban Heat Island - UHI): Según la Agencia de Protección Ambiental de los Estados Unidos (EPA), las islas de calor se forman en áreas urbanas debido a la acumulación de estructuras como edificios, carreteras y otras infraestructuras que absorben y reemiten el calor solar con mayor eficiencia que los paisajes naturales, como los bosques y cuerpos de agua. Esto provoca que las zonas urbanas se conviertan en "islas" de temperatura más elevada en comparación con las áreas circundantes. Las islas de calor pueden surgir en diferentes condiciones climáticas y geográficas, tanto de día como de noche, y su formación no depende de la temporada ni del tamaño de la ciudad en la que se encuentran [7].

Temperatura de la superficie terrestre (Land Surface Temperature - LST): representa la temperatura de la superficie terrestre derivada de mediciones satelitales, específicamente a través de datos infrarrojos térmicos (TIR). Su importancia radica en su capacidad para proporcionar información sobre las variaciones temporales y espaciales del estado de equilibrio de la superficie, siendo fundamental en aplicaciones como el monitoreo de la vegetación, el ciclo hidrológico, el cambio climático y estudios urbanos y ambientales [8].

Emisividad de la Superficie (Land Surface Emissivity - LSE): La emisividad de la superficie terrestre es una propiedad radiativa intrínseca de los materiales que modula la emisión en el infrarrojo térmico y condiciona la estimación de la temperatura superficial a partir de sensores remotos. En términos generales, la LSE puede derivarse de la radiancia emitida medida desde el espacio y actúa como parámetro clave para caracterizar composición y cobertura del terreno en el dominio TIR [9]. Su magnitud es adimensional con valores típicos entre 0 y 1, donde superficies acuáticas y zonas con vegetación presentan valores elevados y zonas urbanizadas o suelo desnudo muestran valores relativamente menores.

3.1.2. Percepción remota y sensores.

Percepción remota: Es la técnica de adquisición de datos de la superficie terrestre a través de sensores instalados en plataformas espaciales. Esta técnica se basa en la interacción electromagnética entre la Tierra y el sensor, que puede capturar la energía reflejada de la luz solar, la emisión propia de la superficie o un haz energéti-

co artificial, generando datos que luego son procesados para obtener información interpretable sobre el terreno. Según Chuvieco [10], la percepción remota no solo implica la adquisición de imágenes, sino también su posterior almacenamiento y procesamiento, ya sea a bordo del satélite o en estaciones receptoras en tierra, para permitir su interpretación en aplicaciones específicas.

Los sensores utilizados en percepción remota se clasifican en pasivos y activos. Los sensores pasivos dependen de fuentes externas de energía, como la luz solar, para captar la radiación reflejada o emitida por la superficie terrestre. En contraste, los sensores activos generan su propia fuente de energía, como un haz de radar o láser, permitiendo la adquisición de datos independientemente de las condiciones de iluminación o climáticas. En el contexto del proyecto de islas de calor, los sistemas satelitales como Landsat y Sentinel están equipados con sensores térmicos que facilitan la captura de datos relacionados con la temperatura superficial, permitiendo identificar y analizar variaciones térmicas en áreas urbanas y rurales.

Sensor pasivo: Según Chuvieco [10], los sensores pasivos "se limitan a recoger la energía electromagnética procedente de las cubiertas terrestres, ya sea ésta reflejada de los rayos solares, ya emitida en virtud de su propia temperatura". Estos dispositivos incluyen cámaras fotográficas, exploradores de barrido y radiómetros de microondas, diseñados para captar información en función de la radiación recibida.

Los sensores pasivos multiespectrales destacan por su capacidad de registrar datos en diferentes regiones del espectro electromagnético, desde el visible hasta el infrarrojo térmico. Este último es particularmente relevante para medir temperaturas superficiales con alta precisión, proporcionando datos esenciales para el análisis de las dinámicas térmicas en el entorno terrestre. Además, satélites como Landsat y MODIS se caracterizan por ofrecer información pública y de calidad, con bandas específicas para el análisis térmico y multiespectral, lo que facilita el acceso a insumos confiables para investigaciones científicas y aplicaciones prácticas a diversas escalas.

Infrarrojo térmico: Según Chuvieco [10], se refiere a la parte del espectro electromagnético que permite detectar el calor emitido por la superficie terrestre y sus distintas coberturas. Este rango se encuentra entre 8 y 14 micrómetros, donde se manifiesta con mayor claridad la emitancia espectral de la superficie en función de su temperatura, aproximadamente 300 K. A diferencia de otros tipos de radiación, la energía captada en el infrarrojo térmico no es producto de la reflexión de la luz solar, sino de la radiación emitida directamente por la superficie terrestre.

Fotointerpretación: Es el proceso de reconocer y analizar objetos o fenómenos

en la superficie terrestre utilizando imágenes captadas por satélites. Este método se basa en observar características como colores, formas, texturas y la relación entre los elementos en el terreno. Las imágenes satelitales contienen información de distintas longitudes de onda de luz, lo que permite distinguir materiales como agua, vegetación o construcciones. A través de herramientas digitales o mediante interpretación visual, la fotointerpretación permite clasificar áreas según su uso, identificar cambios en el tiempo y apoyar estudios como el monitoreo ambiental, la planificación urbana y la gestión de riesgos.

3.1.3. Fuentes de datos.

Landsat: Es un sistema satelital de observación terrestre iniciado en 1972 por la NASA y el Servicio Geológico de los Estados Unidos (USGS), que ha sido pionero en la recolección de datos de alta resolución sobre la superficie de la Tierra. Este sistema ha permitido estudiar cambios en el uso del suelo, monitorizar ecosistemas y evaluar los efectos de las actividades humanas y fenómenos naturales en el medio ambiente.

Según Williams, Goward y Arvidson [11], el programa *Landsat* ha sido fundamental para el desarrollo de la percepción remota global, proporcionando un registro continuo y detallado de las condiciones terrestres. A lo largo de las décadas, varios satélites *Landsat* han sido lanzados para asegurar la continuidad de los datos, siendo el más reciente *Landsat 9*, que se lanzó el 27 de septiembre de 2021. Este nuevo satélite continúa la misión de *Landsat* de monitorear el planeta, asegurando la disponibilidad de datos para futuras investigaciones y aplicaciones.

ERA5 (reanálisis atmosférico): Es la quinta generación de reanálisis del ECMWF; integra observaciones globales heterogéneas mediante asimilación 4D-Var en el IFS para producir estimaciones coherentes y físicamente consistentes del estado de la atmósfera. Proporciona salida horaria para múltiples variables (superficie y niveles de presión/modelo), con resolución horizontal cercana a 31 km, 137 niveles verticales y un conjunto de incertidumbre basado en ensamble [12].

En términos operativos, ERA5 cubre desde 1940 hasta el presente en una malla global de 0.25° y se actualiza casi en tiempo real; su variante ERA5-Land, calculada de forma desacoplada (*offline*) del sistema atmosférico principal, mantiene frecuencia horaria y aumenta el detalle espacial a 0.1°, lo que la hace especialmente útil para aplicaciones terrestres y urbanas [13].

3.1.4. Índices espectrales e indicadores temáticos.

Índices espectrales: son métricas calculadas a partir de la reflectancia registrada en diferentes bandas espectrales por sensores remotos, utilizadas para destacar ciertas características de la superficie terrestre, como la vegetación, el agua y el suelo. Estos índices combinan matemáticamente diferentes bandas para amplificar propiedades específicas de los objetos o superficies en análisis, permitiendo así una evaluación detallada de aspectos ambientales y cambios en la cobertura terrestre.

Índice De Vegetación de Diferencia Normalizada (NDVI): Según el Instituto Nacional de Estadística y Geografía de México (INEGI), lo define como indicador basado en la relación entre la cantidad de luz reflejada por la superficie terrestre en dos bandas del espectro electromagnético: el rojo y el infrarrojo cercano. Valores bajos de reflectancia en la banda roja, combinados con alta reflectancia en el infrarrojo cercano, son indicativos de una mayor actividad fotosintética y, por lo tanto, de una mayor cantidad y densidad de vegetación verde en el área observada [14].

$$\text{NDVI} = \frac{\text{NIR} - \text{RED}}{\text{NIR} + \text{RED}} \tag{1}$$

donde:

- NIR es la reflectancia en el infrarrojo cercano.
- RED es la reflectancia en la banda roja del espectro electromagnético.

Índice de Diferencia Normalizada Edificada (NDBI):Es un índice espectral ampliamente utilizado en percepción remota para detectar y analizar áreas urbanizadas. Este índice aprovecha las diferencias en las propiedades reflectivas de los materiales construidos en comparación con la vegetación y otras coberturas de suelo, destacando las zonas donde predominan superficies impermeables como concreto y asfalto. El cálculo del NDBI se realiza mediante la siguiente fórmula:

$$\text{NDBI} = \frac{\text{SWIR} - \text{NIR}}{\text{SWIR} + \text{NIR}} \tag{2}$$

donde:

- NIR es la reflectancia en el infrarrojo cercano.
- SWIR es la reflectancia en la infrarrojo medio.

Índice de Diferencia Normalizada de Agua (NDWI): Es un índice espectral comúnmente empleado en percepción remota para detectar cuerpos de agua

superficial. Este índice explota las diferencias en las propiedades reflectivas del agua en comparación con la vegetación y el suelo, facilitando la identificación de zonas acuáticas. El cálculo del NDWI se lleva a cabo mediante la siguiente fórmula:

$$\text{NDWI} = \frac{\text{Green} - \text{NIR}}{\text{Green} + \text{NIR}} \tag{3}$$

donde:

- Green es la reflectancia en la banda del verde.
- NIR es la reflectancia en el infrarrojo cercano.

Huella urbana (Global Urban Footprint-GUF) Es una capa raster binaria que representa un mapa global de asentamientos humanos. El GUF proporciona una representación detallada tanto de grandes aglomeraciones urbanas como de pequeñas áreas construidas en regiones rurales, lo que permite estudios globales y comparativos sobre urbanización y patrones de asentamiento [15].

Sistema de monitoreo ambiental: Según Perevochtchikova [16], "el sistema de monitoreo ambiental suele incluir varios subsistemas (o bloques) de medición de diferentes componentes de la naturaleza (como la atmósfera, la biosfera, la hidrosfera y la litosfera). Se evalúan las siguientes características: la calidad del aire, la climatología, la calidad y la cantidad del agua superficial y subterránea y la sedimentación de los cauces, la química y el uso del suelo".

3.1.5. Aprendizaje automático.

Clasificación supervisada: Es un método ampliamente utilizado en el análisis de imágenes satelitales para categorizar los píxeles de una imagen en clases predefinidas, basándose en información previamente conocida. En este enfoque, se seleccionan muestras representativas de cada clase de interés, las cuales son identificadas tanto en la imagen como en el terreno real. Estas muestras, denominadas píxeles muestrales, sirven como referencia para entrenar el modelo de clasificación.

El proceso utiliza algoritmos estadísticos, como análisis de probabilidad o distancias espectrales, para asignar a cada píxel de la imagen una etiqueta correspondiente a una clase específica. Según Arozarena [17], este procedimiento "etiqueta todos los píxeles con criterios estadísticos [...] agrupándolos en las diversas clases definitivas". Esto asegura que las categorías finales reflejen de manera precisa las condiciones del terreno o los objetos reales. Además, una característica clave de la clasificación supervisada es su capacidad para interpretar información compleja en aplicaciones

específicas, como el análisis de coberturas terrestres, el monitoreo ambiental y los estudios de uso del suelo.

Regresión Lineal: Es el modelo más simple y ampliamente utilizados en el aprendizaje supervisado. Su principio fundamental consiste en modelar la relación entre una variable dependiente y y un conjunto de variables independientes $x_1, x_2, \ldots, x_n$, ajustando una función lineal que minimiza la suma de los errores cuadráticos entre los valores observados y los estimados.

La ecuación general se expresa como $y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon$, donde los coeficientes β_i representan la contribución de cada predictor [18].

Random Forest: Es un algoritmo desarrollado por Breiman (2001), es un método de ensamblado que combina múltiples árboles de decisión entrenados sobre diferentes subconjuntos aleatorios de datos y variables. Cada árbol contribuye con un voto independiente, y el promedio o la mayoría de ellos determina la predicción final. Este enfoque mejora la estabilidad y precisión del modelo al reducir el sobreajuste y manejar de manera efectiva datos de alta dimensionalidad [19].

Extra Trees (*Extremely Randomized Trees*): Es un algoritmo de aprendizaje supervisado basado en el ensamblado de múltiples árboles de decisión, propuesto por Geurts et al. (2006). A diferencia del modelo *Random Forest*, que selecciona aleatoriamente subconjuntos de variables y muestras, *Extra Trees* introduce además una aleatorización en los umbrales de división de cada nodo, lo que reduce la varianza del modelo y mejora su capacidad de generalización [20].

Gradient Boosting: Es un método introducido por Friedman (2001), se basa en la construcción secuencial de modelos débiles habitualmente árboles de decisión poco profundos que corrigen los errores cometidos por los anteriores. Cada nuevo árbol se ajusta para minimizar una función de pérdida mediante el gradiente descendente, lo que confiere al modelo alta precisión y capacidad de ajuste en tareas de regresión y clasificación [21].

Support Vector Machine -SVM (*Máquinas de Vectores de Soporte*): Es un modelo de aprendizaje supervisado propuesto por Cortes y Vapnik (1995). Su principio se basa en encontrar un hiperplano óptimo que separe las clases en el espacio de características, maximizando el margen entre los conjuntos de datos. Mediante el uso de funciones kernel, el modelo puede proyectar los datos a espacios de mayor dimensionalidad para resolver problemas no lineales [22].

Multilayer Perceptron (MLP): El modelo pertenece a la familia de redes neuro-

nales artificiales de tipo *feed-forward*. Está compuesto por una capa de entrada, una o más capas ocultas y una capa de salida, donde cada neurona aplica una función de activación no lineal a la combinación ponderada de las entradas. El entrenamiento del modelo se realiza mediante el algoritmo de retropropagación del error (*backpropagation*), ajustando los pesos sinápticos para minimizar una función de pérdida [23].

El MLP es capaz de aproximar relaciones complejas y no lineales entre variables, lo que lo convierte en una alternativa robusta frente a los modelos tradicionales.

3.1.6. Métodos de modelado.

Label Encoding: Es un método que forma parte del módulo de preprocesamiento de la biblioteca *scikit-learn* y permite convertir etiquetas categóricas en valores numéricos enteros, de manera que puedan ser interpretadas por los algoritmos de aprendizaje automático.

Según la documentación oficial de *scikit-learn* [24], este procedimiento asigna a cada categoría un número entero único, manteniendo la correspondencia unívoca entre la variable original y su representación codificada.

GridSearchCV: Es una herramienta del módulo *model_selection* de la biblioteca *scikit-learn* que permite realizar una búsqueda exhaustiva de combinaciones de hiperparámetros mediante validación cruzada. Según la documentación oficial [25], el método ejecuta de forma sistemática todas las combinaciones posibles de parámetros definidos por el usuario, entrenando y evaluando cada modelo a través de un esquema de validación cruzada para identificar la configuración que produce el mejor desempeño según una métrica de evaluación específica. Esta estrategia garantiza un proceso reproducible y optimizado para la selección de hiperparámetros en modelos supervisados.

StratifiedKFold: El método pertenece al módulo *model_selection* de la biblioteca *scikit-learn* y permite realizar la división de los datos para validación cruzada manteniendo la proporción de clases en cada partición. De acuerdo con la documentación oficial [26], este enfoque estratificado asegura que cada conjunto de entrenamiento y validación preserve la distribución de clases del conjunto original, reduciendo el sesgo y mejorando la representatividad de los resultados.

3.1.7. Software y herramientas de procesamiento geográficas.

Google Earth Engine (GEE): Según Gorelick *et al.* [27], es una plataforma en la nube que facilita el acceso a cómputo de alto rendimiento para procesar grandes volúmenes de datos geoespaciales sin las cargas operativas típicas; además, a diferencia de la mayoría de centros de supercomputación, está diseñada para difundir resultados hacia investigadores, tomadores de decisión, ONG, personal de campo y público en general, permitiendo que, una vez desarrollado un algoritmo, se generen productos sistemáticos o se implementen aplicaciones interactivas sin requerir experiencia en desarrollo web. En términos de arquitectura, integra un catálogo multipetabyte listo para análisis, co-localizado con un servicio de cómputo intrínsecamente paralelo y accesible mediante una API y un IDE web que habilitan el prototipado rápido y la visualización de resultados.

PostgreSQL: Según la documentación oficial, es un sistema gestor de bases de datos *relacional y de código abierto* que implementa el estándar SQL e incorpora extensiones avanzadas para tipos de datos, concurrencia multiversión (MVCC), transacciones ACID y replicación; su arquitectura modular y el ecosistema de extensiones permiten ajustar rendimiento y funcionalidad desde casos OLTP hasta analítica, con una documentación versionada que cubre instalación, SQL, administración y operación para cada versión estable [28]. Originalmente fue iniciado en la Universidad de California, Berkeley, como sucesor del proyecto POSTGRES liderado por Michael Stonebraker, y su evolución continúa hoy bajo una comunidad global de desarrolladores [29].

PostGIS: Según [30], es una extensión *de código abierto* que añade al motor relacional PostgreSQL tipos y funciones espaciales para trabajar directamente con datos geográficos. Provee los tipos *geometry* y *geography*, soporte para ráster y topología, y un conjunto amplio de funciones para crear, transformar, analizar y validar geometrías. Asimismo, habilita consultas espaciales consistentes con el sistema de referencia de coordenadas y admite reproyección en el flujo de consulta, integrándose con las capacidades transaccionales del SGBD para el almacenamiento, la consulta y el procesamiento espacial dentro de la base de datos.

QGIS (acrónimo de Quantum Geographic Information System): es un software libre y de código abierto para el análisis, visualización, edición y gestión de información geoespacial. Está desarrollado y mantenido por la comunidad internacional de la QGIS Association bajo licencia GNU General Public License (GPL). Su arquitectura modular permite integrar complementos y herramientas avanzadas para el procesamiento espacial, la visualización cartográfica, el manejo de bases de datos

espaciales y la interoperabilidad con otros sistemas mediante estándares abiertos definidos por el Open Geospatial Consortium (OGC) [31].

QGIS admite tanto datos vectoriales como ráster y se ha consolidado como una de las plataformas SIG más utilizadas en la investigación científica, la gestión ambiental y la planificación territorial. Además, su interfaz de programación en Python (PyQGIS) permite automatizar flujos de trabajo complejos, generar algoritmos personalizados y extender sus capacidades hacia el modelamiento geoespacial y la integración con bibliotecas de aprendizaje automático.

GDAL (Geospatial Data Abstraction Library): es una biblioteca de código abierto para la lectura, escritura y transformación de datos geoespaciales ráster y vectoriales. Según la documentación oficial, GDAL proporciona una API unificada que permite procesar una amplia variedad de formatos, aplicar reproyecciones, recortes, mosaicos y conversiones entre sistemas de referencia espacial [32].

Esta herramienta se incorpora comúnmente en flujos de trabajo de teledetección y SIG automatizados, y desempeñó un papel fundamental en este proyecto al facilitar operaciones de procesamiento masivo de ráster (por ejemplo, fusión, cálculo de índices, entre otros).

GRASS GIS (Geographic Resources Analysis Support System): Es un sistema de software libre de código abierto cuya arquitectura modular permite el procesamiento avanzado de datos ráster y vectoriales para análisis geoespacial, modelamiento y teledetección [33].

PyQGIS: Es la interfaz de programación en Python integrada en el entorno del software libre QGIS, desarrollada por la comunidad de la QGIS Association. Según la documentación oficial, PyQGIS permite automatizar flujos de trabajo espaciales, extender funcionalidades mediante scripts personalizados y construir aplicaciones geoespaciales independientes mediante el API de QGIS [34].

Esta capacidad de programación facilita tareas complejas como la carga de capas ráster y vectoriales, la ejecución de algoritmos de procesamiento espacial y la integración con sistemas de bases de datos geoespaciales y herramientas de aprendizaje automático. La naturaleza de PyQGIS como herramienta de código abierto y su compatibilidad con formatos estándares (como GeoPackage, GeoTIFF, PostGIS) la posicionan como un recurso clave para la investigación aplicada en teledetección y análisis térmico urbano.

La combinación de su naturaleza abierta, su compatibilidad con múltiples forma-

tos (como GeoPackage, GeoTIFF, Shapefile, entre otros) y su activa comunidad de desarrollo lo convierten en una herramienta robusta, interoperable y en constante evolución para el análisis geoespacial.

Geopackage: Es un formato abierto, basado en estándares, independiente de plataforma, portátil y autodescriptivo para el almacenamiento y transferencia de información geoespacial en un solo archivo SQLite. Define un conjunto de convenciones para incluir entidades vectoriales, mosaicos ráster (tile matrix sets) de imágenes y mapas ráster, atributos no espaciales y extensiones. Al cumplir el estándar sin extensiones específicas de proveedor, asegura interoperabilidad entre entornos informáticos empresariales y personales [35]

Bounding Boxes: Es una herramienta de QGIS que permite calcular el cuadro delimitador (*bounding box* o envolvente mínima) de cada entidad contenida en una capa vectorial. De acuerdo con la documentación oficial [36], el algoritmo genera un rectángulo alineado con los ejes principales capaz de contener completamente cada geometría de entrada, ya sea poligonal o lineal.

Esta herramienta resulta especialmente útil en el procesamiento de datos espaciales y en tareas de teledetección, ya que permite establecer marcos espaciales mínimos para operaciones como recortes, intersecciones, muestreos o delimitaciones por área de interés.

Pansharpening: Según [37], el pansharpening es una técnica que combina la alta resolución espacial de las imágenes pancromáticas con la riqueza espectral de las imágenes multiespectrales de baja resolución. Esta fusión tiene como objetivo crear una imagen que integre ambas cualidades, mejorando así la representación de la información espacial y espectral, lo cual resulta útil en aplicaciones como la clasificación de terrenos y el análisis de cambios.

Raster Calculator: Es una herramienta de QGIS que permite realizar operaciones algebraicas sobre valores de píxel en uno o más rásteres de entrada. Según la documentación oficial de QGIS [38], el usuario puede construir expresiones que combinan capas ráster, aplicar operadores matemáticos, lógicos y condicionales, definir la extensión y resolución de salida, y generar un nuevo ráster con los resultados en un formato compatible con GDAL. Esta funcionalidad es clave en el tratamiento de imágenes multiespectrales, ya que permite llevar a cabo cálculos como la conversión de valores digitales a reflectancia o temperatura, la producción de índices espectrales y la integración de variables para modelamiento térmico urbano.

r.fillnulls: Es un módulo que pertenece al conjunto de algoritmos nativos de *GRASS*

GIS y permite rellenar celdas nulas en mapas ráster mediante métodos de interpolación espacial. De acuerdo con la documentación oficial [39], este módulo implementa una interpolación por splines de mínima curvatura o por métodos bilineales y bicúbicos, generando superficies continuas a partir de los valores válidos circundantes.

Su uso resulta esencial en procesos de preprocesamiento de datos multiespectrales, ya que asegura la continuidad espacial del ráster y la eliminación de vacíos generados durante etapas de reproyección o filtrado.

gdal:merge: Es un algoritmo que forma parte del conjunto de herramientas de procesamiento de datos ráster integradas en QGIS y basadas en la biblioteca *GDAL* (Geospatial Data Abstraction Library).

Según la documentación oficial de QGIS [40], este algoritmo permite combinar múltiples archivos ráster en un único archivo de salida, preservando las propiedades espaciales y radiométricas de los insumos originales.

En este estudio, *gdal:merge* fue empleado para generar compuestos multibanda a partir de las bandas reflectivas, facilitando la integración de la información espectral necesaria para el cálculo de índices y la creación de composiciones RGB.

3.1.8. Métricas de evaluación

Coeficiente de determinación R^2: Según la literatura clásica de regresión lineal [41,42], es una medida adimensional de *bondad de ajuste* que cuantifica la proporción de la variabilidad de la variable respuesta explicada por el modelo en comparación con un modelo nulo (que solo utiliza la media). Se define habitualmente como

$$R^2 \;=\; 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \tag{4}$$

donde y_i son los valores observados, $\hat{y}_i$ las predicciones del modelo y $\bar{y}$ la media de la respuesta. Valores cercanos a 1 indican que el modelo explica una fracción elevada de la variabilidad observada, mientras que valores próximos a 0 reflejan un ajuste pobre; no obstante, su interpretación debe contextualizarse según la estructura del modelo, la presencia de no linealidades y la posible sobreparametrización.

Error cuadrático medio de la raíz (RMSE): Es una métrica de error ampliamente utilizada para evaluar el desempeño de modelos de regresión y de predicción continua, definida como la raíz cuadrada del promedio de los errores al cuadrado entre los valores observados y los valores predichos [43]:

$$\text{RMSE} \;=\; \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{5}$$

El RMSE se expresa en las mismas unidades de la variable respuesta, lo que facilita su interpretación práctica como un "error típico" de predicción. Al penalizar de forma cuadrática los errores grandes, es especialmente sensible a valores atípicos, por lo que su uso suele complementarse con otras métricas (como el MAE) para obtener una evaluación más equilibrada del modelo [43].

Precisión (Precision): En el contexto de clasificación supervisada, la precisión cuantifica la proporción de instancias predichas como positivas que son efectivamente positivas [44]. Dado el conteo de verdaderos positivos (TP) y falsos positivos (FP), se define como:

$$\text{Precision} \;=\; \frac{TP}{TP + FP}. \tag{6}$$

Valores altos de precisión indican que el clasificador comete pocos falsos positivos al etiquetar la clase de interés.

Exhaustividad (Recall): También denominada *sensibilidad*, mide la proporción de instancias positivas reales que el modelo es capaz de recuperar correctamente [45]. Dado el número de verdaderos positivos (TP) y falsos negativos (FN), se expresa como:

$$\text{Recall} \;=\; \frac{TP}{TP + FN}. \tag{7}$$

Un valor elevado de *recall* indica que el clasificador omite pocas instancias de la clase positiva (es decir, comete pocos falsos negativos).

Medida F1 (F1-score): Es una métrica compuesta que resume el compromiso entre precisión y exhaustividad mediante su media armónica [46]. Si Precision y Recall representan las métricas anteriores, la medida F1 se define como:

$$F_1 \;=\; 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{8}$$

Esta formulación penaliza de manera simétrica los desequilibrios entre precisión y exhaustividad, por lo que resulta útil cuando se desea un balance entre ambas.

Exactitud (Accuracy): Es una métrica global que cuantifica la proporción de predicciones correctas sobre el total de instancias evaluadas [47]. Dados los verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN), se define como:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{9}$$

La *accuracy* resulta intuitiva y fácil de interpretar, aunque puede ser engañosa en escenarios con clases altamente desbalanceadas, motivo por el cual suele complementarse con precisión, *recall* y F1.

3.2. Antecedentes.

3.2.1. Análisis espacial de islas de calor en la ciudad de Bogotá: los efectos de la urbanización, un estudio desde la percepción remota.

Según la investigación realizada por Cristian Salas Pérez y Daniel Felipe Coy Castro en 2019, la urbanización en Bogotá ha intensificado el fenómeno de islas de calor, generando microclimas urbanos con temperaturas superiores a las de las zonas periféricas. Los autores argumentan que el rápido crecimiento urbano y la reducción de áreas verdes han exacerbado este problema. El objetivo principal del estudio es cuantificar el impacto del uso del suelo sobre la temperatura de superficie y evaluar el rol mitigador de la vegetación en el contexto de un entorno densamente urbanizado [4].

Para los cálculos de temperatura de superficie y análisis de islas de calor, el estudio utiliza imágenes satelitales Landsat desde 1985 hasta 2018 y aplica técnicas avanzadas de clasificación supervisada orientada a objetos. Esta clasificación permite segmentar las áreas urbanas y vegetales dentro de las imágenes, facilitando el monitoreo detallado de cambios en la cobertura del suelo. Los índices espectrales NDVI (Normalized Difference Vegetation Index) y NDBI (Normalized Difference Built-up Index) son empleados para medir la vegetación y las superficies construidas, respectivamente. Estos índices permiten calcular variaciones en la temperatura de superficie asociadas con la distribución de estas coberturas, proporcionando una base sólida para la identificación de áreas con altas temperaturas relativas (islas de calor).

En el análisis estadístico, se utiliza el modelo espacial SARAR (Spatial Autoregressive with Additional Regressors), el cual cuantifica la relación entre la temperatura de superficie y variables como la vegetación y el suelo urbano. Este modelo incorpora la

estructura espacial de los datos para calcular el impacto directo del suelo urbano y la vegetación en la temperatura de superficie, permitiendo una interpretación precisa de la influencia de cada tipo de cobertura. Los resultados muestran que las áreas urbanizadas incrementan la temperatura de superficie en aproximadamente 20°C, mientras que la vegetación logra reducir esta temperatura en cerca de 13°C, lo que demuestra el papel crucial de las áreas verdes en la regulación térmica.

Finalmente, el estudio examina la influencia del clima regional en la temperatura urbana al comparar los datos de temperatura de superficie con índices climáticos, como El Niño y La Niña. La baja correlación encontrada sugiere que el fenómeno de islas de calor en Bogotá es impulsado principalmente por los procesos locales de urbanización. Este hallazgo enfatiza la importancia de conservar áreas verdes como una estrategia de mitigación y destaca el valor de la percepción remota y el análisis espacial en la planificación urbana sostenible.

3.2.2. Analysis of urban heat islands with landsat satellite images and GIS in Kuala Lumpur Metropolitan City.

Según el estudio realizado por Kasniza Jumari et al. en 2023, el fenómeno de las islas de calor urbanas (UHI) en Kuala Lumpur se ha intensificado debido a la expansión urbana, lo que ha motivado el análisis técnico de la temperatura de la superficie terrestre (LST) en función de la urbanización. El estudio emplea imágenes satelitales Landsat-8 de los años 2013 y 2021 y utiliza técnicas de percepción remota térmica en combinación con herramientas de análisis espacial en sistemas de información geográfica (GIS) para identificar y cuantificar la variabilidad térmica en áreas urbanas y rurales [48].

Para calcular la LST, se seleccionaron tres bandas de Landsat-8: Banda 4 (Rojo), Banda 5 (Infrarrojo cercano) y Banda 10 (Térmica), las cuales son esenciales en el procesamiento de índices espectrales y temperatura de brillo (BT). Mediante el índice NDVI (Normalized Difference Vegetation Index), se evaluó la densidad de vegetación y, con el índice NDBI (Normalized Difference Built-up Index), se identificaron áreas construidas. Estos índices permitieron obtener un mapa detallado de variaciones de temperatura de superficie, aplicando fórmulas de radiancia espectral y de transferencia de radiación para ajustar la emisividad y calcular el LST, un paso técnico fundamental para la cuantificación de las UHI.

En términos estadísticos, la investigación implementa el modelo ANOVA para verificar la significancia de los cambios de temperatura observados entre 2013 y 2021.

Este análisis estadístico confirmó un incremento estadísticamente significativo en la intensidad de UHI, con diferencias de temperatura que pasaron de 10.8°C en 2013 a un rango de 16.1°C en 2021 en áreas urbanizadas como Sungai Batu, donde la reducción de vegetación ha incrementado la vulnerabilidad al calentamiento urbano.

El estudio sugiere la efectividad de la cobertura vegetal como mitigador de las UHI, basándose en los resultados del NDVI y en los valores más bajos de LST en áreas boscosas como Bukit Ketumbar. Esta investigación concluye que la expansión de áreas verdes es una estrategia clave para la mitigación del calentamiento urbano, y que el uso combinado de percepción remota térmica y GIS es fundamental en la planificación de entornos urbanos sostenibles [48].

3.2.3. Machine learning for Urban Heat Island (UHI) Analysis: Predicting Land Surface Temperature (LST) in Urban Environments.

El artículo de Tanoori et al. [5] se centra en el análisis de la isla de calor urbana (UHI) y la predicción de la temperatura de la superficie terrestre (LST) en entornos urbanos mediante métodos de aprendizaje automático. La investigación examina cómo la configuración y el uso del suelo en la ciudad de Shiraz, Irán, afectan la distribución del calor. Los autores emplean varias técnicas de machine learning para identificar patrones de temperatura asociados con distintos tipos de cobertura de suelo, como áreas urbanas construidas, vegetación y suelos descubiertos. El objetivo es determinar cómo las características del paisaje urbano influyen en las temperaturas y contribuir a estrategias de planificación que mitiguen el calor y mejoren el confort térmico.

Para lograr este análisis, el estudio emplea cuatro algoritmos de machine learning: Redes Neuronales Profundas (DNN), Extreme Gradient Boosting (XGBoost), Random Forest (RF) y Support Vector Machine (SVM). Estos modelos son evaluados y comparados para determinar cuál ofrece una predicción más precisa de la LST en los diferentes tipos de cobertura del suelo. Los resultados mostraron que los modelos DNN y XGBoost lograron el mejor desempeño en términos de precisión, siendo capaces de capturar relaciones complejas entre la configuración del paisaje y la temperatura superficial, especialmente en áreas de vegetación y suelo urbano.

El estudio también utiliza imágenes satelitales de alta resolución para analizar la LST en la región metropolitana de Shiraz, específicamente datos de Landsat. Estas imágenes permitieron evaluar la influencia de las métricas de configuración del paisaje, como la proporción de áreas construidas y la densidad de bordes, en la

variación de la LST. Estas métricas reflejan la fragmentación y la continuidad de los patrones urbanos, proporcionando información crucial sobre cómo la morfología urbana afecta la distribución de calor.

En conclusión, esta investigación destaca el potencial de los algoritmos de aprendizaje automático para prever la distribución de calor en ciudades con crecimiento acelerado. Los hallazgos subrayan la importancia de integrar métricas de configuración del paisaje en el análisis de la UHI y sugieren que el uso de técnicas avanzadas de predicción, como DNN y XGBoost, puede ser una herramienta valiosa para la planificación urbana y la formulación de políticas de mitigación de calor. Este estudio recomienda la aplicación de estos enfoques en diferentes contextos urbanos y propone la exploración de factores sociales y materiales que también influyen en el UHI.

4. Obtención y procesamiento de imágenes satelitales y estaciones meteorológicas.

En este capítulo se detalla el proceso de delimitación de la zona de estudio y la obtención y preparación de las imágenes satelitales multiespectrales provenientes de las misiones Landsat 8 y 9 (debido a la disponibilidad de datos temporales y espaciales, no se consideraron otras misiones para análisis), utilizadas en el desarrollo del presente estudio. Asimismo, se aborda la consolidación de los registros meteorológicos obtenidos de las estaciones operadas por el IDEAM y la DIMAR. También se describen los criterios de selección de las escenas y las actividades de preprocesamiento aplicadas, incluyendo reproyección, recorte, escalado, limpieza de valores atípicos y reconstrucción de datos nulos. Estas operaciones fueron fundamentales para garantizar la coherencia espacial y temporal de los datos, así como su compatibilidad con los algoritmos de modelamiento empleados en las fases posteriores del análisis.

4.1. Definición de zona de estudio y obtención de datos.

4.1.1. Descarga de límites oficiales municipales.

En la etapa preliminar del estudio, se efectuó la delimitación geoespacial de las zonas de análisis, basándose en los límites municipales de las ciudades de Barranquilla, Cartagena y Santa Marta, con un enfoque particular en las áreas urbanas de cada entidad territorial. Con este propósito, se descargaron los límites municipales oficiales, establecidos en los Planes de Ordenamiento Territorial (POT) de cada ciudad, a través de la plataforma Colombia en Mapas del Instituto Geográfico Agustín Codazzi (IGAC), en formato GeoPackage [35].

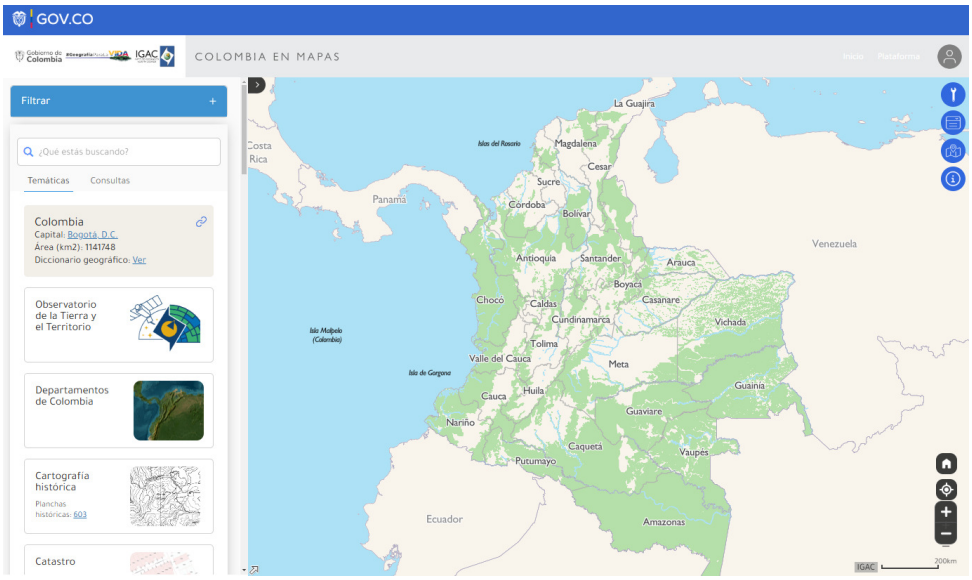


Figura 1: Límites municipales descargados desde la plataforma Colombia en Mapas del IGAC [1]

.

El procesamiento de los límites municipales se efectuó en el software de QGIS, realizando la lectura y visualización de los insumos descargados. Posteriormente, se procedió a la selección y extracción de las áreas urbanas municipales, asegurando así una delimitación coherente y estructurada del espacio urbano a analizar en las ciudades de Barranquilla, Cartagena y Santa Marta.

4.1.2. Definición de las zonas de estudio.

Con el propósito de definir el área de estudio, se generaron *Bounding Boxes* para cada una de las áreas urbanas municipales. Este procesamiento permitió obtener el área rectangular envolvente de las zonas de estudio en cada ciudad. De esta manera se obtuvieron las áreas que definieron y delimitaron el alcance geográfico espacial para el procesamiento de los productos en la ejecución del proyecto.

Las zonas de estudio para las ciudades de Barranquilla (Figura 2), Cartagena (Figura 3) y Santa Marta (Figura 4) se representaron mediante rectángulos de color rojo, dentro de los cuales se observan las áreas urbanas de cada ciudad sobre un mapa satelital.

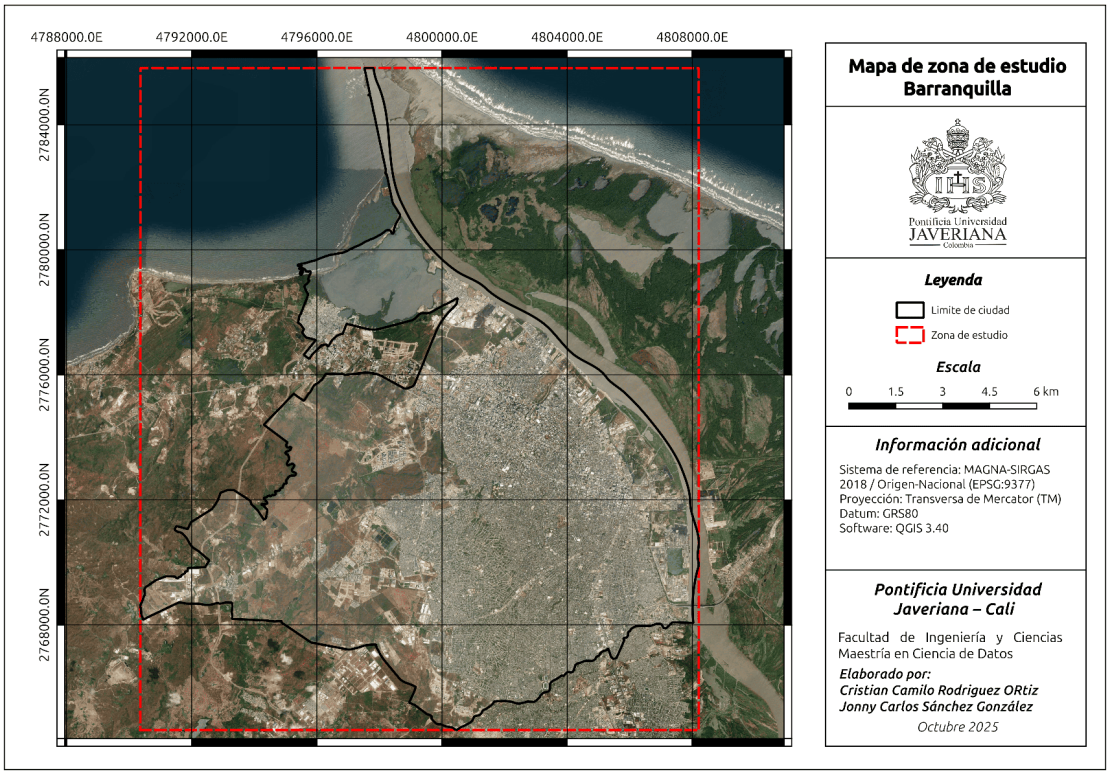


Figura 2: Zona de estudio ciudad de Barranquila.

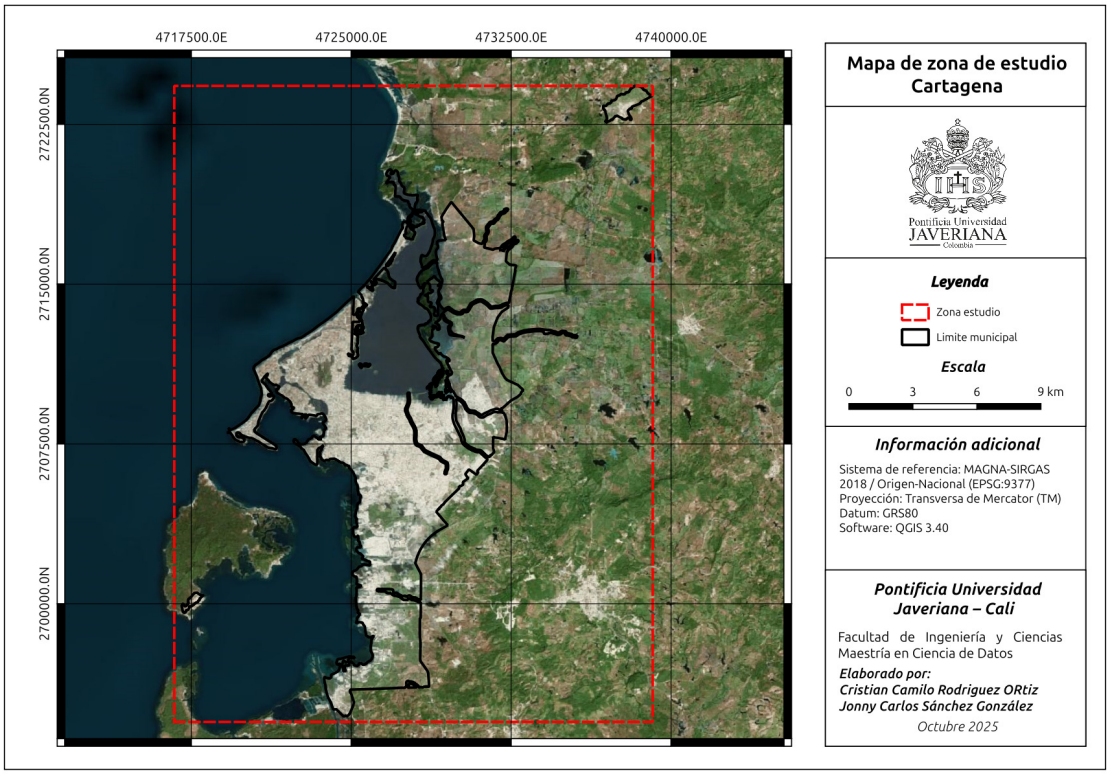


Figura 3: Zona de estudio ciudad de Cartagena

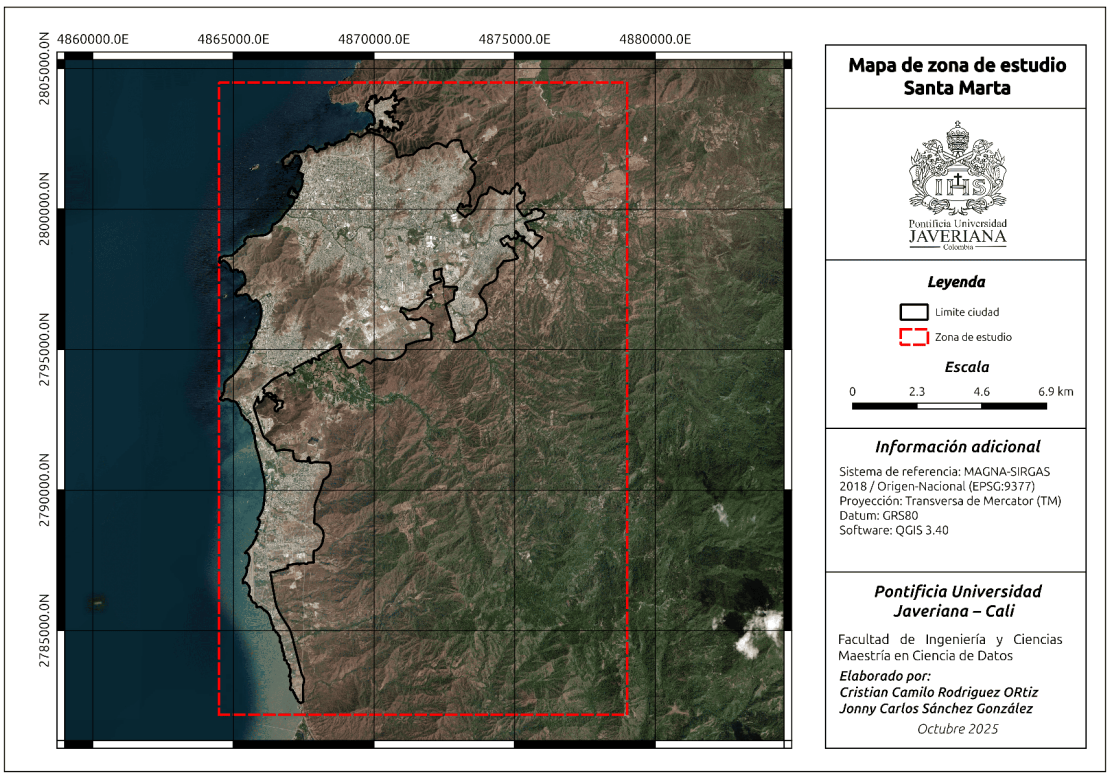


Figura 4: Zona de estudio ciudad de Santa Marta

Alineado con las delimitaciones cartográficas de las zonas de estudio, se calcularon las áreas superficiales del perímetro urbano en cada ciudad en hectáreas. Los valores totales se resumen en la Tabla 1.

Tabla 1: Área del perímetro urbano por ciudad.

Ciudad	**Área (ha)**
Santa Marta	7 391.82
Cartagena	12 942.50
Barranquilla	15 379.10

4.1.3. Descarga de imágenes satelitales multiespectrales.

Como parte del proceso de adquisición de datos satelitales, se descargaron imágenes Landsat 8 y 9 para el periodo de 2015 a 2024. La selección de estas imágenes se hizo con criterios específicos para asegurar la coherencia espacial y temporal del análisis:

- La cobertura espacial de las zonas de estudio previamente delimitadas.
- Umbral máximo de nubosidad en las imágenes satelitales de un 5 %, con el fin

de minimizar la interferencia atmosférica y mejorar la precisión en la extracción de parámetros espectrales y térmicos.

- Imágenes satelitales de productos preprocesados *Nivel 2A*, caracterizadas por incluir corrección atmosférica aplicada mediante modelos de transferencia radiativa. Este nivel proporciona valores calibrados de *reflectancia de superficie* y *temperatura de brillo*, eliminando los efectos de dispersión y absorción atmosférica. Además, las imágenes incorporan *máscaras de calidad (QA bands)* que permiten identificar la presencia de nubes, sombras y otros artefactos que pueden afectar la integridad del análisis espectral y térmico.
- Imágenes capturadas durante el *primer trimestre de cada año*, lo que permite reducir la variabilidad estacional en la temperatura superficial y garantizar homogeneidad y consistencia temporal en los análisis.

El proceso de descarga se realizó desde la plataforma USGS Earth Explorer [49], garantizando que cada una de las escenas cumpliera con los criterios predefinidos. Posteriormente, las imágenes fueron almacenadas y categorizadas de acuerdo a su fecha de adquisición y cobertura geográfica, con el fin de prepararlas para las fases subsiguientes de procesamiento, corrección y análisis. En la Tabla 2 se presenta el inventario de imágenes satelitales que constituye la base de datos utilizada para el análisis multitemporal del fenómeno de islas de calor urbanas.

Tabla 2: Inventario de imágenes satelitales Landsat descargadas.

Barranquilla			**Cartagena**			**Santa Marta**		
Satélite	**Path/Row**	**Fecha**	**Satélite**	**Path/Row**	**Fecha**	**Satélite**	**Path/Row**	**Fecha**
Landsat 8	009/052	2015-04-01	Landsat 8	009/053	2015-04-01	Landsat 8	009/052	2015-04-01
Landsat 8	009/052	2016-01-14	Landsat 8	009/053	2016-01-14	Landsat 8	009/052	2016-01-14
Landsat 8	009/052	2017-01-16	Landsat 8	009/053	2017-01-16	Landsat 8	009/052	2017-01-16
Landsat 8	009/052	2018-02-04	Landsat 8	009/053	2018-02-04	Landsat 8	009/052	2018-02-04
Landsat 8	009/052	2019-10-16	Landsat 8	009/053	2019-10-16	Landsat 8	009/052	2019-10-16
Landsat 8	009/052	2020-03-29	Landsat 8	009/053	2020-03-29	Landsat 8	009/052	2020-03-29
Landsat 8	009/052	2021-12-12	Landsat 8	009/053	2021-12-12	Landsat 8	009/052	2021-12-12
Landsat 9	009/052	2022-02-23	Landsat 8	010/053	2022-02-22	Landsat 9	009/052	2022-02-23
Landsat 8	009/052	2023-01-17	Landsat 8	009/053	2023-01-17	Landsat 8	009/052	2023-01-17
Landsat 8	009/052	2024-01-20	Landsat 8	009/052	2024-01-20	Landsat 8	009/052	2024-01-20

Previo al procesamiento, se realizó un análisis del peso total de las imágenes satelitales Landsat descargadas con el fin de dimensionar el volumen de información manejado durante el proyecto. En total, se recopilaron veinte escenas multiespectrales, con un tamaño combinado de aproximadamente 16.61GB.

El Tabla 3 presenta el detalle de las imágenes por ciudad y fecha de captura, junto con su peso individual, evidenciando la magnitud del conjunto de datos empleado en las etapas de preprocesamiento y modelamiento.

Tabla 3: Inventario de imágenes satelitales Landsat descargadas (con peso).

Barranquilla			**Cartagena**			**Santa Marta**		
Satélite	**Fecha**	**Peso [MB]**	**Satélite**	**Fecha**	**Peso [MB]**	**Satélite**	**Fecha**	**Peso [MB]**
Landsat 8	2015-04-01	825.28	Landsat 8	2015-04-01	933.67	Landsat 8	2015-04-01	825.28
Landsat 8	2016-01-14	766.09	Landsat 8	2016-01-14	917.96	Landsat 8	2016-01-14	766.09
Landsat 8	2017-01-16	806.68	Landsat 8	2017-01-16	915.58	Landsat 8	2017-01-16	806.68
Landsat 8	2018-02-04	758.29	Landsat 8	2018-02-04	915.12	Landsat 8	2018-02-04	758.29
Landsat 8	2019-10-16	785.78	Landsat 8	2019-10-16	921.26	Landsat 8	2019-10-16	785.78
Landsat 8	2020-03-29	822.58	Landsat 8	2020-03-29	924.51	Landsat 8	2020-03-29	822.58
Landsat 8	2021-12-12	824.56	Landsat 8	2021-12-12	926.50	Landsat 8	2021-12-12	824.56
Landsat 9	2022-02-23	805.32	Landsat 8	2022-02-22	769.16	Landsat 9	2022-02-23	805.32
Landsat 8	2023-01-17	767.64	Landsat 8	2023-01-17	941.44	Landsat 8	2023-01-17	767.64
Landsat 8	2024-01-20	841.42	Landsat 8	2024-01-20	914.97	Landsat 8	2024-01-20	841.42
Total Barranquilla:		8003.64 MB	**Total Cartagena:**		9080.17 MB	**Total Santa Marta:**		8003.64 MB

Notas. Los pesos provienen de los directorios originales. Las escenas 009/052 se emplean para Barranquilla y Santa Marta; por ello, el **total global único** de descargas no es la suma de las tres columnas, sino el de los 20 directorios analizados: **16.61 GB**.

No se utilizaron imágenes satelitales de otras plataformas debido a que, para el desarrollo de este proyecto, era indispensable contar con un sensor térmico operativo capaz de proporcionar mediciones consistentes de temperatura de la superficie terrestre. En la actualidad, muy pocas misiones satelitales ofrecen bandas térmicas calibradas y con disponibilidad histórica continua. La serie Landsat constituye una de las pocas plataformas que integran un canal térmico de libre acceso, con resolución espacial y radiométrica adecuadas para estudios urbanos y ambientales.

4.2. Procesamiento de imágenes satelitales.

4.2.1. Reproyección de imágenes satelitales.

Las imágenes multiespectrales fueron reproyectadas desde el sistema de referencia *EPSG:32618 (WGS84 / UTM zona 18N)* al sistema oficial adoptado en Colombia, *EPSG:9377*, conforme a lo establecido por el Instituto Geográfico Agustín Codazzi (IGAC) [50].

El proceso de reproyección de las imágenes satelitales fue automatizado mediante un script desarrollado en *PyQGIS*, el cual recorre de forma iterativa los directorios del proyecto, identifica los archivos en formato *.TIF* y aplica la transformación espacial

al sistema de referencia *EPSG:9377* utilizando las herramientas de *GDAL*.

En la Figura 5 se presenta el diagrama que resume la secuencia lógica implementada en el script de reproyección.

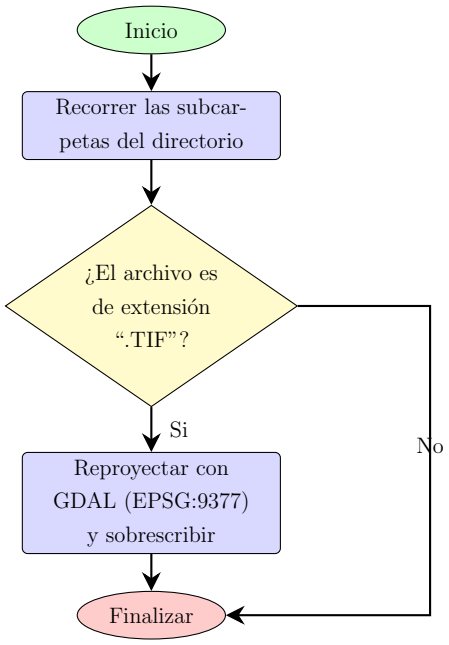


Figura 5: Diagrama del proceso para la reproyección de imágenes satelitales al sistema EPSG:9377.

El script desarrollado para la reproyección de las imágenes satelitales se encuentra disponible públicamente en el repositorio de GitHub del proyecto: `https://github.com/jonnykewin/tesis_MCD/blob/main/1_reproject_raster.py`.

4.2.2. Recorte de imágenes satelitales.

Después de reproyectar las imágenes, se recortaron las escenas satelitales *Landsat* usando los *Bounding Boxes* que ya se habían hecho para las zonas urbanas de Barranquilla, Cartagena y Santa Marta. Esto sirvió para delimitar las imágenes reproyectadas y asegurar que los análisis se centraran solo en las zonas de estudio.

El procedimiento fue automatizado mediante un script desarrollado en *PyQGIS*, en conjunto con las utilidades del paquete *GDAL*. Esta combinación de herramientas permitió una delimitación espacial precisa, eficiente y reproducible sobre el conjunto de imágenes satelitales.

En la Figura 6 se presenta el diagrama que resume las fases del proceso de recorte implementado.

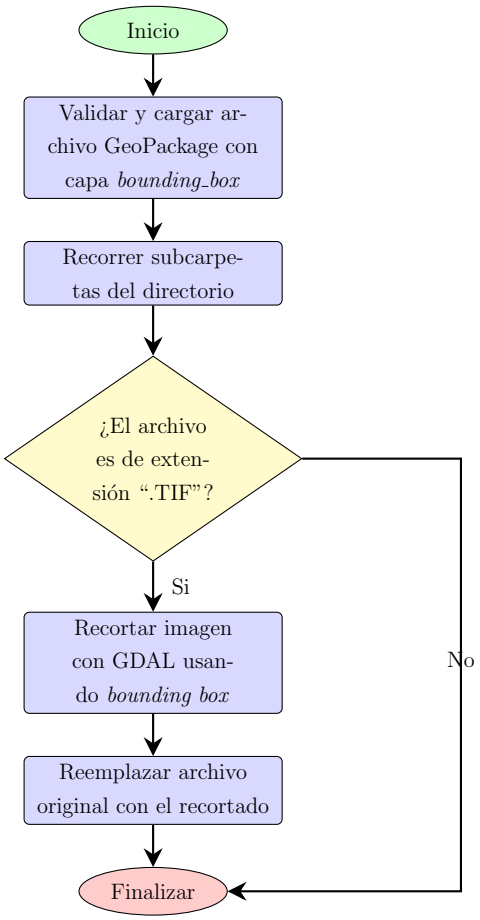


Figura 6: Diagrama del proceso para el recorte de imágenes satelitales utilizando *Bounding Box*.

El script desarrollado para la recorte de las imágenes satelitales se encuentra disponible públicamente en el repositorio de GitHub del proyecto: `https://github.com/jonnykewin/tesis_MCD/blob/main/2_clip_raster.py`

4.2.3. Escalado de imágenes satelitales: conversión de valores digitales a reflectancia y temperatura.

Como parte del preprocesamiento de las imágenes multiespectrales, se realizó la conversión de los valores digitales (*Digital Numbers* – DN) a parámetros físicos de *reflectancia de superficie* y *temperatura (kelvin)* . Este procedimiento es requerido para garantizar la precisión en el cálculo de índices espectrales como el NDVI y el NDBI, así como en la estimación de la temperatura de la superficie terrestre (LST).

La transformación se ejecutó de manera automatizada mediante un script desarrollado en *PyQGIS*, utilizando la herramienta *Raster Calculator*, integrada en el entorno de procesamiento de QGIS. Las fórmulas de escalado aplicadas se basaron

en las especificaciones técnicas publicadas por el Servicio Geológico de los Estados Unidos (USGS) [51], las cuales establecen los factores de conversión para las bandas *SR_B1–SR_B7*, *ST_B10* y *ST_EMIS*.

En la Figura 7 se presenta el diagrama que resume las fases del proceso de escalado implementado.

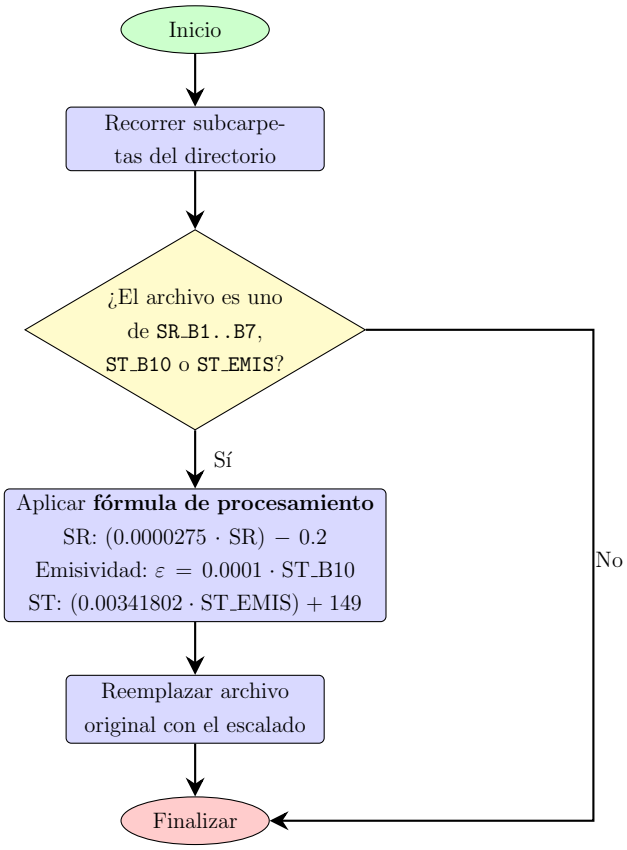


Figura 7: Diagrama de proceso escalado de imágenes (SR, emisividad o ST).

El script desarrollado para el escalado de las imágenes satelitales se encuentra disponible públicamente en el repositorio de GitHub del proyecto `https://github.com/jonnykewin/tesis_MCD/blob/main/3_scale_bands.py`

4.2.4. Limpieza de outliers.

Con el propósito de asegurar la calidad de los datos radiométricos y evitar la presencia de valores atípicos (*outliers*) en las bandas multiespectrales de las imágenes *Landsat*, se implementó un proceso de limpieza sistemática.

Esta etapa consistió en identificar y excluir los valores que se encontraban fuera del rango reflectivo esperado (menores a 0 o mayores a 1), los cuales fueron reemplazados por un valor nulo (-9999). El procedimiento fue automatizado mediante

un script en *PyQGIS* que utilizó el algoritmo *Raster Calculator* dentro del entorno de procesamiento de QGIS, permitiendo una depuración eficiente y homogénea de todas las escenas satelitales reprocesadas.

En la Figura 8 se presenta el diagrama que resume las fases del proceso de limpieza aplicado.

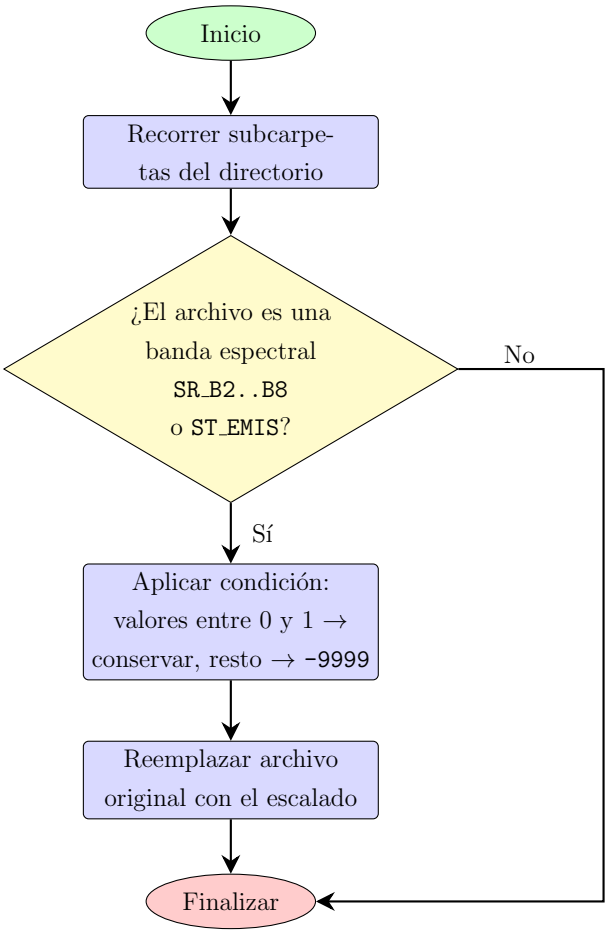


Figura 8: Diagrama del proceso para la limpieza de *outliers*.

El script desarrollado para el limpieza de outliers de las imágenes satelitales se encuentra disponible públicamente en el repositorio de GitHub del proyecto `https://github.com/jonnykewin/tesis_MCD/blob/main/5_clean_outliers.py`

4.2.5. Llenado de datos vacíos o nulos.

Posterior al proceso de limpieza de valores atípicos, se implementó una etapa de llenado de valores nulos con el fin de garantizar la continuidad espacial de las bandas multiespectrales y térmicas. Este procedimiento permitió evitar discontinuidades en las imágenes, especialmente en zonas donde la exclusión de *outliers* había generado

datos vacíos.

El proceso fue automatizado mediante un script desarrollado en *PyQGIS*, el cual recorre las carpetas de trabajo, identifica las bandas espectrales correspondientes (*SR_B2–SR_B7* y *ST_EMIS*) y aplica el algoritmo *r.fillnulls* de GRASS GIS, integrado en el entorno de procesamiento de QGIS. Este algoritmo interpola los valores faltantes mediante un método de superficie spline, conservando la coherencia espacial de la reflectancia y la emisividad entre píxeles contiguos.

En la Figura 9 se presenta el diagrama que resume las fases del proceso de llenado de datos nulos implementado.

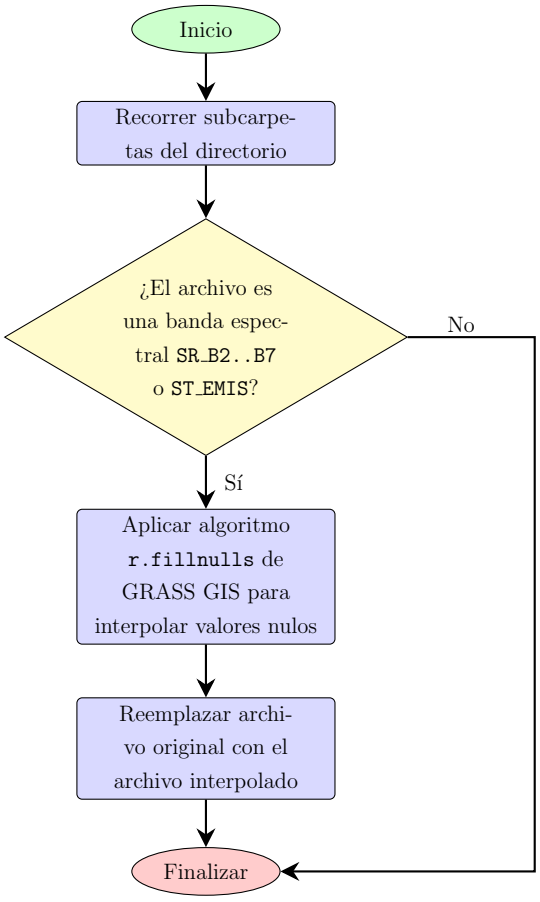


Figura 9: Diagrama del proceso para el llenado de valores nulos en las imágenes multiespectrales.

El script desarrollado para el llenado de datos vacíos de las imágenes satelitales se encuentra disponible públicamente en el repositorio de GitHub del proyecto `https://github.com/jonnykewin/tesis_MCD/blob/main/6_fill_data.py`

4.2.6. Generación de compuesto RGB.

Como parte del preprocesamiento visual, se generaron composiciones RGB a partir de las bandas reflectivas *SR_B4* (rojo), *SR_B3* (verde) y *SR_B2* (azul). Esta fusión proporciona una visualización en color natural de la escena satelital, útil para la interpretación inicial del área de estudio y la verificación cualitativa de las coberturas geográficas en la imagen.

La composición se produjo de forma automatizada mediante un script en *PyQGIS*, utilizando el algoritmo *gdal:merge* del entorno de procesamiento de QGIS, con la opción `SEPARATE=True` para conservar cada banda como capa independiente dentro del archivo resultante y el tipo de salida `UInt16` para mantener la fidelidad radiométrica.

En la Figura 10 se muestra el flujo resumido del proceso de integración de bandas para la obtención del compuesto RGB.

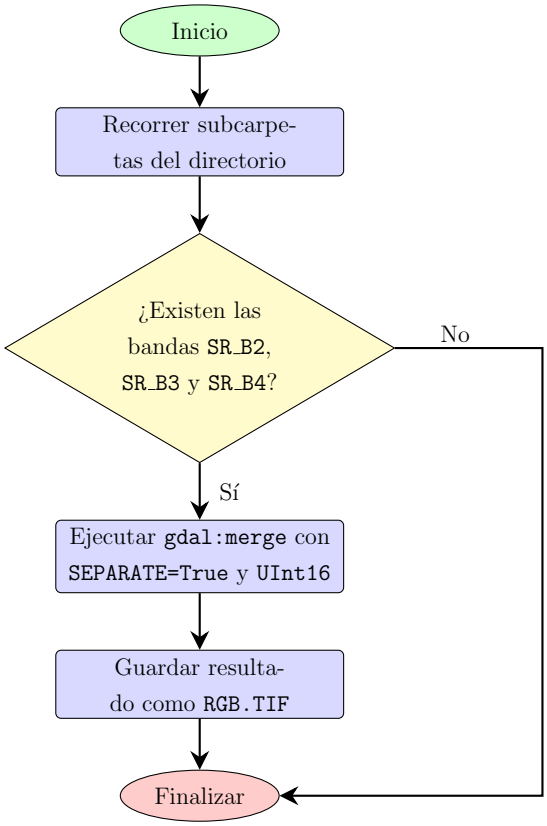
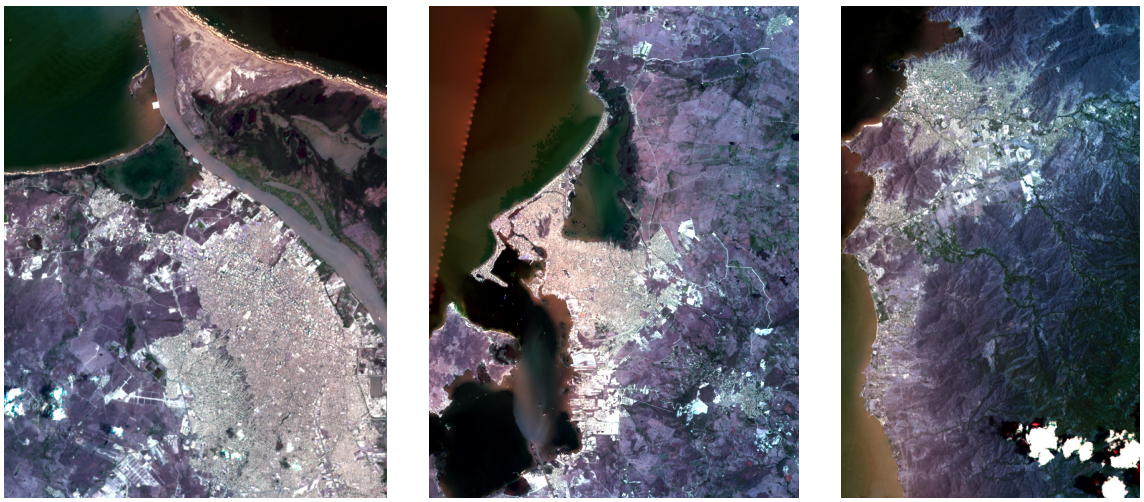


Figura 10: Diagrama del proceso para la integración de bandas y generación del compuesto RGB.

El script desarrollado para la creación de la composición RGB de las imágenes satelitales se encuentra disponible públicamente en el repositorio de GitHub del proyecto

https://github.com/jonnykewin/tesis_MCD/blob/main/7_index_lst_rgb.py

A continuación se presentan las composiciones RGB generadas para las ciudades de Barranquilla, Cartagena y Santa Marta, a partir de imágenes satelitales capturadas en el año 2015, como se muestra en la Figura 11. Estas composiciones permiten una visualización general del territorio y facilitan una evaluación preliminar del contenido espacial presente en cada escena.



(a) Barranquilla (b) Cartagena (c) Santa Marta

Figura 11: Composición RGB en 2015 para Barranquilla, Cartagena y Santa Marta.

4.3. Procesamiento de estaciones meteorológicas.

4.3.1. Estaciones: datos oficiales (IDEAM y DIMAR).

Los datos de estaciones meteorológicas representaron un insumo fundamental para el desarrollo del proyecto. No obstante, el acceso y procesamiento de esta información presentó varios retos relacionados con el volumen de registros, la dispersión de las fuentes y la heterogeneidad en los formatos.

En el caso del IDEAM, la descarga inicial correspondió a un archivo `CSV` en bruto de aproximadamente 11 GB, proveniente del Catálogo Nacional de Estaciones. Este conjunto de datos contenía más de 506 000 registros. Por su parte, los datos de la Dirección General Marítima (DIMAR) fueron obtenidos a través del portal CECOLDO, el cual provee hipervínculos individuales para cada estación. En ambos casos, los datos fueron cargados a una base de datos `PostgreSQL` con extensión `PostGIS`, donde se estructuraron según un esquema estandarizado.

Durante la importación, se realizó un filtrado temporal para conservar únicamente

los registros cuyas fechas coincidieran con las capturas de imágenes *Landsat* entre 2015 y 2024. Asimismo, se identificaron valores atípicos en las mediciones de temperatura, como `-99999`, que fueron eliminados mediante consultas `SQL` para garantizar la calidad de la información.

4.3.2. Integración con conjuntos alternos (ERA5/ERA5-Land vía GEE).

Tras depurar los datos oficiales (IDEAM y DIMAR) se evidenciaron vacíos temporales y una cobertura espacial insuficiente en sectores costeros y marinos. Esta limitación fue especialmente crítica en el caso de la ciudad de Cartagena, donde inicialmente solo se contaba con registros entre 2015 y 2019, con una densidad de muestras muy baja sobre el área de estudio, lo que restringía el análisis multianual y la caracterización robusta del comportamiento térmico. Para preservar la comparabilidad interanual y mejorar la representatividad espacial, el conjunto se complementó con series horarias de reanálisis obtenidas en Google Earth Engine: ERA5-Land para superficie continental y ERA5 para celdas oceánicas.

Para cada fecha, se seleccionó la observación más cercana a las 10:00 hora local (UTC−5). La variable temperatura del aire a 2 m (`temperature_2m`, K) se transformó a °C restando 273,15. Cuando fue necesario, se definieron estaciones virtuales en el mar para cubrir puntos estratégicos de la franja costera. Los resultados se exportaron en `CSV` e ingresaron en PostGIS con el mismo esquema de los datos oficiales, preservando campos de procedencia y trazabilidad. No se aplicó corrección de sesgo; los valores de ERA5/ERA5-Land se utilizaron como fuente complementaria estandarizada para completar continuidad temporal y espacial.

4.3.3. Estructuración e integración de datos.

Una vez depurados, los registros se estandarizaron en archivos `CSV` y se incorporaron a un esquema específico dentro de la base geoespacial. La tabla resultante se diseñó con los siguientes campos:

- **fuente**: procedencia del dato (p. ej., `ideam`, `dimar`, `era5`, `era5_land`).
- **ciudad**: municipio asociado a la estación.
- **fecha_toma**: `timestamp` normalizado en `UTC`, que integra fecha y hora de observación.
- **codigo_estacion**: identificador único por estación.
- **medicion**: temperatura en °C (`numeric`).

- **geometria**: tipo punto.

Para la consolidación, se implementó un proceso ETL en `SQL`. Se gestionaron tablas independientes para cada fuente (`estaciones_dimar`, `estaciones_ideam`, `estaciones_era5_hourly` y `estaciones_era5_land`) y, a partir de ellas, se construyó la tabla final `estaciones_consolidadas`, la cual integró de manera homogénea los registros oficiales y los conjuntos alternos.

El resultado fue un conjunto único, estandarizado y espacialmente habilitado, con el esquema `(fuente, ciudad, fecha_toma, codigo_estacion, medicion, geometria)`. La Tabla 4 mostró un extracto representativo con mediciones validadas en Barranquilla, Cartagena y Santa Marta, junto con sus coordenadas. Este *dataset* sirvió como insumo para la fase de validación cruzada de la Temperatura de la Superficie Terrestre (LST) derivada de imágenes satelitales.

Tabla 4: Fragmento del *dataset* consolidado de estaciones meteorológicas (primeros 10 registros).

fuente	ciudad	fecha_toma	codigo_estacion	medicion	geometria
ideam	cartagena	2019-01-06 11:00:00.000	14015020	28.9	POINT (-75.516 10.447)
ideam	cartagena	2016-01-14 23:00:00.000	14015020	26.2	POINT (-75.516 10.447)
ideam	cartagena	2018-02-04 01:00:00.000	14015020	24.7	POINT (-75.516 10.447)
ideam	barranquilla	2020-03-29 17:00:00.000	29004520	27.8	POINT (-74.785 11.00638889)
ideam	barranquilla	2019-01-06 13:00:00.000	29004520	30.5	POINT (-74.785 11.00638889)
ideam	cartagena	2017-01-16 14:00:00.000	14015020	29.1	POINT (-75.516 10.447)
ideam	santa marta	2020-03-29 04:00:00.000	15015120	27.5	POINT (-74.18591667 11.22305556)
ideam	barranquilla	2023-01-17 05:00:00.000	29004520	24.6	POINT (-74.785 11.00638889)
ideam	cartagena	2019-01-06 10:00:00.000	14015020	30.7	POINT (-75.516 10.447)
ideam	cartagena	2015-04-01 04:00:00.000	14015020	25.8	POINT (-75.516 10.447)

Finalmente, se presentó la distribución espacial de las estaciones oficiales (IDEAM y DIMAR) y de las estaciones virtuales provenientes de ERA5/ERA5-Land mediante mapas por ciudad, que incluían el límite municipal y la zona de estudio (EPSG:9377), con el fin de evaluar su cobertura y proximidad a las áreas urbanas y costeras. Estos resultados se ilustraron para las ciudades de Barranquilla (Figura 12), Santa Marta (Figura 13) y Cartagena (Figura 14).

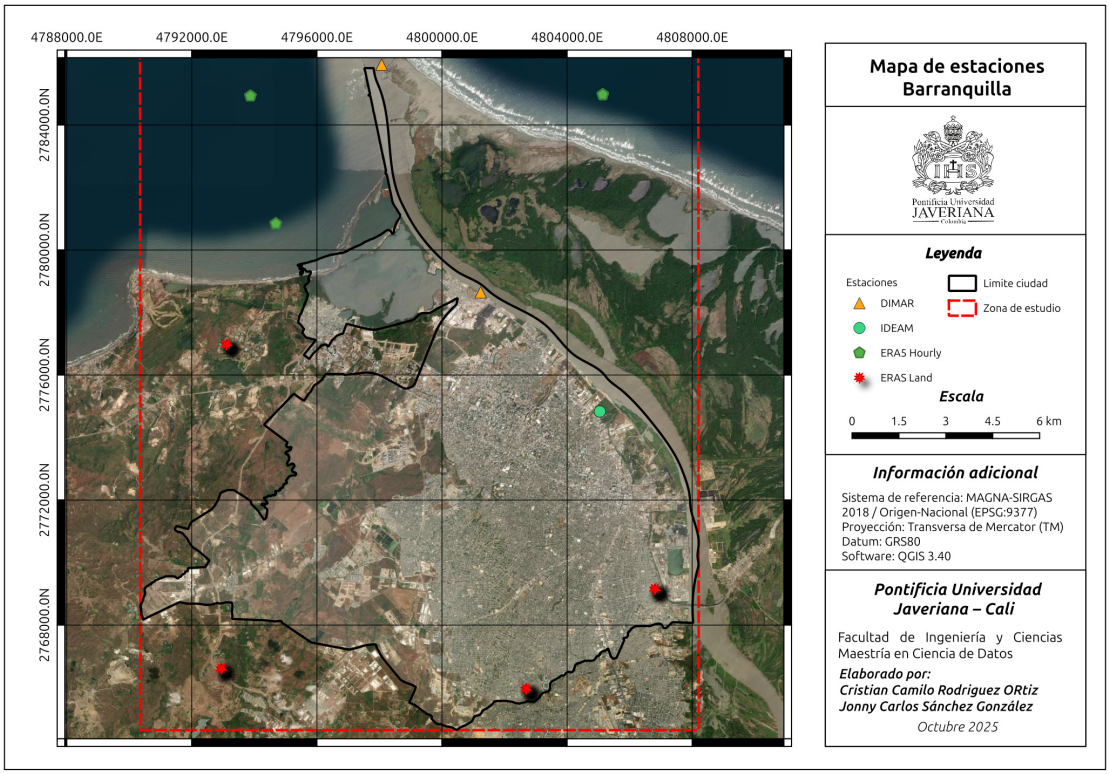


Figura 12: Mapa de estaciones — Barranquilla.

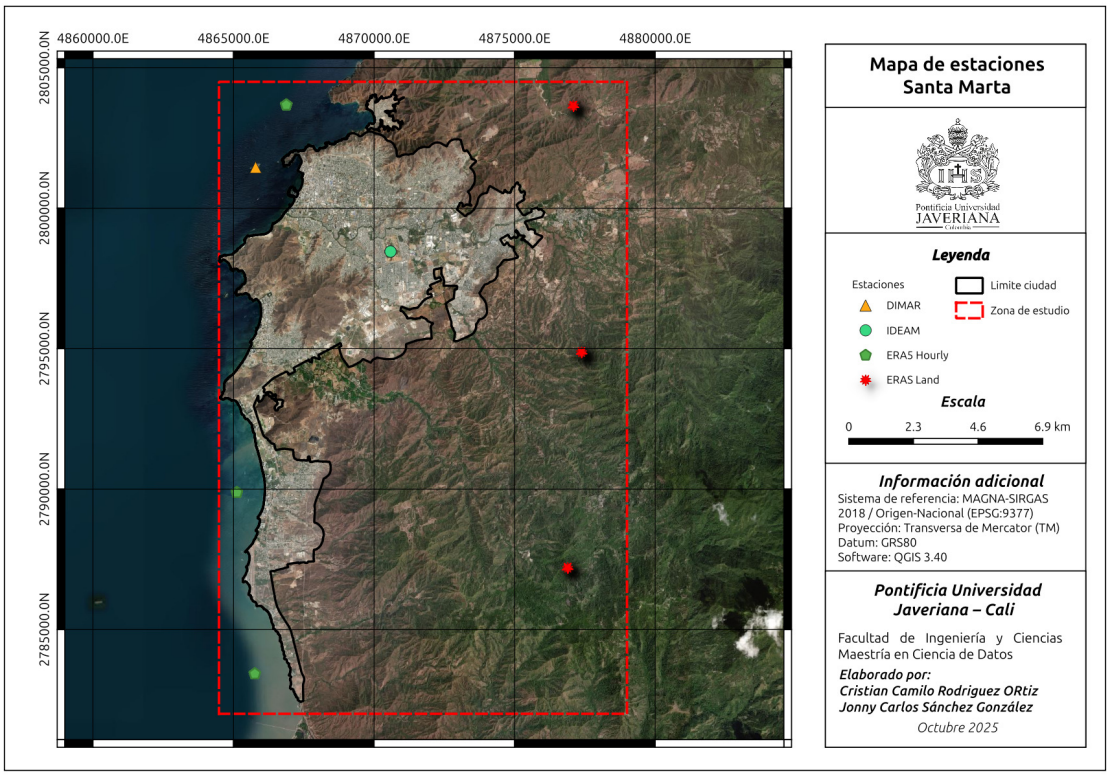


Figura 13: Mapa de estaciones — Santa Marta.

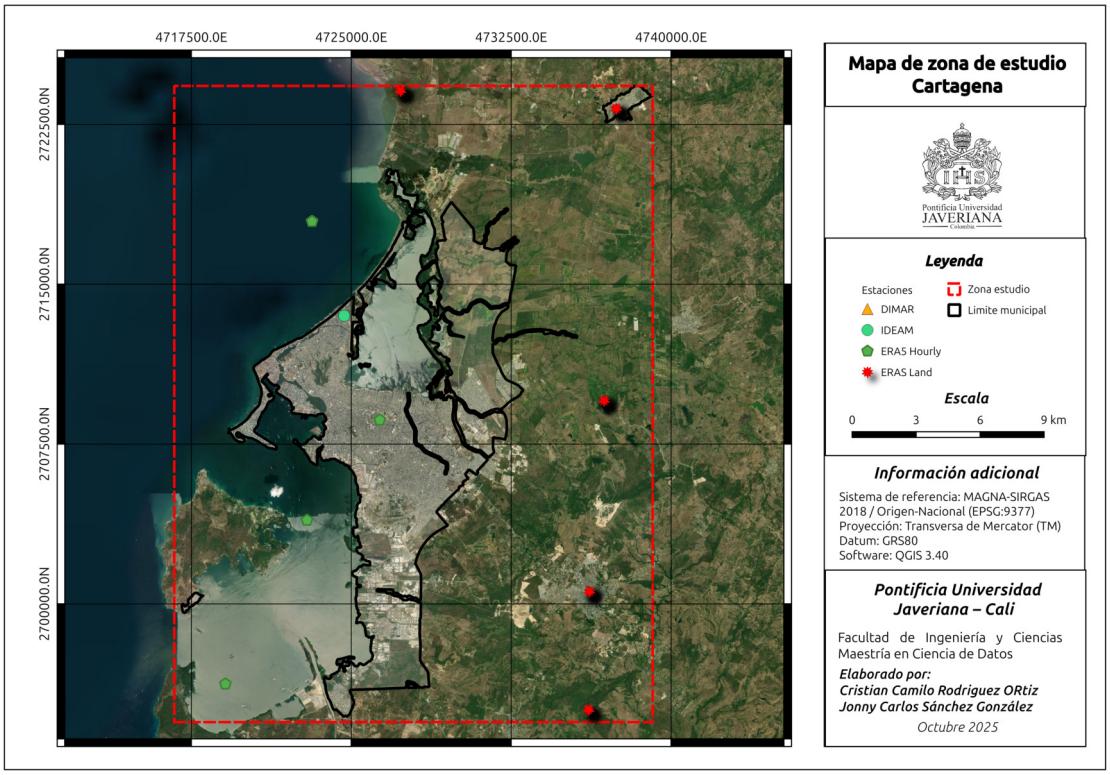


Figura 14: Mapa de estaciones — Cartagena.

5. Procesamiento de datos y creación de modelos.

El presente capítulo detalla las etapas de procesamiento y modelamiento dirigidas a la generación de productos derivados y modelos predictivos. Se presenta el cálculo de los índices espectrales *NDVI*, *NDBI* y *NDWI*, así como la estimación de variables biofísicas como la *emisividad superficial (LSE)* y la *temperatura de la superficie terrestre (LST)*. Adicionalmente, se desarrollaron modelos supervisados de clasificación de coberturas y estimación de la temperatura, basados en algoritmos de *machine learning* y *deep learning*. La implementación de estos modelos, en conjunto con el empleo de herramientas de código abierto como *QGIS*, *PyQGIS* y *scikit-learn*, permitió la automatización del flujo de trabajo y la garantía de la reproducibilidad del proceso.

5.1. Cálculo de índices espectrales.

5.1.1. Cálculo del índice de vegetación NDVI.

Como parte del conjunto de insumos requeridos para la estimación de la temperatura, se realizó el cálculo del Índice de Vegetación de Diferencia Normalizada (NDVI). Este índice espectral es ampliamente utilizado en percepción remota para evaluar y monitorear la presencia, salud y densidad de vegetación en entornos urbanos y rurales [52]. El NDVI destaca la actividad fotosintética al aprovechar las diferencias significativas entre la reflectancia de la vegetación en la banda del infrarrojo cercano (NIR) y la banda del rojo (RED).

El cálculo del NDVI se llevó a cabo mediante la ecuación (1), adaptada específicamente a las bandas multiespectrales *SR_B4* (rojo) y *SR_B5* (infrarrojo cercano), provenientes de imágenes satelitales. El procesamiento fue automatizado mediante un script en *PyQGIS*, utilizando el algoritmo *gdal:rastercalculator* dentro del entorno de QGIS.

En la Figura 15 se presenta el diagrama que resume la secuencia lógica implementada en el script.

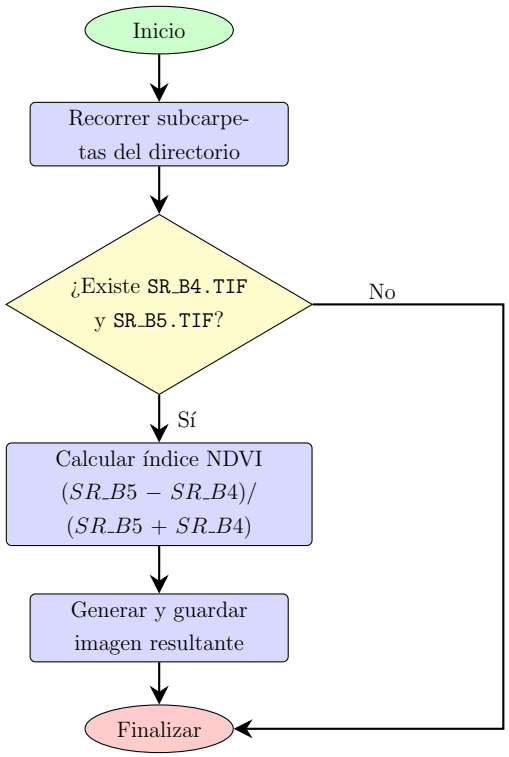
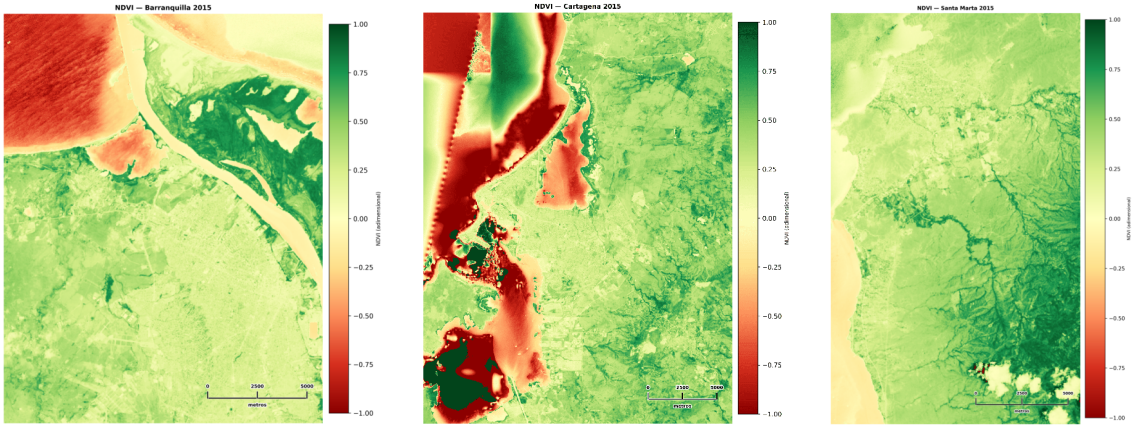


Figura 15: Diagrama del proceso para el cálculo del índice NDVI.

El script desarrollado para el cálculo del NDVI de las imágenes satelitales se encuentra disponible públicamente en el repositorio de GitHub del proyecto `https://github.com/jonnykewin/tesis_MCD/blob/main/7_index_lst_rgb.py`

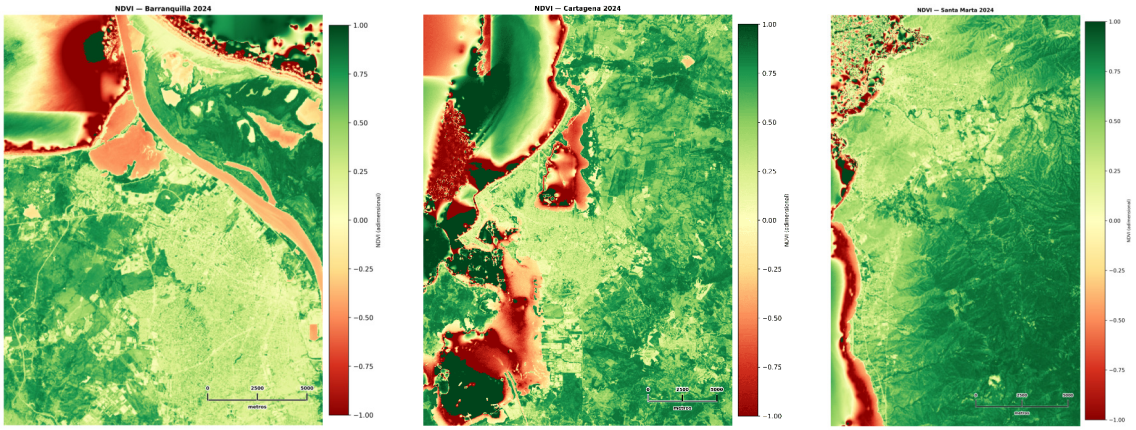
A continuación, se presentan las imágenes que comparan el índice de vegetación de diferencia normalizada (NDVI) correspondiente a los años 2015 (Figura 16) y 2024 (Figura 17) en las ciudades de Barranquilla, Cartagena y Santa Marta. Para la visualización de los resultados, se aplicó un gradiente cromático continuo que varía desde el rojo oscuro hasta el verde intenso, permitiendo una interpretación visual clara de los gradientes de vegetación.

En esta codificación, los valores altos ($\approx +1$), correspondientes a áreas con alta densidad vegetal y elevada actividad fotosintética, se muestran en tonos verde oscuro; los valores intermedios, próximos a 0, se visualizan en tonos amarillos y verde claro, representando zonas con vegetación escasa o cobertura mixta; mientras que los valores bajos o negativos (≈ -1), característicos de áreas urbanas, cuerpos de agua o suelos desnudos, se representan en tonalidades rojizas.



(a) Barranquilla (b) Cartagena (c) Santa Marta

Figura 16: NDVI en 2015 para Barranquilla, Cartagena y Santa Marta.



(a) Barranquilla (b) Cartagena (c) Santa Marta

Figura 17: NDVI en 2024 para Barranquilla, Cartagena y Santa Marta.

5.1.2. Cálculo del Índice de Áreas Construidas NDBI.

Con el objetivo de identificar superficies urbanizadas e impermeables en el área de estudio, se calculó el Índice de Diferencia Normalizada de Áreas Construidas (NDBI). Este índice espectral es ampliamente utilizado para detectar infraestructura urbana como concreto, asfalto y edificaciones, al contrastar zonas construidas frente a áreas vegetadas o naturales [53].

El cálculo del NDBI se realizó mediante la ecuación (2), utilizando las bandas multiespectrales *SR_B6* (infrarrojo de onda corta – SWIR) y *SR_B5* (infrarrojo cercano – NIR) de las imágenes. El proceso fue automatizado mediante un script en *PyQGIS*, aplicando el algoritmo *gdal:rastercalculator* dentro del entorno de QGIS.

En la Figura 18 se presenta el diagrama que resume la secuencia lógica implementada en el script.

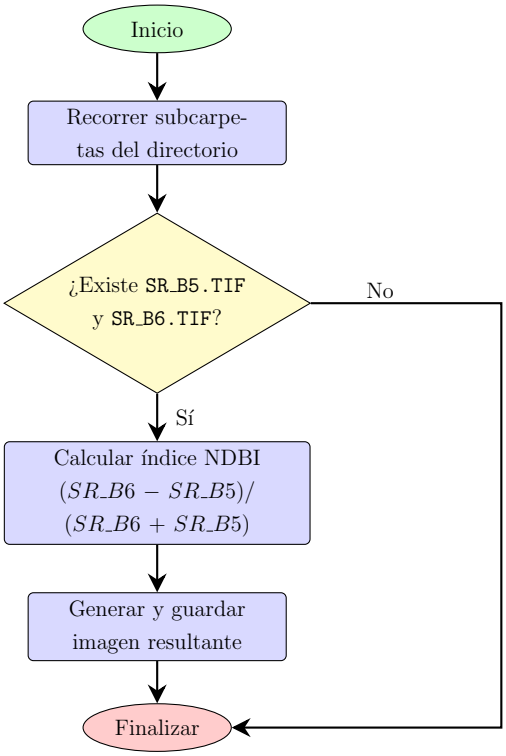
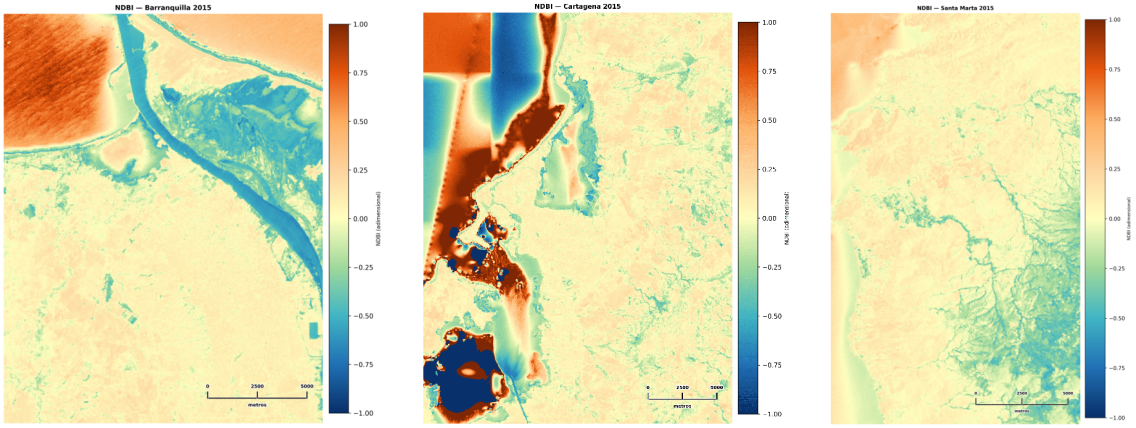


Figura 18: Diagrama del proceso para el cálculo del índice NDBI.

El script desarrollado para el cálculo del NDBI de las imágenes satelitales se encuentra disponible públicamente en el repositorio de GitHub del proyecto `https://github.com/jonnykewin/tesis_MCD/blob/main/7_index_lst_rgb.py`

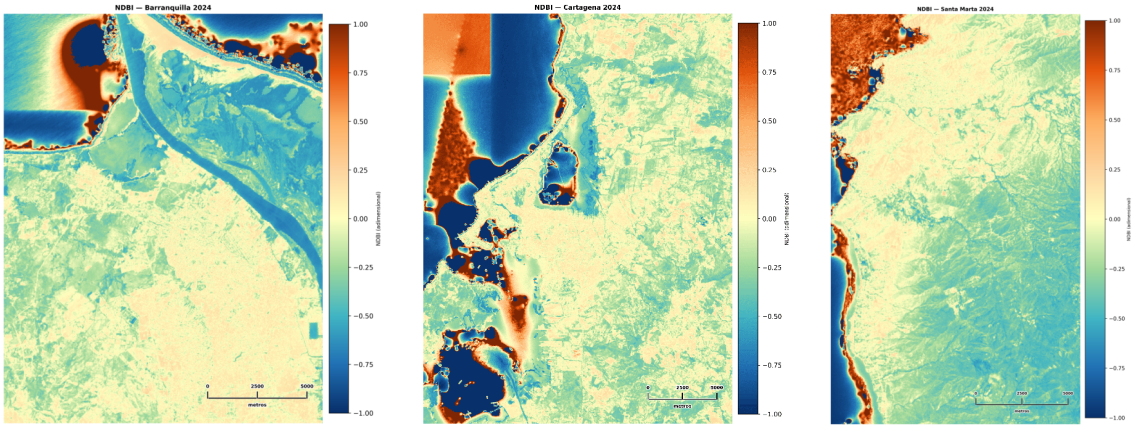
A continuación se presentan las imágenes que comparan el NDBI correspondiente a los años 2015 (Figura 19) y 2024 (Figura 20) en las ciudades de Barranquilla, Cartagena y Santa Marta. Para la visualización de los resultados, se aplicó un gradiente cromático continuo que varía desde azul oscuro hasta naranja intenso, facilitando la interpretación de los gradientes de urbanización.

En esta codificación, los valores altos ($\approx +1$), correspondientes a zonas con alta densidad de superficies construidas e impermeables como edificaciones, vías y pavimentos, se muestran en tonalidades naranjas y ocres; los valores intermedios, cercanos a 0, en tonos verdosos o amarillentos, indican áreas de transición entre suelo natural y urbano; mientras que los valores bajos o negativos (≈ -1), asociados a vegetación densa o cuerpos de agua, se representan en tonos azul oscuro.



(a) Barranquilla (b) Cartagena (c) Santa Marta

Figura 19: Índice de Construcción Normalizado (NDBI) en 2015 para Barranquilla, Cartagena y Santa Marta.



(a) Barranquilla (b) Cartagena (c) Santa Marta

Figura 20: Índice de Construcción Normalizado (NDBI) en 2024 para Barranquilla, Cartagena y Santa Marta.

5.1.3. Cálculo del índice de humedad NDWI.

Con el objetivo de identificar cuerpos de agua y zonas con alto contenido de humedad superficial en el área de estudio, se calculó el Índice de Diferencia Normalizada de Humedad (NDWI). Este índice espectral es ampliamente utilizado en percepción remota para resaltar masas de agua, como ríos, lagos y humedales, mediante la relación entre la reflectancia en las bandas del verde y el infrarrojo cercano [54].

El cálculo del NDWI se efectuó utilizando la ecuación (3), empleando las bandas multiespectrales *SR_B3* (verde) y *SR_B5* (infrarrojo cercano – NIR) de las imágenes satelitales. La implementación fue automatizada a través de un script en *PyQGIS*,

haciendo uso del algoritmo *gdal:rastercalculator* dentro del entorno de procesamiento de QGIS.

En la Figura 21 se presenta el diagrama que resume la secuencia lógica implementada en el script.

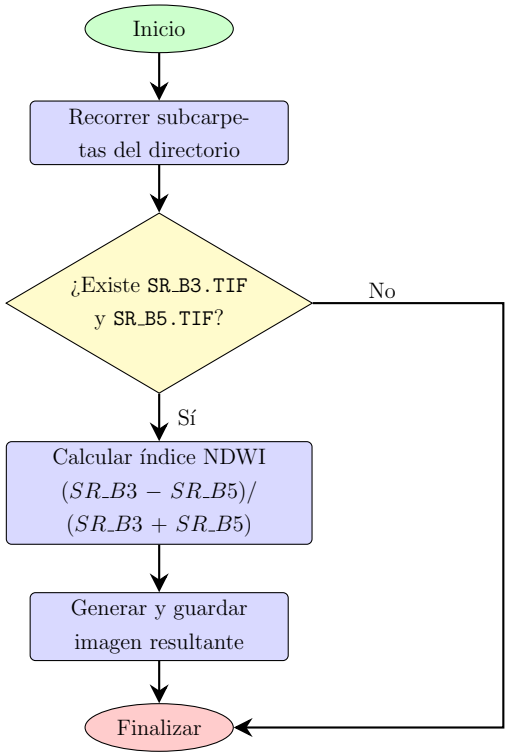


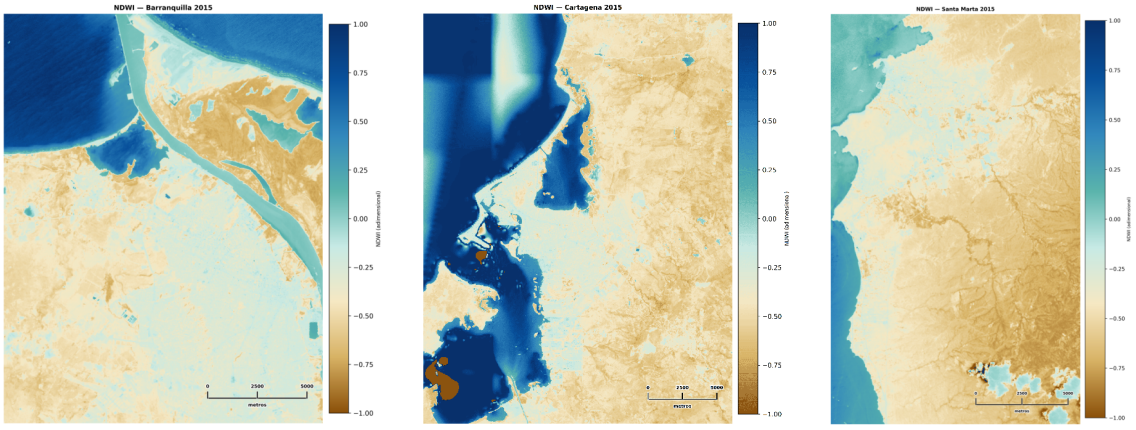
Figura 21: Diagrama del proceso para el cálculo del índice NDWI.

El script desarrollado para el cálculo del NDWI de las imágenes satelitales se encuentra disponible públicamente en el repositorio de GitHub del proyecto `https://github.com/jonnykewin/tesis_MCD/blob/main/7_index_lst_rgb.py`

A continuación se presentan las imágenes que comparan el NDWI correspondiente a los años 2015 (Figura 22) y 2024 (Figura 23) en las ciudades de Barranquilla, Cartagena y Santa Marta. Para la visualización de los resultados, se aplicó un gradiente cromático continuo que varía desde el marrón hasta el azul, facilitando la interpretación de las zonas con alta humedad superficial o presencia de agua.

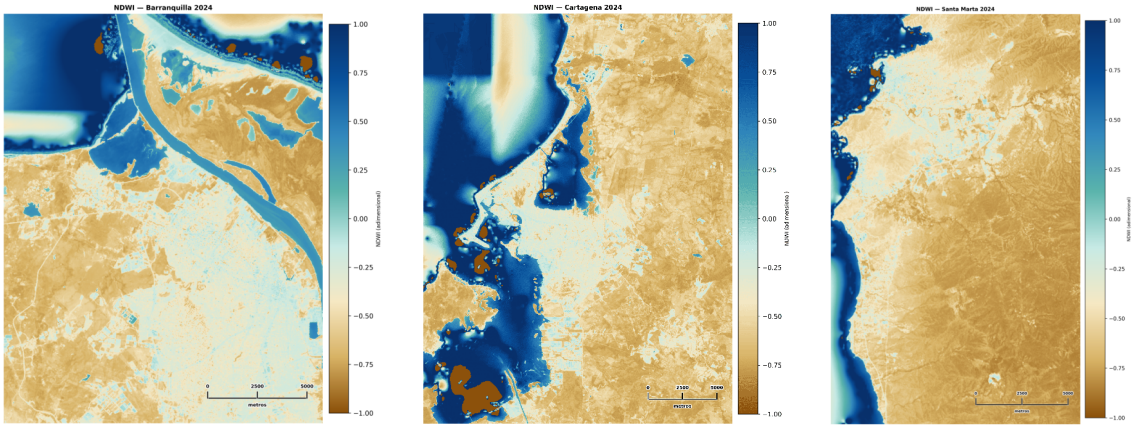
Los valores negativos (≈ -1), correspondientes a superficies secas, suelo desnudo o áreas urbanizadas, se representan en marrones claros; los valores intermedios, próximos a 0, se muestran en cian o verde agua, indicando condiciones de humedad moderada o mezcla de coberturas; mientras que los valores positivos (hasta $+1$), característicos de cuerpos de agua o zonas saturadas, se expresan en azules cada

vez más oscuros, alcanzando su máxima intensidad en aguas abiertas. Este esquema cromático facilita la interpretación espacial del contenido hídrico y la identificación de gradientes desde áreas secas hasta regiones con mayor disponibilidad de agua.



(a) Barranquilla (b) Cartagena (c) Santa Marta

Figura 22: Índice de Diferencia Normalizada de Agua (NDWI) en 2015 para Barranquilla, Cartagena y Santa Marta.



(a) Barranquilla (b) Cartagena (c) Santa Marta

Figura 23: Índice de Diferencia Normalizada de Agua (NDWI) en 2024 para Barranquilla, Cartagena y Santa Marta.

5.2. Cálculo de LST y LSE.

5.2.1. Cálculo de la Emisividad de la Superficie Terrestre (LSE).

La emisividad superficial (*Land Surface Emissivity (LSE)*) se obtuvo a partir de la banda *ST_EMIS*, la cual es proporcionada radiométricamente corregida y acompañada por su factor de escala en los metadatos del producto Landsat. Esta variable

representa la capacidad de las superficies terrestres para emitir energía térmica y constituye un parámetro esencial en la estimación precisa de la temperatura de la superficie terrestre (LST) [55].

La banda *ST_EMIS* fue procesada mediante su reescalación de niveles digitales a unidades físicas, aplicando el factor de escala provisto por el producto:

$$LSE = ST\_EMIS \times 0.0001 \tag{10}$$

donde *LSE* corresponde a la emisividad adimensional. Este procedimiento fue automatizado utilizando el algoritmo *gdal:rastercalculator* dentro del entorno de procesamiento de QGIS, ejecutando la expresión sobre cada píxel de la banda *ST_EMIS*.

En la Figura 24 se presenta el diagrama que resume la secuencia lógica implementada en el script.

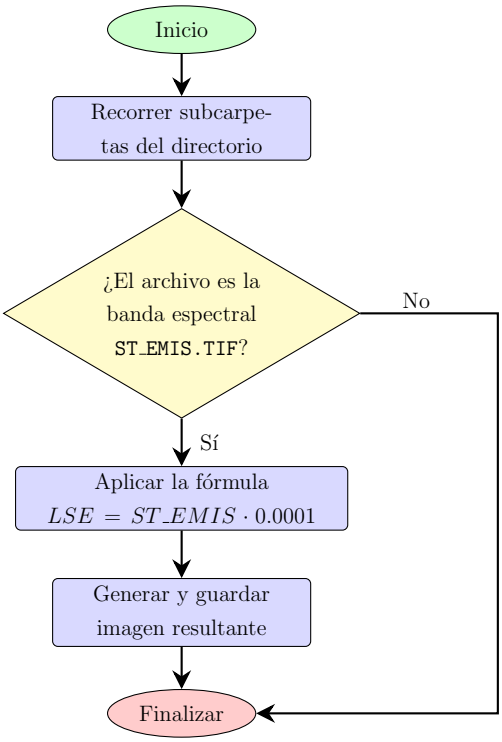
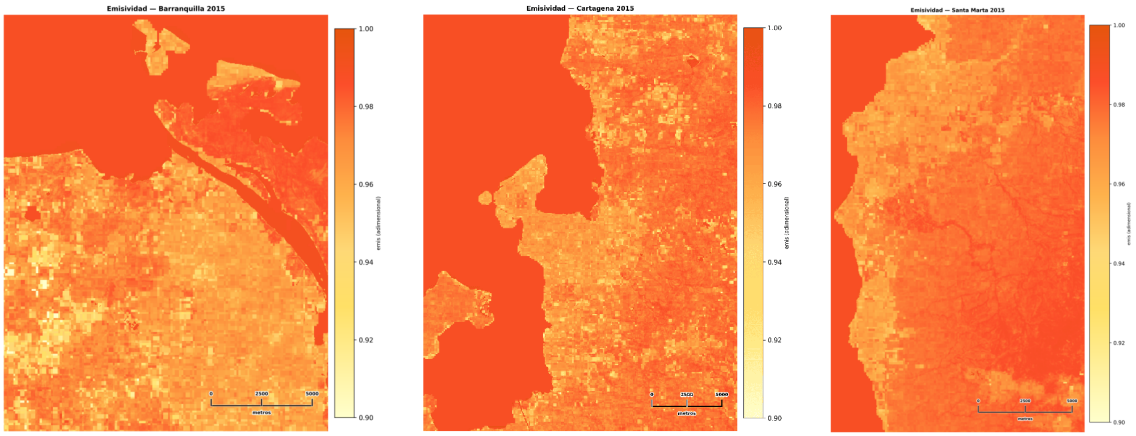


Figura 24: Diagrama del proceso para el cálculo del LSE.

El script desarrollado para el cálculo del LSE de las imágenes satelitales se encuentra disponible públicamente en el repositorio de GitHub del proyecto `https://github.com/jonnykewin/tesis_MCD/blob/main/3_scale_bands.py`

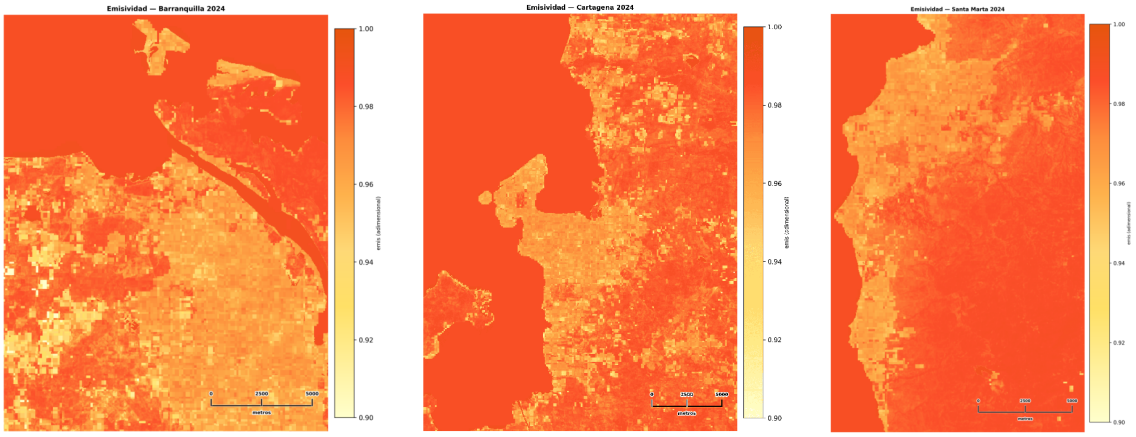
A continuación, se presentan las imágenes que muestran la emisividad superficial terrestre (LSE) para los años 2015 (Figura 25) y 2024 (Figura 26) en las ciudades de

Barranquilla, Cartagena y Santa Marta. Estas representaciones permiten observar la distribución espacial de la emisividad, la cual está relacionada con la radiación térmica emitida por las coberturas terrestres. Los valores más altos de emisividad, cercanos a 1, suelen asociarse con coberturas vegetales y cuerpos de agua, mientras que los valores más bajos, próximos a 0.90, se vinculan con suelos desnudos, áreas pavimentadas o superficies artificiales.



(a) Barranquilla (b) Cartagena (c) Santa Marta

Figura 25: Emisividad superficial terrestre (LSE) en 2015 para Barranquilla, Cartagena y Santa Marta.



(a) Barranquilla (b) Cartagena (c) Santa Marta

Figura 26: Emisividad superficial terrestre (LSE) en 2024 para Barranquilla, Cartagena y Santa Marta.

5.2.2. Cálculo de la Temperatura de la Superficie Terrestre (LST).

La *Temperatura de la Superficie Terrestre* (LST) se obtuvo a partir de la banda térmica *ST_B10* de las imágenes satelitales. Esta banda ya se encuentra corregida

radiométricamente y expresada en unidades de Kelvin, conforme a los estándares definidos por el Servicio Geológico de los Estados Unidos (USGS).

Para facilitar la interpretación y el análisis espacial de los resultados, los valores de temperatura fueron convertidos de Kelvin a grados Celsius mediante la siguiente fórmula:

$$\text{LST } (°\text{C}) = \text{LST } (K) - 273.15 \tag{11}$$

Este cálculo fue automatizado utilizando el algoritmo *gdal:rastercalculator* dentro del entorno de procesamiento de QGIS. El producto final representa la temperatura superficial en grados Celsius, constituyendo un insumo clave para la detección y análisis de islas de calor urbanas, dada su relevancia en la cuantificación del balance energía–superficie y los efectos de urbanización [56].

En la Figura 27 se presenta el diagrama que resume la secuencia lógica implementada en el script.

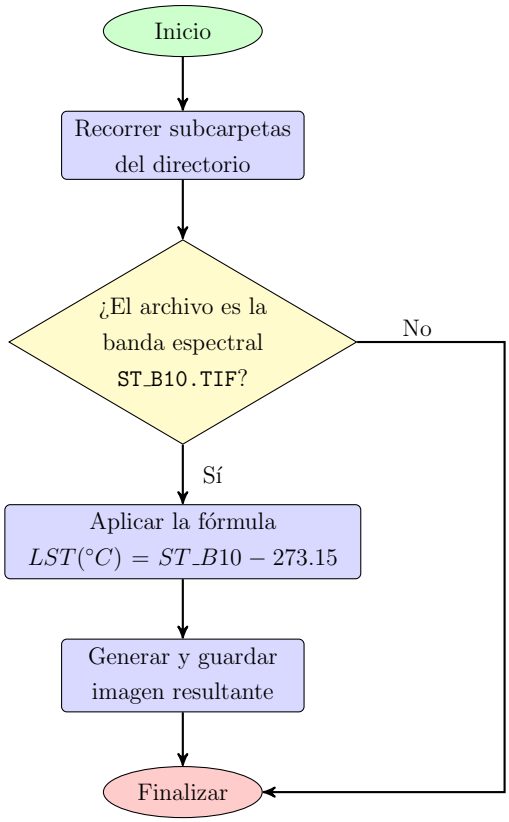


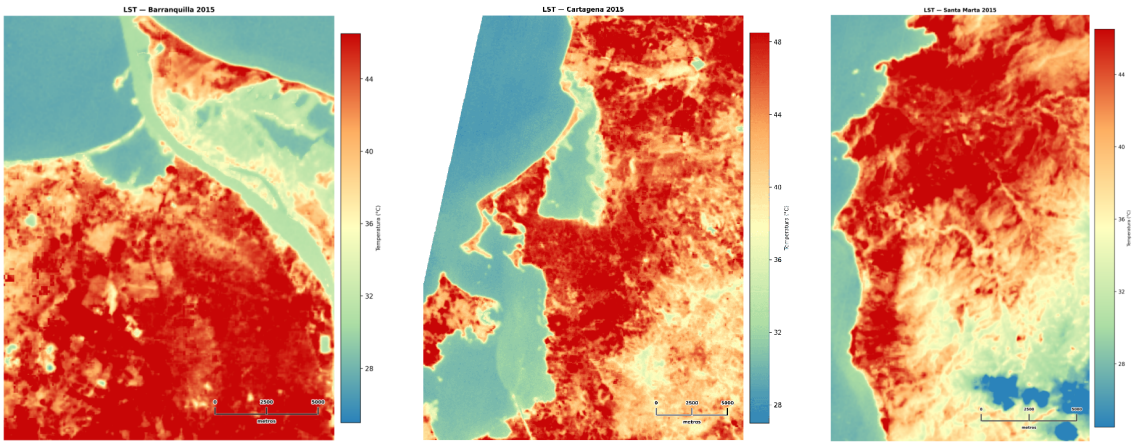
Figura 27: Diagrama del proceso para el cálculo de la Temperatura de la Superficie Terrestre (LST).

El script desarrollado para el cálculo del LST de las imágenes satelitales se encuentra disponible públicamente en el repositorio de GitHub del proyecto `https://github.com/jonnykewin/tesis_MCD/blob/main/7_index_lst_rgb.py`

A continuación, se presentan las imágenes que comparan la temperatura superficial terrestre (LST) correspondiente a los años 2015 (Figura 28) y 2024 (Figura 29) en las ciudades de Barranquilla, Cartagena y Santa Marta. Estas representaciones permiten observar las variaciones espaciales y temporales asociadas a los procesos de urbanización y a la presencia de coberturas vegetales o cuerpos de agua.

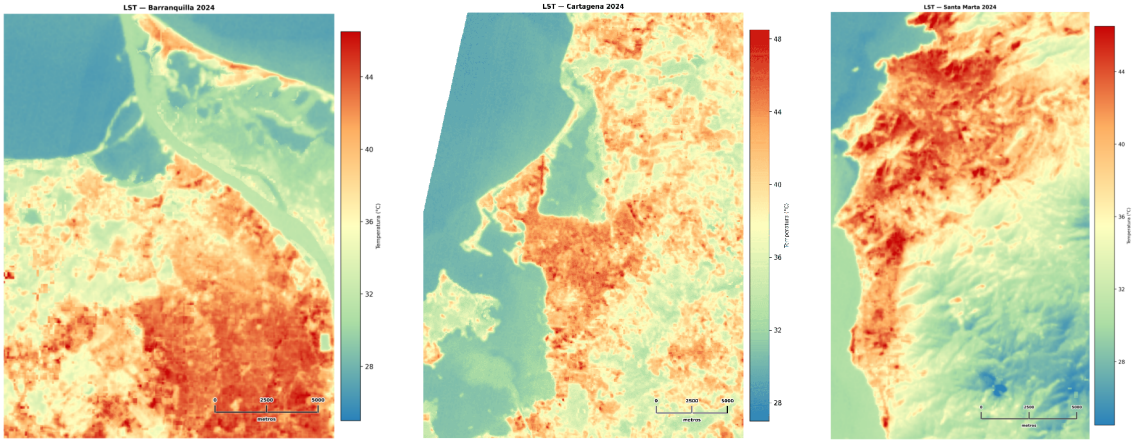
La visualización del LST emplea una paleta cromática continua que facilita la interpretación de los gradientes térmicos en el territorio. Los valores más bajos de temperatura, generalmente inferiores a los 28 °C, se asocian con cuerpos de agua, zonas costeras y áreas vegetadas, y se representan en tonos azul oscuro y cian. Las temperaturas intermedias, entre 30 y 36 °C, se visualizan en gamas de verde y amarillo claro, indicando transiciones entre superficies naturales y áreas parcialmente urbanizadas. Por su parte, las temperaturas más altas, superiores a 40 °C, se muestran en colores naranja y rojo intenso, característicos de zonas urbanas densamente edificadas, pavimentos y suelos expuestos a alta radiación solar.

Esta representación cromática permite identificar de forma rápida y visual las áreas con mayor acumulación térmica, facilitando la detección de islas de calor urbanas y el análisis comparativo de su evolución entre los periodos analizados.



(a) Barranquilla (b) Cartagena (c) Santa Marta

Figura 28: Temperatura superficial terrestre (LST) en 2015 para Barranquilla, Cartagena y Santa Marta.



(a) Barranquilla (b) Cartagena (c) Santa Marta

Figura 29: Temperatura superficial terrestre (LST) en 2024 para Barranquilla, Cartagena y Santa Marta.

5.3. Construcción de datos etiquetados.

Como insumo fundamental para la etapa de entrenamiento de modelos supervisados, se seleccionó una nube de puntos georreferenciados etiquetados, clasificados en seis coberturas del suelo: *vegetación*, *construcción*, *suelo desnudo*, *agua dulce*, *agua salada* y *nubes*. La asignación de clases se realizó por fotointerpretación sobre la composiciones RGB, a través del entorno gráfico de QGIS, apoyándose en criterios visuales y espectrales.

Para asegurar la homogeneidad del proceso, se definió un estándar en la estructura de carpetas, nombres de archivos y configuración de proyectos en QGIS, el cual fue replicado de forma sistemática para cada ciudad (Barranquilla, Cartagena y Santa Marta) y para los años 2015, 2018, 2020, 2022 y 2024. Cada conjunto de puntos fue almacenado en archivos GeoPackage (`.gpkg`), constituyendo así una base de datos espacial etiquetada por ciudad y periodo temporal. Un ejemplo de este proceso de etiquetado por fotointerpretación puede observarse en la Figura 30, donde se muestran los puntos clasificados según las coberturas del suelo definidas.

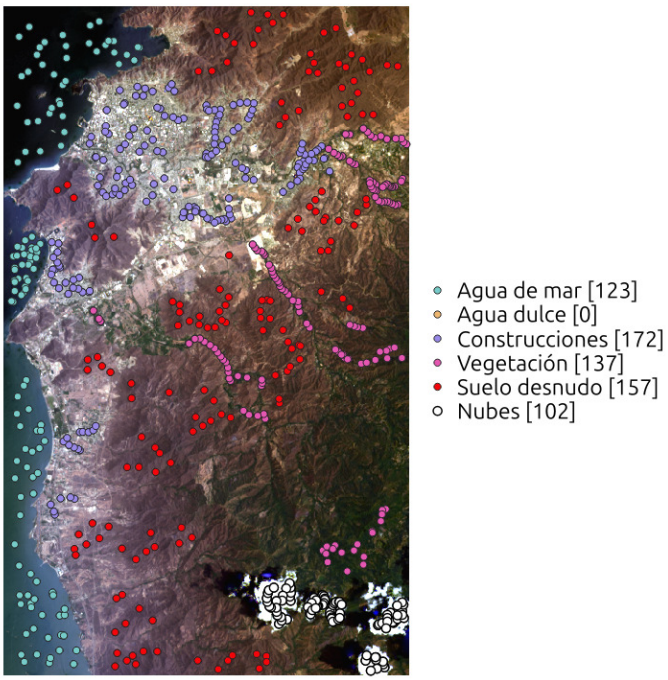


Figura 30: Ejemplo de puntos etiquetados por fotointerpretación para cada una de las coberturas del suelo definidas.

Una vez consolidados los puntos de muestreo, se procedió a la extracción de los valores espectrales y térmicos mediante el algoritmo `Sample Raster Values` de QGIS. Este procedimiento se ejecutó de forma automatizada y secuencial, replicando la misma lógica en todas las combinaciones espacio-temporales.

Para garantizar un flujo reproducible, se implementó un script en Python denominado `prueba.py`, basado en el módulo `processing` de QGIS y en el uso de herramientas auxiliares del sistema. El script fue diseñado para recorrer automáticamente todas las subcarpetas del directorio de trabajo, identificar aquellas que contienen el archivo base `datos.gpkg`, y aplicar la herramienta `native:rastersampling` sobre una lista predefinida de insumos raster. Dicha lista incluía las bandas SR_B2 a SR_B8, el producto térmico LST y los índices espectrales NDVI, NDBI y NDWI.

A cada raster se le asignó un prefijo específico para diferenciar las columnas generadas en la tabla de atributos, y la salida de cada iteración se almacenó como un archivo GeoPackage intermedio. Estos archivos se generaron de forma progresiva, acumulando en cada paso los valores espectrales o térmicos correspondientes, hasta consolidar un único archivo final denominado `datos_final.gpkg`. Este archivo contiene la totalidad de los valores extraídos de los productos raster asociados a cada punto de muestreo. Además, el script implementa una rutina de limpieza automática que elimina todos los archivos temporales intermedios, preservando únicamente el

archivo enriquecido final, optimizando así el almacenamiento y la organización del proyecto.

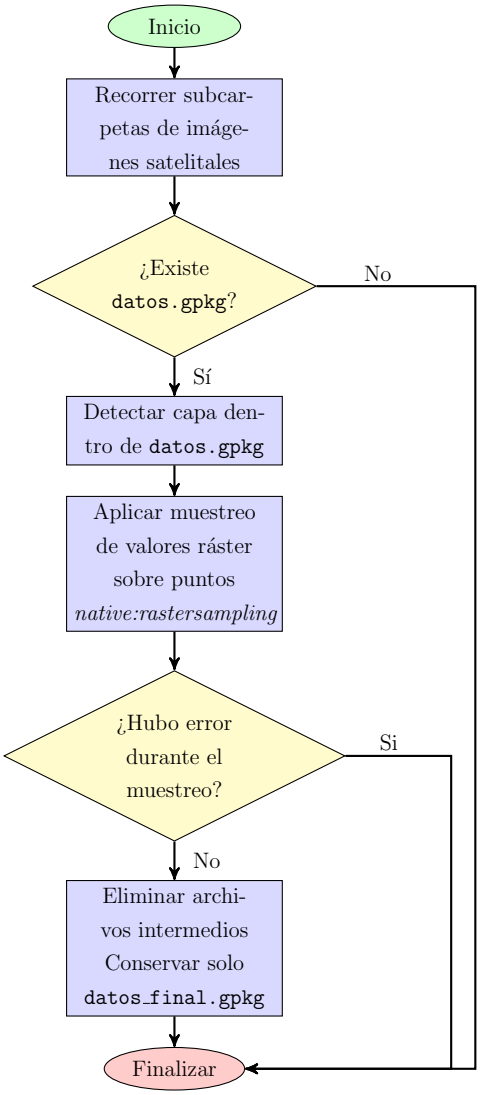


Figura 31: Esquema del proceso automatizado de muestreo de valores ráster.

Una vez ejecutado el script en cada una de las carpetas correspondientes a los proyectos organizados por ciudad y temporalidad, se obtuvo un conjunto estructurado de datos espaciales. Cada archivo consolidado contenía como atributos principales: la clase de cobertura asignada (etiqueta), el año de la imagen procesada, los valores espectrales de las bandas B2 a la B8, la temperatura de la superficie terrestre (LST), los índices espectrales NDVI, NDBI, NDWI y, en algunos casos, metadatos adicionales según el flujo de trabajo implementado.

A continuación, se presenta una visualización del formato de visulización del conjunto de datos correspondiente al año 2015 para la ciudad de Santa Marta. Este ejemplo

permite observar cómo los valores raster fueron asociados a cada punto etiquetado de forma estructurada y espacialmente coherente.

Tabla 5: Visualización del conjunto de datos etiquetado (Santa Marta, 2015).

fid	cobertura	año	lst1	ndvi1	ndbi1	b21	b31	b41	b51	b61	b71	b81	ndwi1
1	Construcciones	2015	44.806	0.242	0.042	0.087	0.141	0.165	0.273	0.297	0.239	0.155	-0.357
2	Construcciones	2015	45.753	0.240	0.065	0.077	0.134	0.164	0.270	0.308	0.244	0.207	-0.394
3	Construcciones	2015	45.479	0.317	0.014	0.059	0.111	0.147	0.284	0.292	0.204	0.154	-0.449

Notas: Estructura de ejemplo de 3 datos

Finalizado el proceso de enriquecimiento, todos los archivos resultantes fueron integrados en una base de datos relacional PostgreSQL con la extensión PostGIS habilitada. La arquitectura de la base se diseñó siguiendo una lógica modular y escalable: se creó un esquema independiente para cada ciudad `barranquilla`, `cartagena` y `santa_marta` dentro de la base de datos `maestria`, y dentro de cada uno se almacenaron las tablas correspondientes a los años procesados: 2015, 2018, 2020, 2022 y 2024. Cada tabla contiene los puntos etiquetados, junto con los valores extraídos de los productos raster y su respectiva geometría espacial.

La carga de estos conjuntos de datos se realizó utilizando el complemento *DB Manager* de QGIS, asegurando la preservación de los atributos y la geometría puntual. Aunque los puntos de muestreo fueron capturados inicialmente bajo el sistema de coordenadas geográficas WGS84 (`EPSG:4326`), se llevó a cabo una transformación previa al ingreso a la base de datos hacia el sistema proyectado oficial colombiano (`EPSG:9377`), siguiendo los lineamientos técnicos del Instituto Geográfico Agustín Codazzi (IGAC). Esta conversión garantizó la compatibilidad con otros insumos vectoriales utilizados en las etapas posteriores de análisis espacial.

Con el objetivo de facilitar el manejo integrado de la información geoespacial, se llevó a cabo la materialización de una tabla final consolidada por ciudad. Esta operación consistió en unificar todas las tablas anuales contenidas dentro de cada esquema de la base de datos PostgreSQL/PostGIS correspondientes a los años 2015, 2018, 2020, 2022 y 2024 mediante una operación `UNION ALL`. El resultado fue la creación de una nueva tabla denominada `datos_final` dentro del esquema de cada ciudad, que agrupa en una sola estructura todos los puntos etiquetados recolectados a lo largo del proyecto.

Esta tabla consolidada sirvió como punto de partida para las siguientes etapas del análisis, permitiendo exportar el conjunto completo de datos en formato `.csv` para su posterior procesamiento en el entorno de Python. La Tabla 6 presenta la cantidad

total de registros integrados en la tabla `datos_final`, distribuidos por ciudad y tipo de cobertura.

Tabla 6: Distribución de muestras por ciudad, tipo de cobertura y total por ciudad.

Ciudad	Cobertura	Número de muestras	Total por ciudad
Cartagena	Agua de mar	552	3 444
	Agua dulce	392	
	Construcciones	954	
	Suelo desnudo	664	
	Vegetación	882	
Barranquilla	Agua de mar	390	2 943
	Agua dulce	544	
	Construcciones	662	
	Suelo desnudo	798	
	Vegetación	549	
Santa Marta	Agua de mar	461	2 865
	Construcciones	636	
	Nubes	469	
	Suelo desnudo	597	
	Vegetación	702	

5.4. Modelo supervisado para la clasificación de coberturas.

Para la clasificación de las coberturas superficiales en el área de estudio, se implementaron algoritmos de *machine learning* (ML) y *deep learning* (DL), utilizando las muestras etiquetadas generadas mediante fotointerpretación sobre las imágenes satelitales.

Variables y preparación de datos. Las variables predictoras empleadas fueron las bandas espectrales *B2, B3, B4, B6*, la *temperatura superficial (LST)*, y los índices espectrales *NDVI* y *NDBI*. La variable objetivo (*label*) se codificó numéricamente mediante *Label Encoding*.

Entrenamiento y validación. Se entrenaron los modelos *Random Forest* [19], *Extra Trees* [20], *Support Vector Machines (SVC)* [22], *Gradient Boosting* [21] y una *Multilayer Perceptron (MLP)* [23]. El ajuste de hiperparámetros se realizó mediante

búsqueda en rejilla (*Grid Search*) combinada con validación cruzada estratificada (*StratifiedKFold*) con $k = 10$ iteraciones, para estimar el desempeño y prevenir el sobreajuste. El conjunto total de datos se dividió en *bloque de entrenamiento (out-of-fold) (70 %)* y *conjunto de prueba (hold-out) (30 %)*, garantizando independencia entre las fases de ajuste y evaluación.

Modelos evaluados. A continuación, se resumen los hiperparámetros óptimos obtenidos para cada modelo:

Tabla 7: Mejores hiperparámetros (según validación cruzada).

Modelo	**Hiperparámetros (CV)**
Random Forest	`max_depth=None, min_samples_leaf=1, min_samples_split=2, n_estimators=200`
Extra Trees	`max_depth=None, min_samples_leaf=1, min_samples_split=2, n_estimators=300`
Gradient Boosting	`learning_rate=0.1, max_depth=2, n_estimators=400`
SVC	`C=2, gamma=scale, kernel=rbf`
MLP	`activation=tanh, alpha=0.01, hidden_layer_sizes=(128,64,32), solver=adam`

Evaluación de desempeño. Una vez definidos los hiperparámetros óptimos, los modelos fueron reentrenados sobre el conjunto de entrenamiento completo y evaluados sobre el bloque de prueba (*hold-out*). Se reportaron las métricas *Accuracy* y *F1-score* tanto para el entrenamiento con validación cruzada (*OOF-CV*) como para la evaluación final en prueba (*TEST*), junto con los tiempos de ejecución, como se muestra en la Tabla 8.

Tabla 8: Desempeño y tiempos por modelo en entrenamiento con validación cruzada (OOF-CV) y en prueba (TEST).

Modelo	**TRAINING (OOF-CV)**			**TEST (hold-out)**		
	Accuracy	F1-score	Time[s]	Accuracy	F1-score	Time[s]
SVC	0.943	0.942	1.066	0.943	0.940	3.827
MLP	0.974	0.974	22.056	0.977	0.977	15.563
Random Forest	0.973	0.976	2.200	0.974	0.976	9.083
Gradient Boosting	0.971	0.973	19.826	0.971	0.973	103.265
Extra Trees	**0.978**	**0.982**	**1.630**	**0.978**	**0.980**	**9.008**

Notas: OOF-CV = estimación *out-of-fold* con validación cruzada; TEST = evaluación final en el conjunto de prueba.

Interpretación de resultados. Los resultados indican que el modelo Extra Trees obtuvo el mejor rendimiento global. Este comportamiento puede atribuirse a su capacidad de manejar grandes volúmenes de variables y su eficiencia en la reducción del sobreajuste, manteniendo a su vez un bajo tiempo de entrenamiento. Esto concuerda con la literatura, donde el algoritmo *Extra Trees* muestra gran estabilidad y capacidad para manejar variables correlacionadas mediante una mayor aleatoriedad en la selección de divisiones [57].

Para complementar el análisis cuantitativo, se presentan las métricas de desempeño del modelo en la Tabla 9 y las matrices de confusión generadas para el modelo Extra Trees en *test (hold-out)* (Figura 32) y *training con validación cruzada (out-of-fold)* (Figura 33). Estas matrices permiten identificar visualmente los patrones de error y las confusiones entre clases, proporcionando una herramienta para interpretar el desempeño del modelo.

Tabla 9: Métricas de desempeño por tipo de cobertura para el modelo entrenado.

Cobertura	Precisión	Recall	F1-score
Agua de mar	0.978	0.988	0.983
Agua dulce	0.979	0.965	0.972
Construcciones	0.983	0.979	0.981
Nubes	1.000	1.000	1.000
Suelo desnudo	0.970	0.965	0.967
Vegetación	0.975	0.984	0.980
Accuracy	0.978	0.978	0.978
Macro avg	0.981	0.980	0.980
Weighted avg	0.978	0.978	0.978

Notas: Las métricas corresponden al desempeño del clasificador sobre el conjunto de prueba.

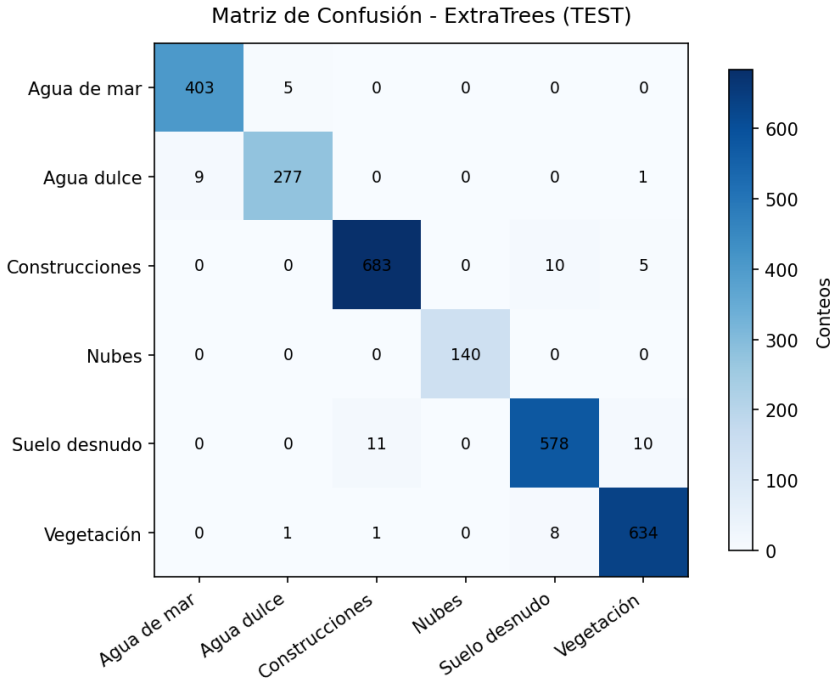


Figura 32: Matriz de confusión del modelo ExtraTrees (Test).

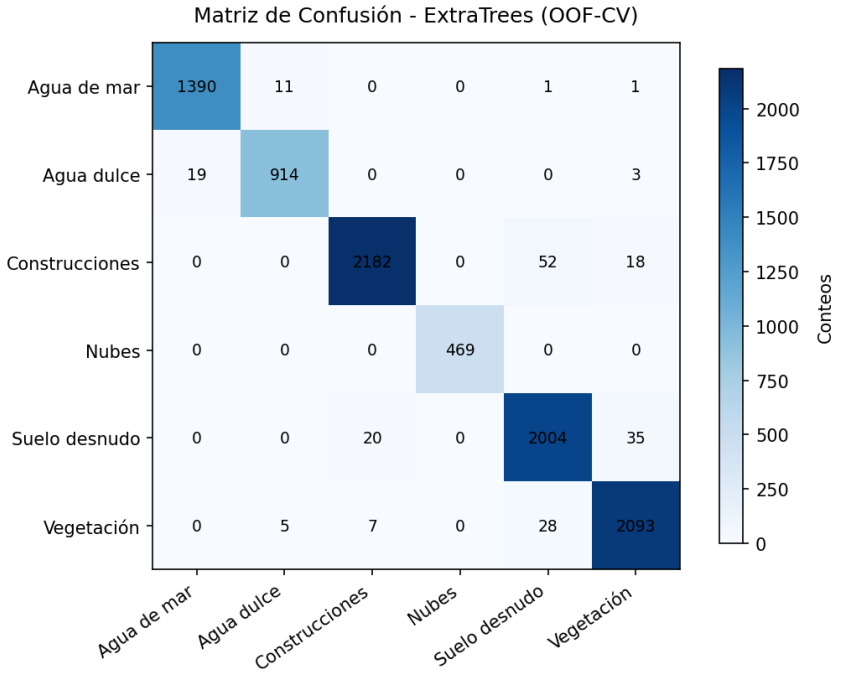


Figura 33: Matriz de confusión del modelo ExtraTrees (Training).

Existe un alto desempeño en todas las clases, con aciertos especialmente claros en *Nubes* y *Construcciones* tanto en los datos de entrenamiento como pruebas. Aun así, se observan confusiones entre pares de coberturas con firmas espectrales cercanas y mezclas de pixeles.

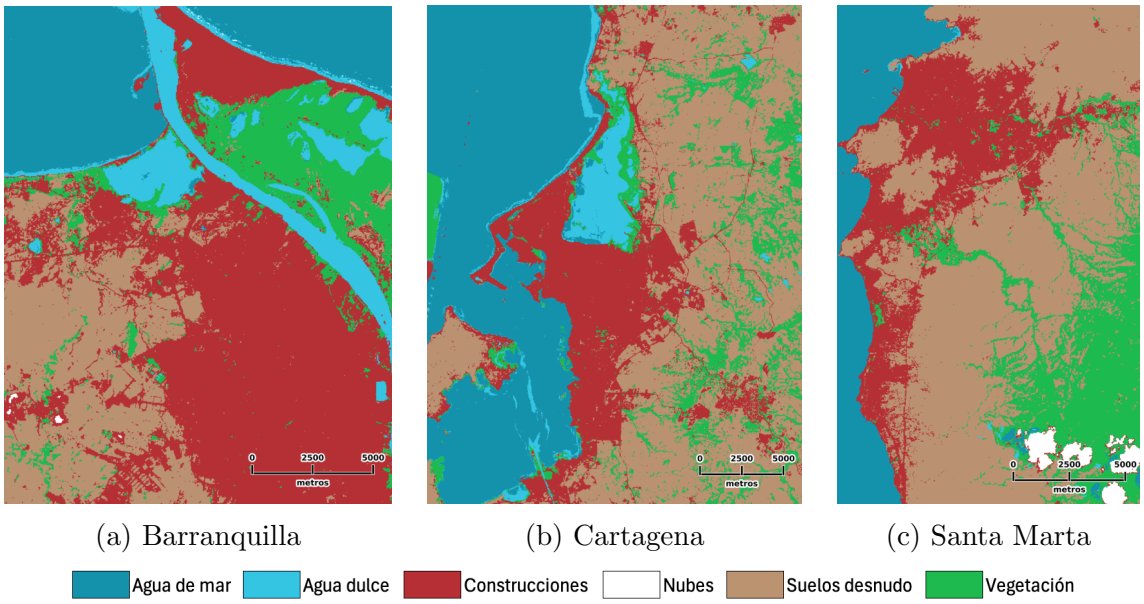
A continuación se resumen los casos más relevantes:

- Agua de mar y Agua dulce. Se observan pocos intercambios entre ambas clases. Este comportamiento es compatible con escenarios de turbidez o presencia de solutos que alteran la reflectancia del agua continental, así como con píxeles que combinan agua y borde costero. En tales situaciones, la separación basada únicamente en firmas espectrales puede verse afectada.
- Agua de mar y Agua dulce. Se observan pocos intercambios entre ambas clases. Este comportamiento es compatible con escenarios de turbidez o presencia de solutos que alteran la reflectancia del agua continental, así como con píxeles que combinan agua y borde costero. En tales situaciones, la separación basada únicamente en firmas espectrales puede verse afectada.
- Construcciones frente a Suelo desnudo y Vegetación. Una parte de los píxeles clasificados como *Construcciones* se asigna incorrectamente a *Suelo desnudo* o *Vegetación*. La confusión con *Suelo desnudo* ocurre con superficies claras de origen humano, como explanadas, terraplenes o cubiertas reflectantes, que

se asemejan a suelos expuestos. La confusión con *Vegetación* puede deberse a techos con vegetación, materiales envejecidos o píxeles que contienen pequeñas áreas verdes adyacentes.

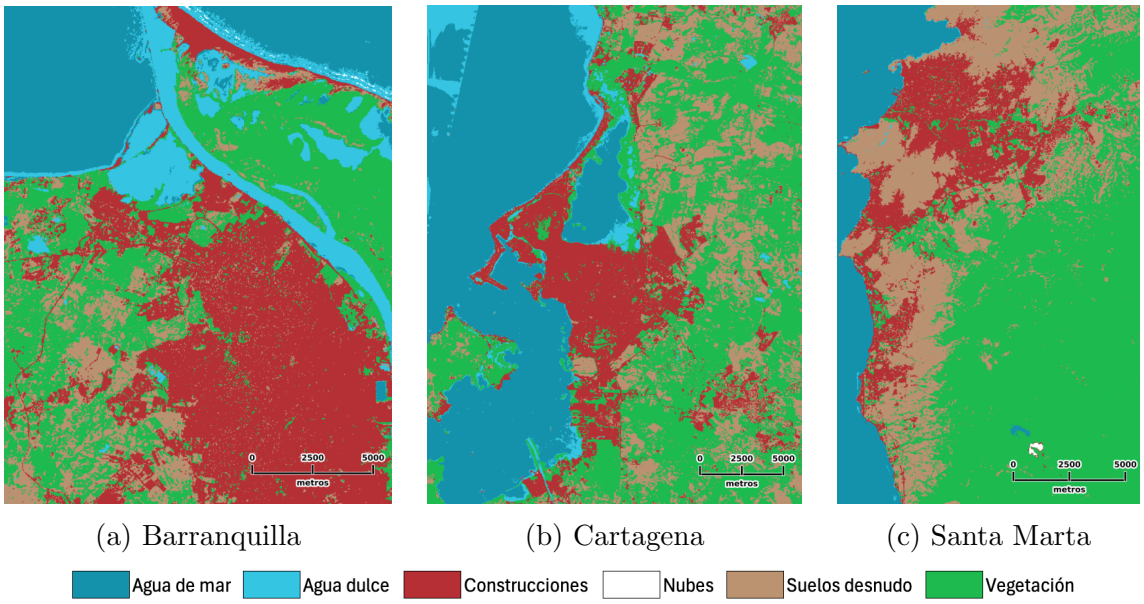
- Suelo desnudo frente a Construcciones y Vegetación. Parte de *Suelo desnudo* se categoriza como *Construcciones* o *Vegetación*. En el primer caso, suelos compactados pueden mostrar firmas similares a superficies urbanas. En el segundo, coberturas escasas, como pastizales dispersos o cultivos con suelo visible, aumentan moderadamente la reflectancia en el infrarrojo cercano, acercándose a patrones vegetales.
- Vegetación frente a Suelo desnudo y Agua dulce. Las confusiones desde *Vegetación* hacia *Suelo desnudo* suelen concentrarse en áreas agrícolas con hileras separadas o en coberturas discontinuas; hacia *Agua dulce* pueden originarse en sectores ribereños o con vegetación flotante, donde el píxel integra componentes de ambas clases.

Las confusiones observadas fueron típicas en teledetección a la resolución espacial empleada y reflejaron, principalmente, mezcla subpíxel y similitud espectral [58]. A continuación, se presentan las clasificaciones de coberturas terrestres correspondientes a los años 2015 (Figura 34) y 2024 (Figura 35) en las ciudades de Barranquilla, Cartagena y Santa Marta, utilizando el algoritmo *ExtraTrees*.



(a) Barranquilla (b) Cartagena (c) Santa Marta

Figura 34: Clasificación de coberturas en 2015 para Barranquilla, Cartagena y Santa Marta.



(a) Barranquilla (b) Cartagena (c) Santa Marta

Figura 35: Clasificación de coberturas en 2024 para Barranquilla, Cartagena y Santa Marta.

5.5. Mapas de densidad de cobertura.

Con el objetivo de caracterizar la configuración espacial de las coberturas clasificadas y reconocer patrones de concentración o dispersión, se generaron *mapas de densidad* para cada cobertura partir de la clasificación obtenido mediante el modelo *Extra Trees*.

Este procedimiento permitió cuantificar, para cada píxel (i, j), la proporción de celdas pertenecientes a la misma clase dentro de una vecindad cuadrada de 5×5 píxeles, lo que equivale a un análisis de vecindad local.

El flujo de procesamiento se ejecutó de forma automatizada mediante un script en *PyQGIS* integrado al entorno de QGIS, utilizando *GDAL*, así: El procedimiento constó de las siguientes etapas principales:

- Máscaras binarias por clase. A partir del raster de clasificación se derivó, para cada cobertura, un raster binario, esto produce $cobertura \in \{0, 1\}$.
- Agregación local (promedio). Sobre cada cobertura se aplicó la ventana móvil de $5*5$ calculando el promedio de la vecidad de las coberturas sobre cada pixel (i, j) este ultimo paso genera el *mapa de densidad por cobertura.*

Los mapas resultantes permitieron identificar zonas de alta continuidad espacial (valores próximos a 1) y áreas de transición o borde (valores intermedios o bajos),

y aportaron información complementaria para la modelación de temperatura y el análisis espacial de las islas de calor urbanas. La Figura 36 presentó el flujo general implementado para la generación de los mapas de densidad de coberturas.

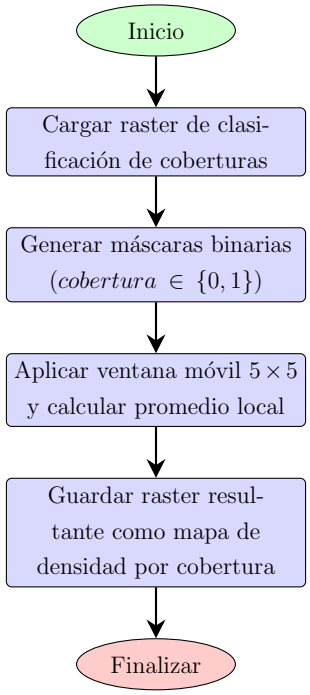
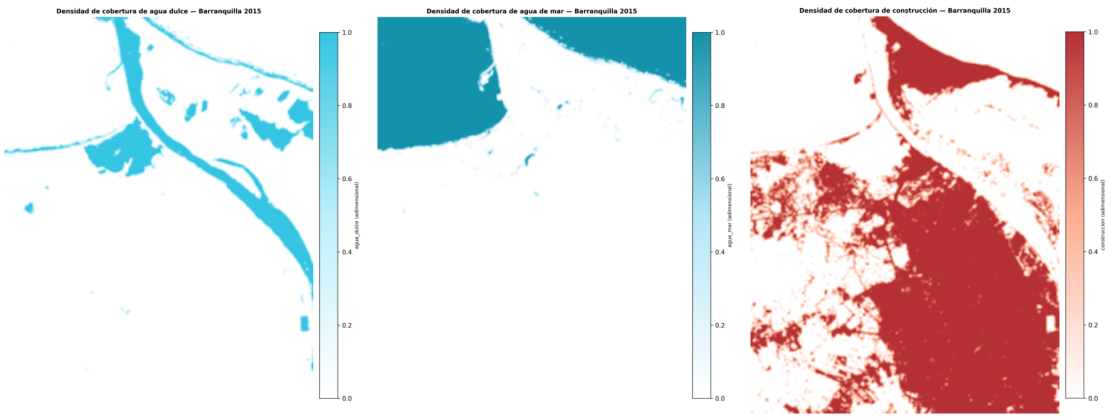


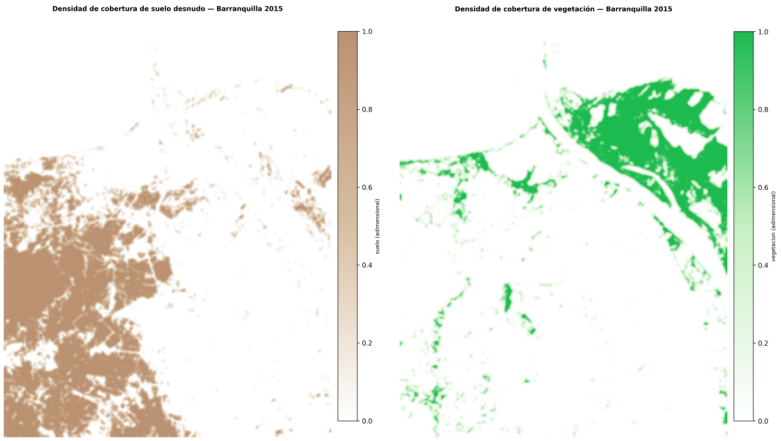
Figura 36: Diagrama del proceso para la generación de mapas de densidad de coberturas (ventana móvil 5×5).

El script desarrollado para la creación de los mapas de densidad se encuentra disponible públicamente en el repositorio de GitHub del proyecto `https://github.com/jonnykewin/tesis_MCD/blob/main/9_cover_lse.py`

A continuación, se presentan como ejemplo los mapas de densidad generados para cada tipo de cobertura agua dulce, agua de mar, construcción, suelo desnudo y vegetación en la ciudad de Barranquilla, correspondientes al año 2015 Figura 37. Cada mapa emplea una paleta de colores específica que facilita la interpretación visual de la distribución espacial de las coberturas. Los valores se expresan en una escala adimensional de 0 a 1, donde los tonos más intensos indican una mayor densidad o presencia relativa de cada cobertura en el territorio.



(a) Agua dulce. (b) Agua de mar. (c) Construcción



(d) Suelo. (e) Vegetación.

Figura 37: Mapas de densidad de coberturas Barranquilla 2015.

5.6. Integración de estaciones meteorológicas.

5.6.1. Exportación de estaciones a GeoPackage por ciudad y año.

Con el objetivo de organizar la información meteorológica en un formato geoespacial ligero y portable, se implementó un script en Python que se conectaba a la tabla *estaciones_consolidadas* en PostgreSQL/PostGIS. Este script automatizaba la creación de archivos GeoPackage (.gpkg) estructurados por combinaciones de ciudad y año.

El procedimiento incluyó cuatro etapas: (i) identificación de combinaciones válidas, (ii) extracción de subconjuntos desde la base de datos, (iii) verificación del sistema de coordenadas (EPSG:4326) y (iv) exportación a carpetas organizadas por ciudad y año. Cada archivo resultante contenía una capa denominada *estaciones_meteorologicas*, con los campos: fuente, ciudad, fecha_toma, codigo_estacion, medicion y geometría. Como resultado, se obtuvo una colección de archivos indepen-

dientes, optimizados para su uso en procesos de validación, análisis multitemporal y visualización cartográfica.

5.6.2. Enriquecimiento de estaciones con variables satelitales.

Los GeoPackage generados fueron enriquecidos con variables derivadas de productos ráster obtenidos en etapas anteriores del proyecto. Para ello, se diseñó un script en Python, ejecutado en QGIS, que empleaba el algoritmo *rastersampling* para muestrear automáticamente los valores de múltiples rásteres en la ubicación de cada estación.

Entre las capas utilizadas se incluyeron: temperatura de la superficie terrestre (LST), emisividad, índices espectrales (NDVI, NDBI, NDWI) y coberturas temáticas derivadas de clasificación supervisada (vegetación, construcción, suelo desnudo, agua dulce, agua salada y nubes). El proceso fue recursivo: tras cada iteración, el archivo se actualizaba con nuevas columnas según el producto procesado (por ejemplo, *lst_*, *ndvi_*). El script también gestionó la limpieza automática de archivos temporales, lo que garantizó eficiencia y reproducibilidad.

Cada estación fue asociada a las variables satelitales correspondientes a su ubicación geográfica, consolidando así un conjunto de datos robusto que integró mediciones en territorio con productos derivados de percepción remota. Esta integración permitió calcular con mayor precisión valores de temperatura superficial del suelo (LST), emisividad, índices espectrales y coberturas del terreno.

5.6.3. Integración y depuración final del dataset.

Una vez generados y enriquecidos los archivos GeoPackage, se implementó un procedimiento de importación recursiva a la base de datos `PostgreSQL/PostGIS`. Para ello, un script en Python buscó automáticamente los archivos *estaciones_meteorologicas_final.gpkg* dentro de la estructura de directorios por ciudad y año, extrajo sus capas y las cargó en tablas individuales dentro del esquema *estaciones_nuevas*. Durante la importación se añadieron metadatos de contexto (ciudad y año) y se definieron índices espaciales y temáticos para optimizar el acceso y las consultas posteriores.

Posteriormente, todas las tablas fueron consolidadas en una única tabla denominada *estaciones_datos*, que unificó los campos base (fuente, ciudad, fecha_toma, código de estación y medición) con los valores derivados de productos satelitales: temperatura superficial (LST), índices espectrales (NDVI, NDBI, NDWI), emisividad y cobertu-

ras temáticas (vegetación, construcción, suelo desnudo, agua dulce, agua salada y nubes).

En la depuración se eliminaron los registros nulos, originados en estaciones cuya localización quedaba fuera del área de cobertura de los rásteres y, por tanto, producían celdas sin información al muestrear. Esta exclusión garantizó que la base final incluyera únicamente observaciones válidas y espacialmente consistentes.

Finalmente, se aplicó un filtro temporal para seleccionar, por cada estación y día, el registro más cercano a las 10:00 hora local (UTC−5), coherente con la hora de adquisición de las escenas *Landsat*.Inicialmente se disponía de 31 visualizaciones comparables; tras integrar las series complementarias de ERA5/ERA5-Land y completar la cobertura en zonas costeras y marinas, el conjunto consolidado alcanzó 260 registros, cada uno representando un escenario de contraste entre mediciones en territorio/reanalizadas y productos de percepción remota.

5.7. Modelo supervisado para la determinación de temperatura.

Con el fin de estimar la temperatura del aire en el área de estudio, se desarrolló un modelo supervisado que integra técnicas de *machine learning* y *deep learning*.

Si bien el sensor térmico de Landsat proporciona una estimación de la temperatura de superficie, estos valores no necesariamente coinciden con la temperatura del aire medida por estaciones meteorológicas. La temperatura de superficie terrestre puede ser igual o incluso superior a la temperatura del aire, dependiendo de la cobertura del suelo [59]. El sensor Landsat captura la radiación emitida por la superficie, incluyendo tejados, pavimento, vegetación, entre otros. Estas coberturas pueden retener y emitir calor de manera distinta al aire, particularmente en zonas urbanas densas con materiales altamente absorbentes. En consecuencia, se construyó un modelo supervisado que estima la temperatura del aire a partir de observaciones de estaciones meteorológicas y las características derivadas y construidas en el estudio.

Variables y preparación de datos. Las variables explicativas incluyeron la *temperatura de la superficie terrestre (LST)*, la *emisividad superficial (ε)*, los índices espectrales (*NDVI*, *NDBI* y *NDWI*), los mapas de densidad de cobertura *(Agua de mar, Agua dulce, Construcciones, Nubes, Suelo desnudo, Vegetación)*, la *fecha de captura* y las coordenadas geográficas de las estaciones meteorológicas (*norte, este*). Para cada fecha, se realizó la intersección espacial entre las estaciones y los

raster correspondientes, obteniendo así el vector de predictores $\mathbf{x}_n$ y la temperatura observada y_n en cada ubicación.

Entrenamiento y validación. El conjunto total de datos se dividió en *bloque de entrenamiento (80 %)* y *conjunto de prueba (hold-out) (20 %)*, considerando la cantidad limitada de observaciones. Sobre el bloque de entrenamiento se aplicó una validación cruzada estratificada con $k = 10$ iteraciones. La búsqueda de hiperparámetros se llevó a cabo mediante búsqueda exhaustiva en rejilla (*Grid Search*), utilizando como métrica de selección el menor *RMSE*.

Modelos evaluados. Se entrenaron y compararon los siguientes algoritmos: *Linear Regression*, *Random Forest*, *Extra Trees*, *Support Vector Machines (SVC)*, *Gradient Boosting* y una *Multilayer Perceptron (MLP)*. Los hiperparámetros óptimos obtenidos para cada modelo se muestran en la Tabla 10.

Tabla 10: Mejores hiperparámetros (según validación cruzada).

Modelo	**Hiperparámetros (CV)**
Regresión Lineal	...
Random Forest	`max_depth=None, min_samples_leaf=1, min_samples_split=2, n_estimators=200`
Extra Trees	`max_depth=15, min_samples_leaf=1, min_samples_split=2, n_estimators=500`
Gradient Boosting	`learning_rate=0.1, max_depth=3, n_estimators=200`
SVC	`C=10.0, model_epsilon=0.1, kernel=rbf`
MLP	`activation=tanh, alpha=0.01, hidden_layer_sizes=(128,64,32), solver=lbfgs`

Evaluación de desempeño. Una vez elegidos los mejores hiperparámetros, cada modelo se reentrenó con el bloque de entrenamiento completo y se evaluó sobre el conjunto de prueba *(hold-out)*. Se reportaron las métricas *RMSE*, *MAE* y R^2 tanto para el entrenamiento con validación cruzada *(OOF-CV)* como para la evaluación en prueba *(TEST)*, con el fin de contrastar la estabilidad del ajuste frente a su rendimiento fuera de muestra (Tabla 11).

Tabla 11: Desempeño y tiempos por modelo en entrenamiento con validación cruzada (OOF-CV) y en prueba (TEST).

Modelo	TRAINING (OOF-CV)				TEST (hold-out)			
	R^2	RMSE[°C]	MAE[°C]	Time[s]	R^2	RMSE[°C]	MAE[°C]	Time[s]
Regresión Lineal	0.555	1.658	0.972	0.030	-0.012	2.662	1.237	0.033
MLP Regressor	0.865	0.912	0.613	24.273	0.728	1.379	0.806	174.407
SVR	0.834	1.012	0.707	0.027	0.756	1.306	0.890	0.077
Random Forest	0.926	0.675	0.454	0.744	0.910	0.793	0.536	6.575
Extra Trees	**0.942**	**0.600**	**0.420**	**1.222**	**0.921**	**0.745**	**0.522**	**16.564**
Gradient Boosting	0.938	0.617	0.439	0.268	0.935	0.676	0.488	4.427

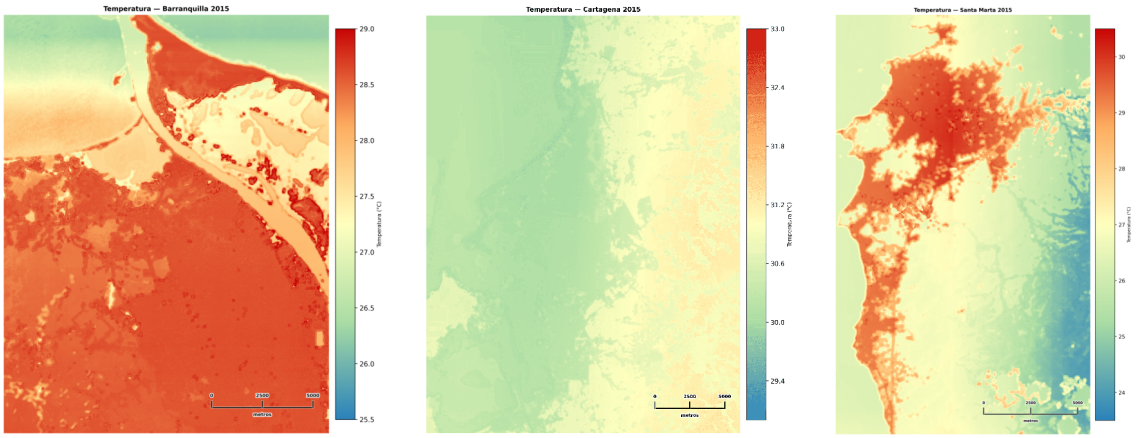
Interpretación de resultados. Los datos se organizaron de manera lógica y comparable, lo que permitió evaluar el desempeño de distintos algoritmos de regresión. El *Gradient Boosting* se destacó como el modelo con mejor rendimiento en las métricas estadísticas *(R^2, RMSE y MAE)*. Esto coincide con estudios recientes que destacan el éxito de los algoritmos de boosting en la predicción de la temperatura superficial (LST) y la identificación de islas de calor urbano. Investigaciones como las revisiones sobre el efecto de isla de calor urbano y la predicción con inteligencia artificial refuerzan la idea de que el aprendizaje automático, especialmente los métodos de boosting, ofrecen alta precisión y escalabilidad en el análisis de datos de teledetección aplicados a las islas de calor [60].

Sin embargo, aunque el *Gradient Boosting* lideró en métricas, su aplicación en la generación de mapas produjo efectos visuales en forma de grilla o cuadriculado. Esta situación puede explicarse porque, como se señala en la documentación oficial de scikit-learn, las predicciones de los árboles de decisión no son continuas ni suaves, sino que corresponden a aproximaciones constantes por tramos, lo que en aplicaciones espaciales puede traducirse en superficies cuadriculadas.

Para resolver este problema, se eligió el modelo ***Extra Trees*** como modelo final para la predicción de temperatura. Aunque no superó al *Gradient Boosting* en todas las métricas, *Extra Trees* produjo mapas de superficie más suaves y sin patrones de grilla o cuadriculado visibles. Además, mantuvo un desempeño estadístico muy competitivo. En resumen, la selección del *Extra Trees* como modelo final se basa en un equilibrio entre precisión numérica y coherencia espacial. Si bien *Gradient Boosting* es conocido por su capacidad predictiva, en este caso se priorizó la estabilidad visual de los mapas generados, lo que resulta en un producto final más adecuado para el análisis de las islas de calor urbano en el área de estudio.

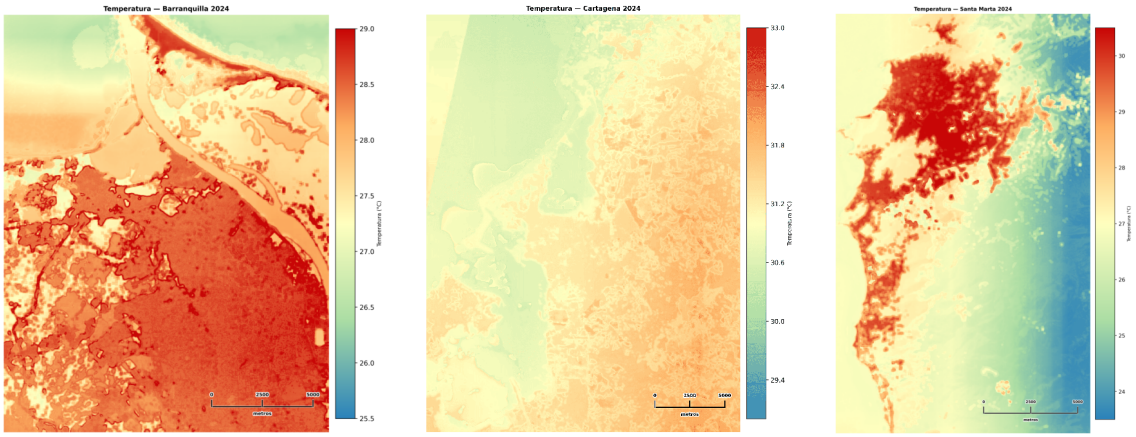
A continuación, se presentan comparaciones visuales de la temperatura estimada

correspondiente a los años 2015 (Figura 38) y 2024 (Figura 39) en las ciudades de Barranquilla, Cartagena y Santa Marta, generadas mediante el algoritmo *Extra Trees*.



(a) Barranquilla (b) Cartagena (c) Santa Marta

Figura 38: Estimación de temperatura en 2015 para Barranquilla, Cartagena y Santa Marta.



(a) Barranquilla (b) Cartagena (c) Santa Marta

Figura 39: Estimación de temperatura en 2024 para Barranquilla, Cartagena y Santa Marta.

5.8. Evaluación del modelo de temperatura frente a estaciones meteorológicas.

Para garantizar una evaluación confiable y trazable, la validación del modelo de estimación de temperatura se realizó exclusivamente con las estaciones terrestres pertenecientes al IDEAM y DIMAR, descartando las estaciones virtuales. Esta decisión se fundamenta en que las estaciones en tierra proporcionan mediciones directas

y "calibradas" por sensores físicos, mientras que las estaciones virtuales constituyen estimaciones derivadas de interpolaciones. En consecuencia, el análisis se basó únicamente en los registros con observaciones en sitio, de manera que el error calculado refleje de forma precisa el desempeño del modelo frente a mediciones reales.

Las estaciones empleadas para la evaluación del modelo de temperatura se presentan en la Tabla 12. Los registros se encuentran organizados por ciudad y fecha de toma.

Tabla 12: Estaciones utilizadas para la evaluación del modelo de temperatura (IDEAM y DIMAR).

Ciudad	**Fecha de toma**	**Fuente**	**Medición estación [°C]**	**Temperatura modelo [°C]**
Barranquilla	2015-04-01 10:00	IDEAM	28.5	28.50
Barranquilla	2017-01-16 10:00	IDEAM	27.9	27.89
Barranquilla	2018-02-04 10:00	IDEAM	30.0	29.99
Barranquilla	2019-01-06 10:00	IDEAM	29.6	29.59
Barranquilla	2020-03-29 07:00	DIMAR	26.4	26.40
Barranquilla	2020-03-29 10:00	IDEAM	30.2	29.08
Barranquilla	2020-03-29 13:00	DIMAR	28.5	28.50
Barranquilla	2021-02-12 10:00	IDEAM	30.9	30.89
Barranquilla	2021-02-12 13:00	DIMAR	28.12	28.12
Barranquilla	2021-02-12 13:00	DIMAR	28.6	28.60
Barranquilla	2022-02-23 13:00	DIMAR	27.5	27.50
Barranquilla	2022-02-23 13:00	DIMAR	27.81	27.81
Barranquilla	2023-01-17 10:00	IDEAM	30.4	30.39
Barranquilla	2023-01-17 13:00	DIMAR	28.0	28.17
Cartagena	2015-04-01 10:00	IDEAM	31.8	30.01
Cartagena	2016-01-14 10:00	IDEAM	30.8	30.80
Cartagena	2017-01-16 10:00	IDEAM	29.7	29.70
Cartagena	2018-02-04 10:00	IDEAM	29.6	29.60
Cartagena	2019-01-06 10:00	IDEAM	30.7	30.83
Santa Marta	2015-04-01 10:00	IDEAM	32.6	30.25
Santa Marta	2015-04-01 13:00	DIMAR	27.0	27.00
Santa Marta	2016-01-14 13:00	DIMAR	26.1	26.10
Santa Marta	2017-01-16 10:00	IDEAM	29.2	29.20
Santa Marta	2017-01-16 13:00	DIMAR	27.5	26.22
Santa Marta	2018-02-04 10:00	IDEAM	30.9	29.72
Santa Marta	2019-01-06 10:00	IDEAM	31.2	31.80
Santa Marta	2019-01-06 13:00	DIMAR	24.7	27.17
Santa Marta	2020-03-29 10:00	IDEAM	31.8	31.80
Santa Marta	2021-02-12 13:00	DIMAR	26.7	26.70
Santa Marta	2022-02-23 13:00	DIMAR	27.3	27.30

Notas: Se incluyen únicamente las estaciones terrestres pertenecientes a IDEAM y DIMAR empleadas en la validación del modelo de temperatura. Los registros se presentan en orden ascendente según la fecha de toma.

Metodología de evaluación Para cada estación y fecha, se calculó la diferencia entre la temperatura estimada y la observada ($e_i = \hat{y}_i - y_i$). A partir de estos valores, se determinó el *error cuadrático medio* (RMSE, por sus siglas en inglés), que representa la magnitud promedio de las desviaciones del modelo respecto a las mediciones observadas:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{12}$$

Este indicador permite evaluar la precisión global del modelo: valores más bajos de RMSE indican una mayor concordancia entre las temperaturas estimadas y las registradas en campo.

Resultados La Tabla 13 muestra los resultados de RMSE obtenidos tras el filtrado de estaciones terrestres. En general, el modelo presenta un error promedio reducido y homogéneo entre ciudades, lo que respalda su capacidad de generalización espacial en la estimación de la temperatura a partir de variables satelitales.

Tabla 13: Error cuadrático medio (RMSE) del modelo de temperatura por ciudad (solo estaciones IDEAM/DIMAR).

Ciudad	**n**	**RMSE [°C]**
Barranquilla	14	0,303
Cartagena	5	0,801
Santa Marta	11	1,168
Global	*30*	**0,806**

Notas: n = número de observaciones; RMSE = raíz del error cuadrático medio.

6. Islas de calor y análisis multitemporal.

Este capítulo presenta los resultados finales del estudio, enfocándose en la identificación, distribución y evolución de las islas de calor urbanas en las ciudades Barranquilla, Cartagena y Santa Marta. Mediante la aplicación de modelos desarrollados, se evaluaron los patrones térmicos y su correlación con las coberturas del suelo, efectuando un análisis multitemporal de las variaciones entre los años 2015 y 2024. Se incorporan mapas comparativos de temperatura y coberturas, acompañados de la interpretación espacial de las zonas con mayor concentración térmica y su relación con los procesos de urbanización y pérdida de vegetación.

6.1. Identificación de Islas de Calor Urbanas (ICU).

6.1.1. Insumos y preprocesamiento.

Para cada escena (ciudad–fecha) se empleó como insumo principal un campo térmico de temperatura de superficie $T(x)$ generado por el modelo, en línea con la literatura que ha analizado ICU a partir de campos térmicos derivados [61, 62]. Como soporte para la construcción de *máscaras* basadas en coberturas se utilizaron capas de densidad normalizadas en $[0, 1]$. A partir de estos insumos se establecieron umbrales fijos para tres máscaras binarias; cada máscara devolvía 1 cuando se cumplía la condición y 0 en caso contrario:

$$U(x) = \begin{cases} 1, & \text{si construcciones}(x) \geq 0.6, \\ 0, & \text{en otro caso,} \end{cases} \quad \text{(urbano)}$$

$$V(x) = \begin{cases} 1, & \text{si vegetación}(x) \geq 0.6, \\ 0, & \text{en otro caso,} \end{cases} \quad \text{(vegetación densa)}$$

$$E(x) = \begin{cases} 1, & \text{si agua salada}(x) > 0.5 \ \vee \ \text{agua dulce}(x) > 0.5 \ \vee \ \text{nubes}(x) > 0.5, \\ 0, & \text{en otro caso,} \end{cases} \quad \text{(exclusión).} \tag{13}$$

Los valores sin información en cualquiera de las capas se asignaron a la máscara de exclusión, es decir, $E(x) = 1$, para evitar falsos positivos en etapas posteriores.

6.1.2. Línea base térmica y anomalía ΔT.

La línea base térmica local se estima sobre vegetación densa libre de exclusión:

$$\mathcal{B} = \{\, T(x) \;:\; V(x) = 1,\; E(x) = 0 \,\}. \tag{14}$$

Se adopta como referencia el percentil 25 de $\mathcal{B}$:

$$T_{\text{base}} = \mathrm{P}_{25}(\mathcal{B}), \tag{15}$$

y se define la anomalía por píxel como:

$$\Delta T(x) = T(x) - T_{\text{base}}. \tag{16}$$

El uso de vegetación como referencia fría y local está respaldado por estudios que recomiendan bases naturales para comparar el contraste urbano [3]. Los valores sin información se excluyen de todos los cálculos estadísticos.

6.1.3. Umbralización adaptativa e intensidad de ICU.

La detección se restringe al dominio urbano válido

$$M = \{\, x \;:\; U(x) = 1,\; E(x) = 0,\; \Delta T(x) \in \mathbb{R} \,\}. \tag{17}$$

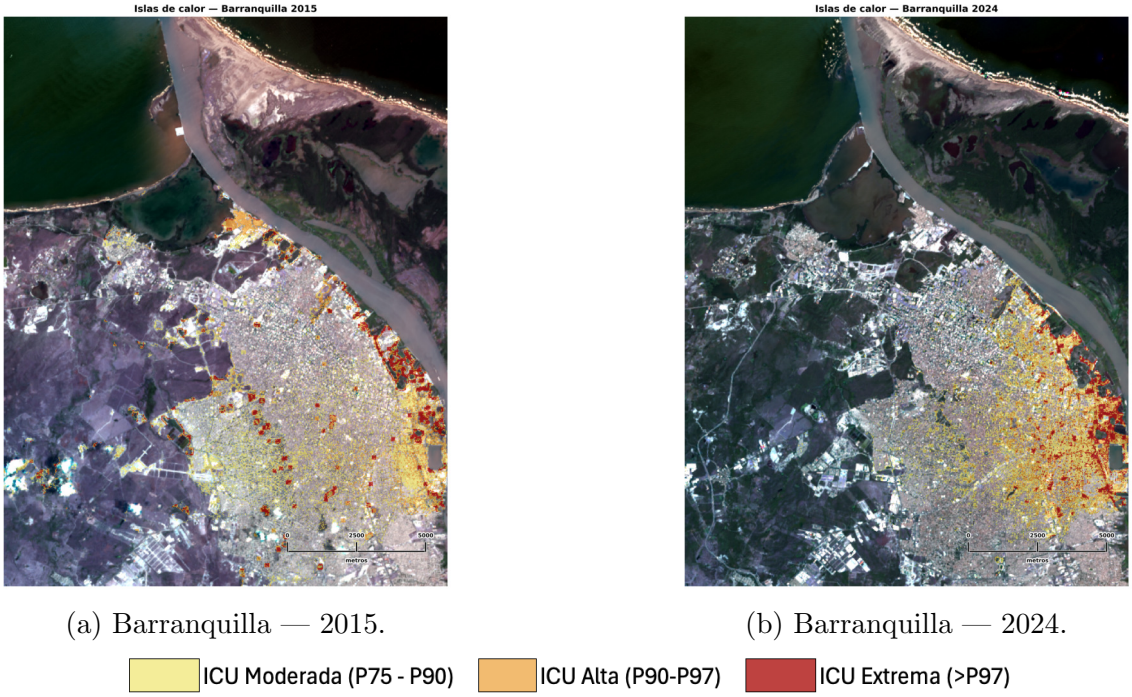
Sobre los valores $\{\Delta T(x) : x \in M\}$ se calculan percentiles adaptativos del propio día (P_{75}, P_{90}, P_{97}), y la intensidad se clasifica como:

$$\text{ICU\_intensity}(x) = \begin{cases} 0, & x \notin M \text{ o } \Delta T(x) < P_{75}, \\ 1, & P_{75} \le \Delta T(x) < P_{90}, \\ 2, & P_{90} \le \Delta T(x) < P_{97}, \\ 3, & \Delta T(x) \ge P_{97}. \end{cases} \tag{18}$$

Este esquema percentilar intraurbano permitía detectar "hotspots" térmicos relativos al contexto del día [63]. En este marco, $T(x)$ correspondía a los campos térmicos de temperatura de superficie generados por el modelo. Las ICU se representaron mediante mapas de zonificación para las ciudades de Barranquilla (Figura 40), Cartagena (Figura 41) y Santa Marta (Figura 42) en los años extremos del periodo (2015 y 2024).

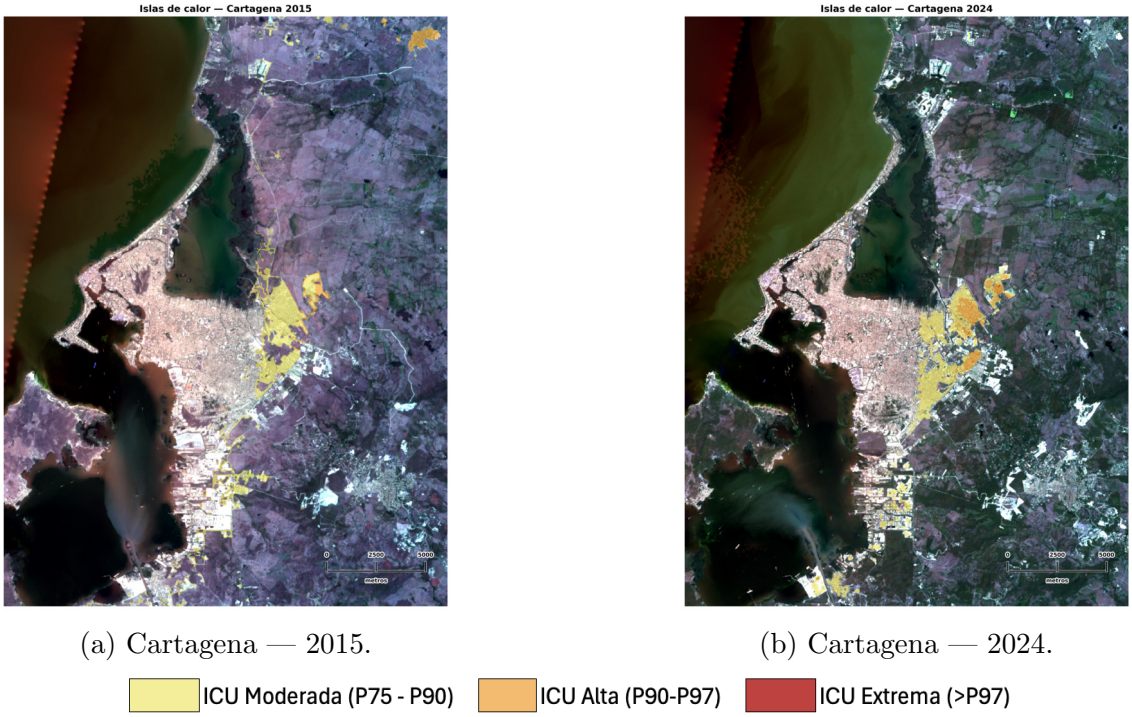
Las clases de intensidad consideradas fueron *moderada*, *alta* y *extrema*, codificadas

cromáticamente en amarillo, naranja y rojo, correspondientes a los intervalos *ICU moderada* (P75–P90), *ICU alta* (P90–P97) e *ICU extrema* (≥P97), respectivamente.



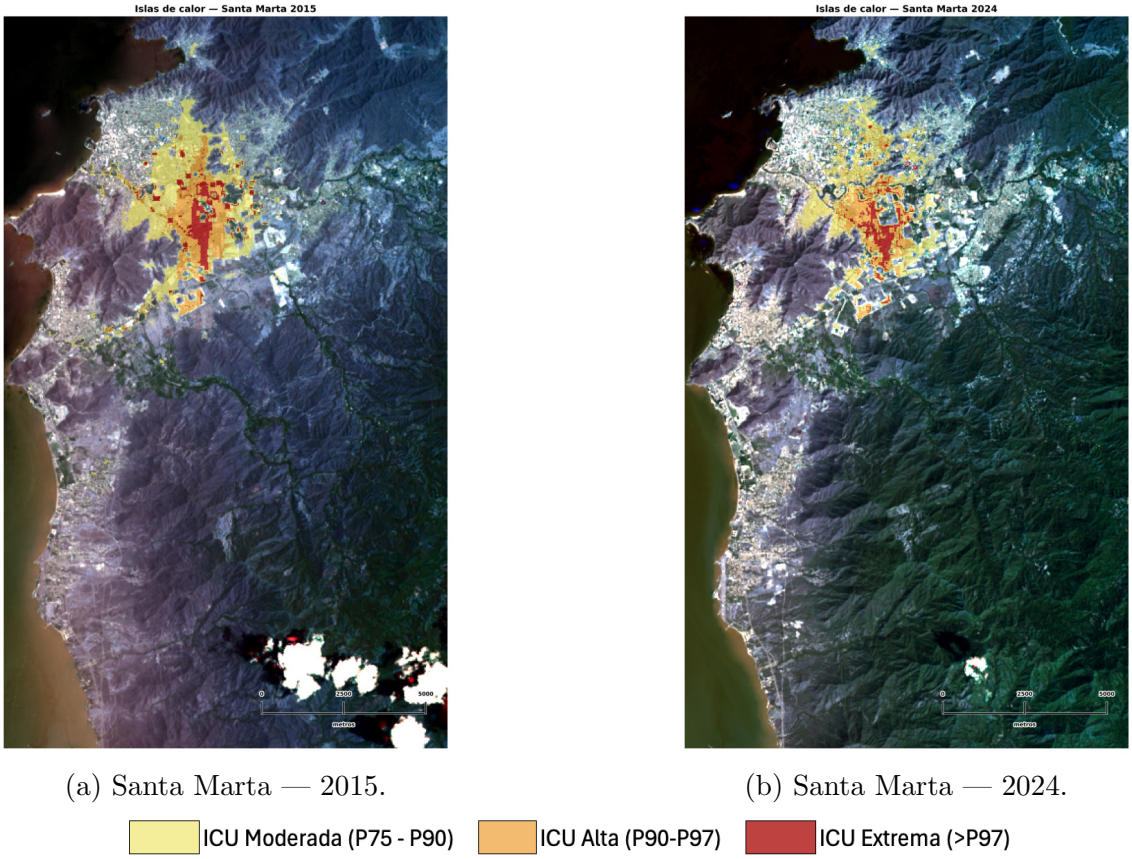
(a) Barranquilla — 2015. (b) Barranquilla — 2024.

Figura 40: Zonificación de islas de calor para la ciudad de Barranquilla.



(a) Cartagena — 2015. (b) Cartagena — 2024.

Figura 41: Zonificación de islas de calor para la ciudad de Cartagena.



(a) Santa Marta — 2015.

(b) Santa Marta — 2024.

Figura 42: Zonificación de islas de calor para la ciudad de Santa Marta.

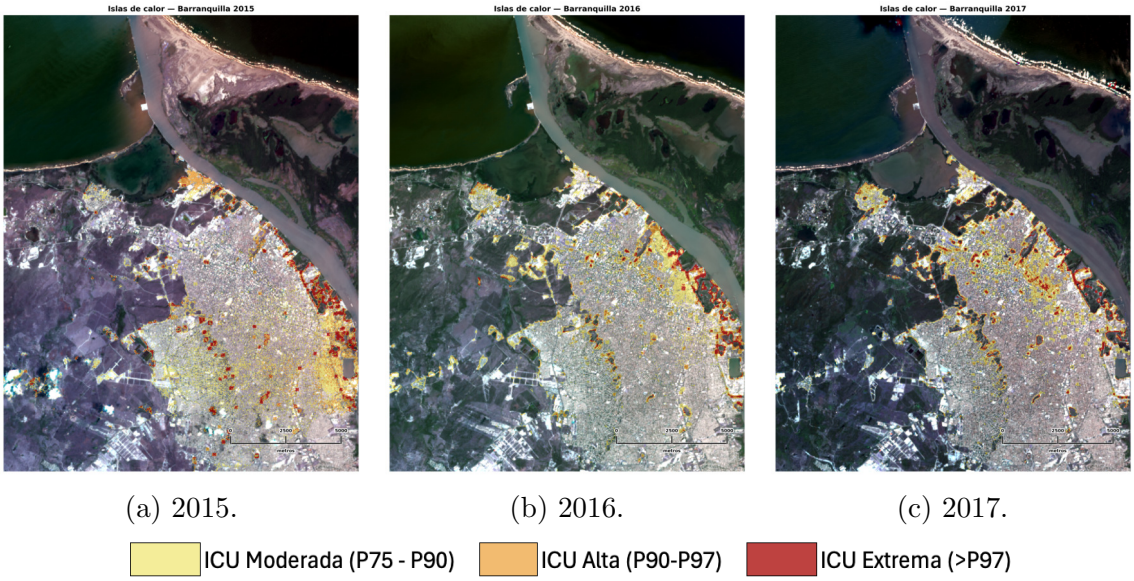
6.2. Análisis multitemporal.

6.2.1. Resultados cartográficos comparativos.

Para iniciar el análisis multitemporal de la evolución de las Islas de Calor Urbanas en las ciudades de la costa Caribe (2015–2024), se abordó primero una lectura estrictamente espacial. Los mosaicos cartográficos permitieron revisar la configuración y el cambio de los patrones intraurbanos, reconociendo sectores donde la concentración del fenómeno se intensificó, se desplazó o se fragmentó entre años representativos, como antesala a su cuantificación y evaluación estadística.

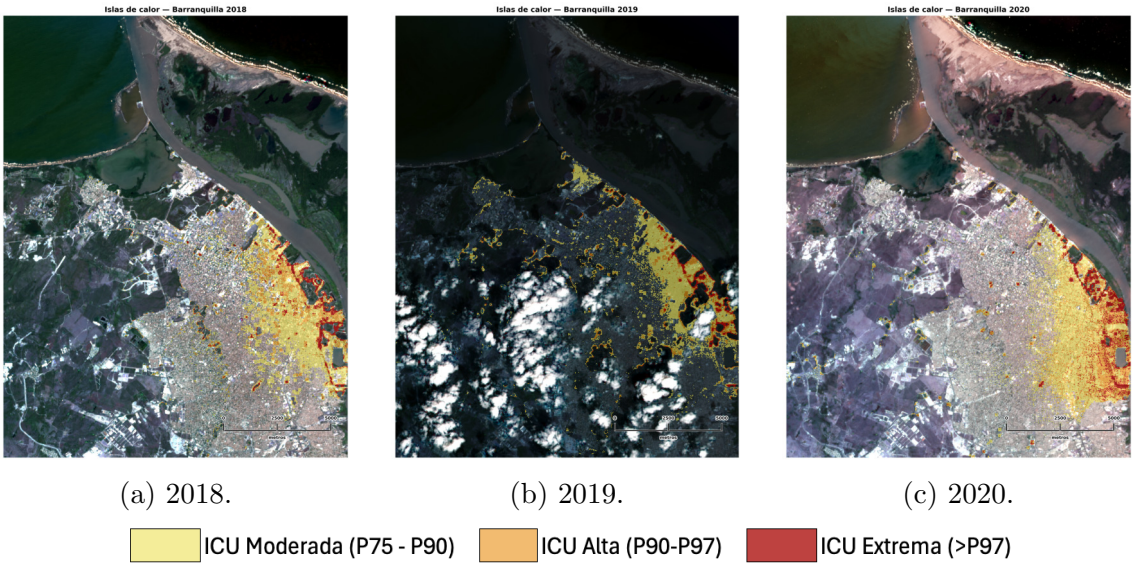
Barranquilla. En Barranquilla se observó un eje cálido dominante sobre el costado oriental del tejido urbano, con mayor expresión de las clases *alta* y *extrema* hacia el frente costero–fluvial y sectores densamente construidos. Entre 2015 y 2017 (Figura 43) predominó un patrón de manchas *moderadas*, con focos cálidos puntuales distribuidos dentro del tejido. Hacia el periodo 2018–2020 (Figura 44) se advirtió la consolidación del borde cálido oriental y la ampliación de parches *altos/extremos*,

en continuidad con corredores viales y áreas de mayor compactación. El año 2019 presentó nubosidad residual que redujo la legibilidad fina del patrón intraurbano, por lo que su interpretación se tomó con cautela. Finalmente, entre 2021 y 2024 (Figura 45), el frente oriental mantuvo la mayor carga térmica relativa, con expansión hacia el suroriente y la emergencia de núcleos secundarios dispersos, coherentes con procesos de densificación y continuidad del mosaico urbano. En síntesis, el análisis evidenció una concentración persistente de las ICU hacia el oriente de la ciudad, particularmente en la ronda del río Magdalena.



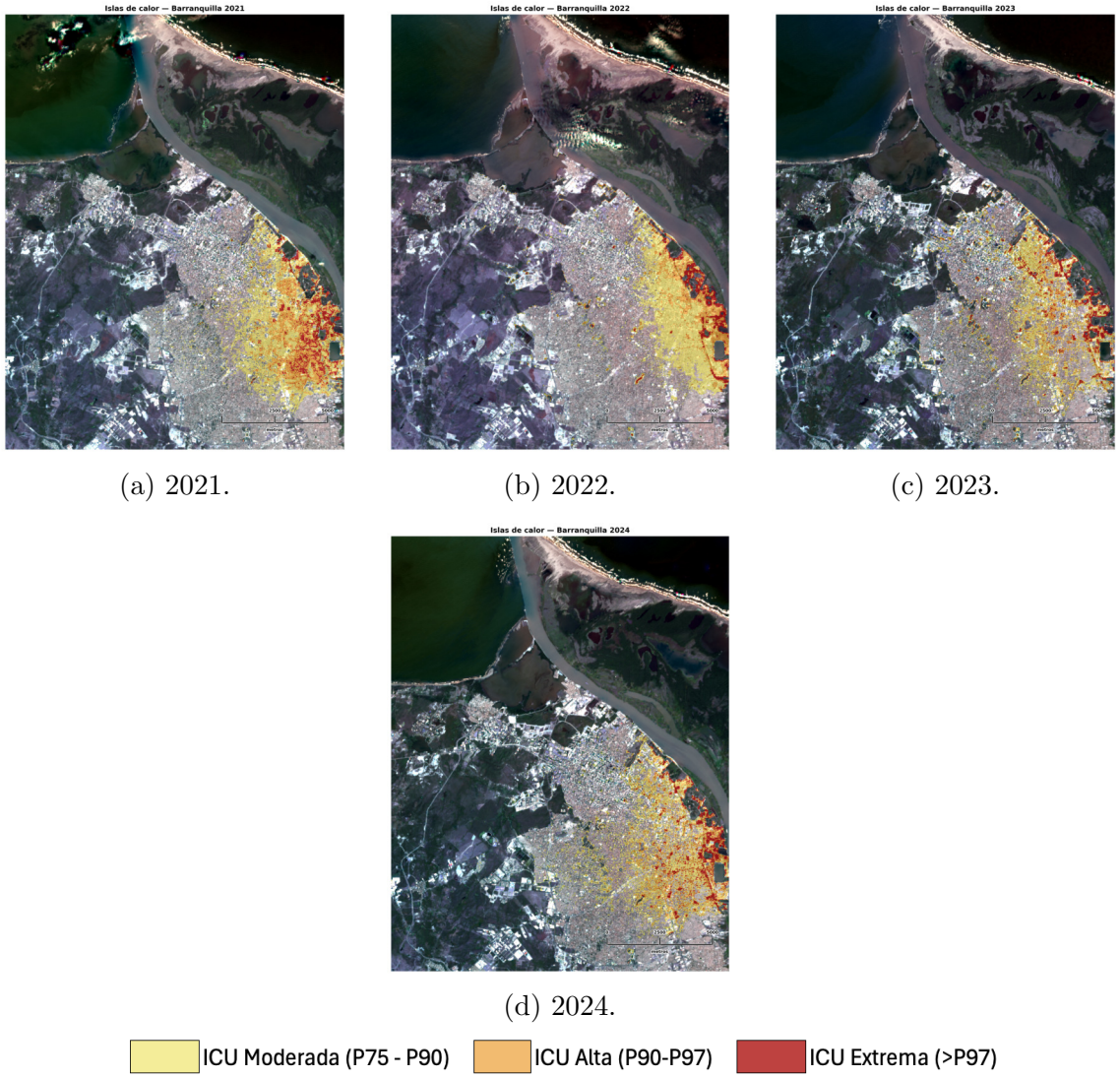
(a) 2015. (b) 2016. (c) 2017.

Figura 43: Evolución cartográfica de ICU — Barranquilla (2015–2017).



(a) 2018. (b) 2019. (c) 2020.

Figura 44: Evolución cartográfica de ICU — Barranquilla (2018–2020).



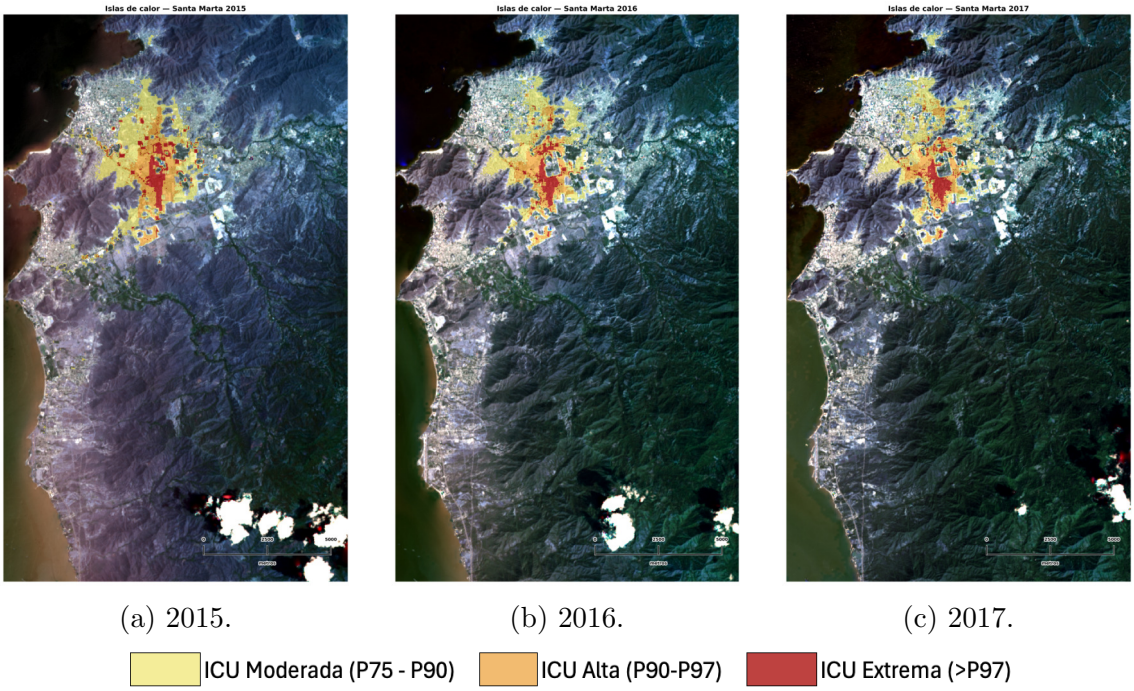
(a) 2021. (b) 2022. (c) 2023.

(d) 2024.

Figura 45: Evolución cartográfica de ICU — Barranquilla (2021–2024). Leyenda y cortes uniformes.

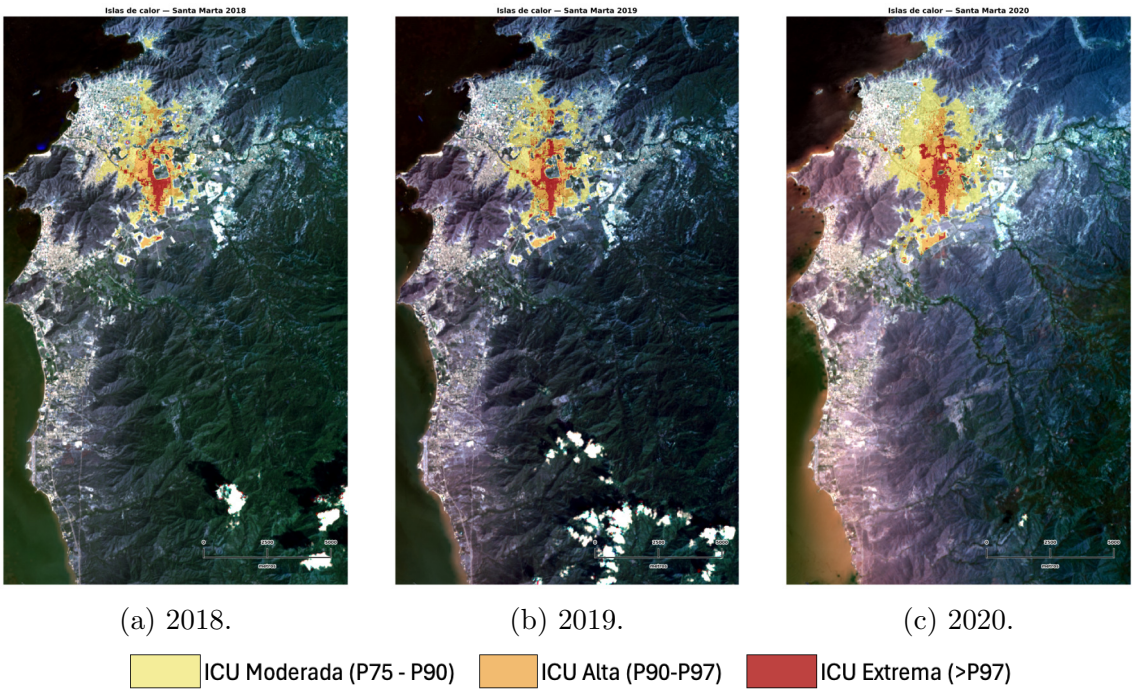
Santa Marta. En Santa Marta se identificó un patrón cálido estable con máxima expresión en el tejido urbano central y su entorno inmediato, mientras que el sector de Gaira mostró escasa o nula presencia de clases *alta/extrema* a lo largo del periodo. Entre 2015 y 2017 (Figura 46) predominó una configuración concentrada en el núcleo central, con extensiones *moderadas* hacia el nororiente, especialmente en dirección al piedemonte de la Sierra Nevada. En 2018–2020 (Figura 47) el núcleo cálido se densificó y ganó continuidad, manteniéndose anclado al centro urbano y a los principales ejes viales. Para 2021–2024 (Figura 48), el patrón persistió: las clases *alta/extrema* se estructuraron como un bloque compacto sobre el centro y el norte inmediato, con leves expansiones radiales y fragmentación secundaria; el año 2022 presentó nubosidad/oscurecimiento al sureste que limitó la lectura fina. En conjun-

to, los mapas evidenciaron una concentración reiterada del fenómeno en el centro de Santa Marta, sin un desplazamiento sostenido hacia Gaira.



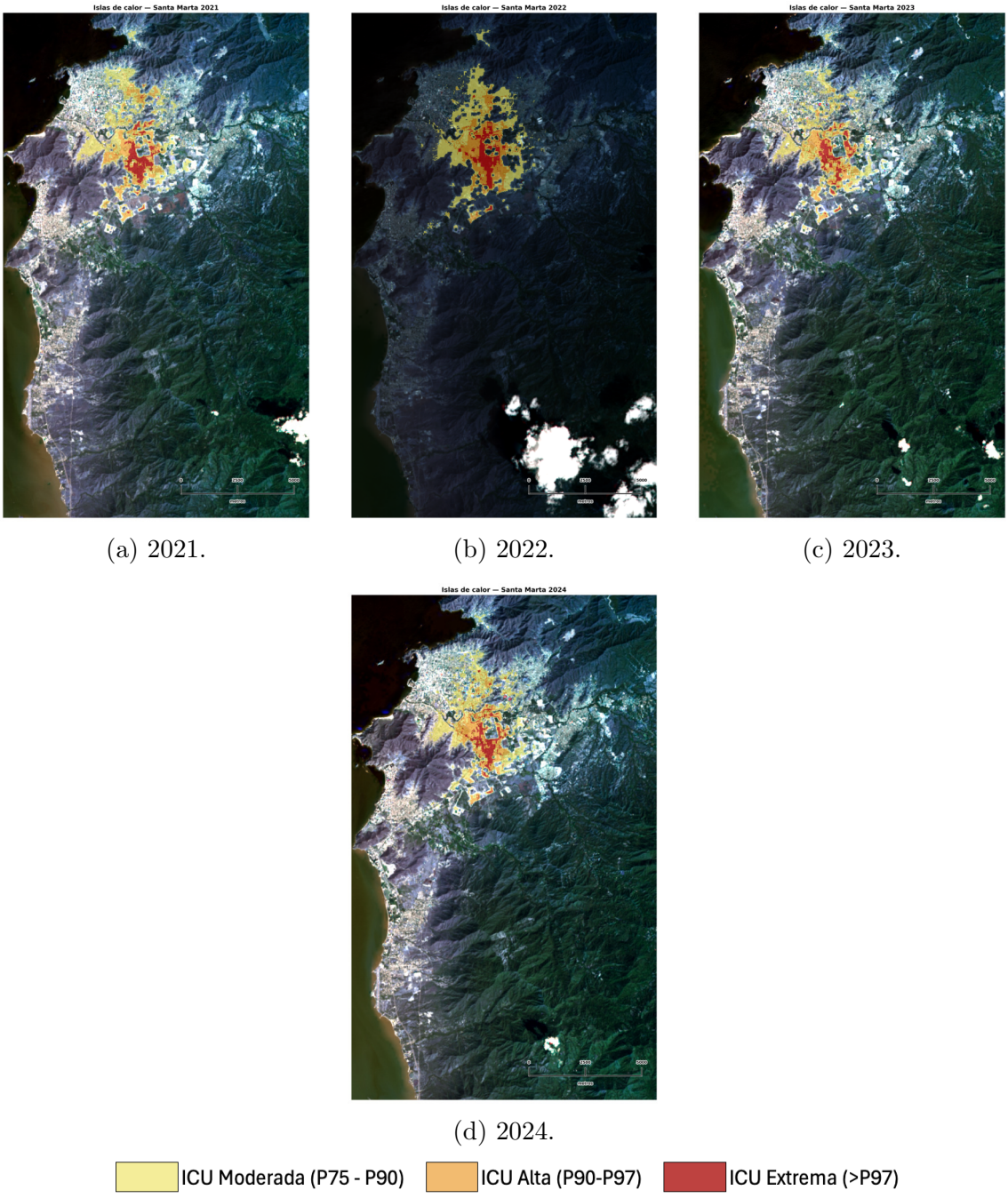
(a) 2015. (b) 2016. (c) 2017.

Figura 46: Evolución cartográfica de ICU — Santa Marta (2015–2017).



(a) 2018. (b) 2019. (c) 2020.

Figura 47: Evolución cartográfica de ICU — Santa Marta (2018–2020).



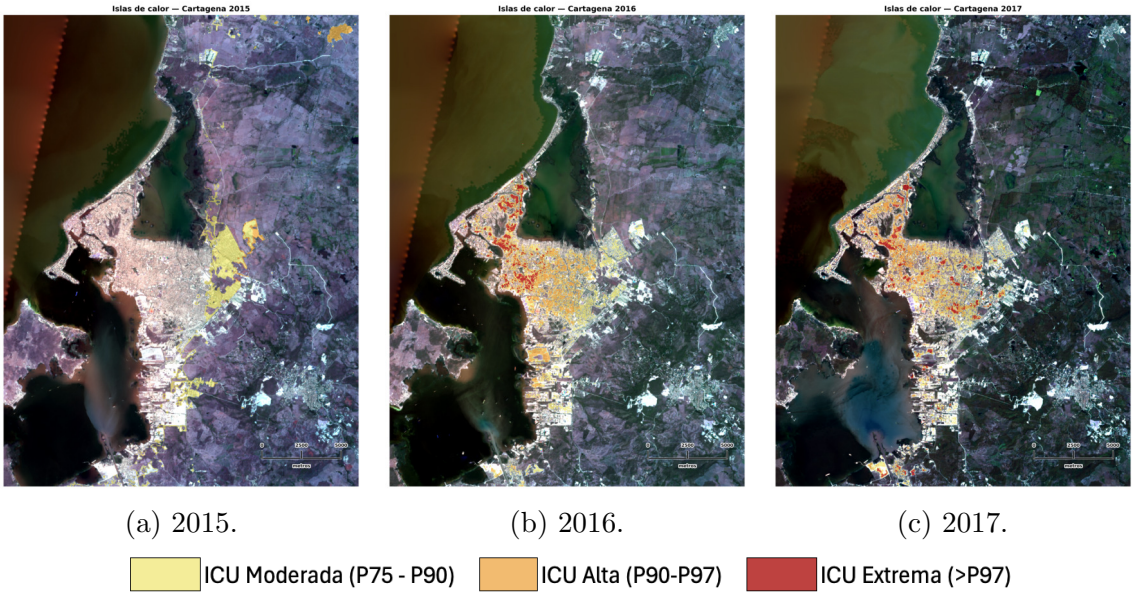
(a) 2021. (b) 2022. (c) 2023.

(d) 2024.

Figura 48: Evolución cartográfica de ICU — Santa Marta (2021–2024).

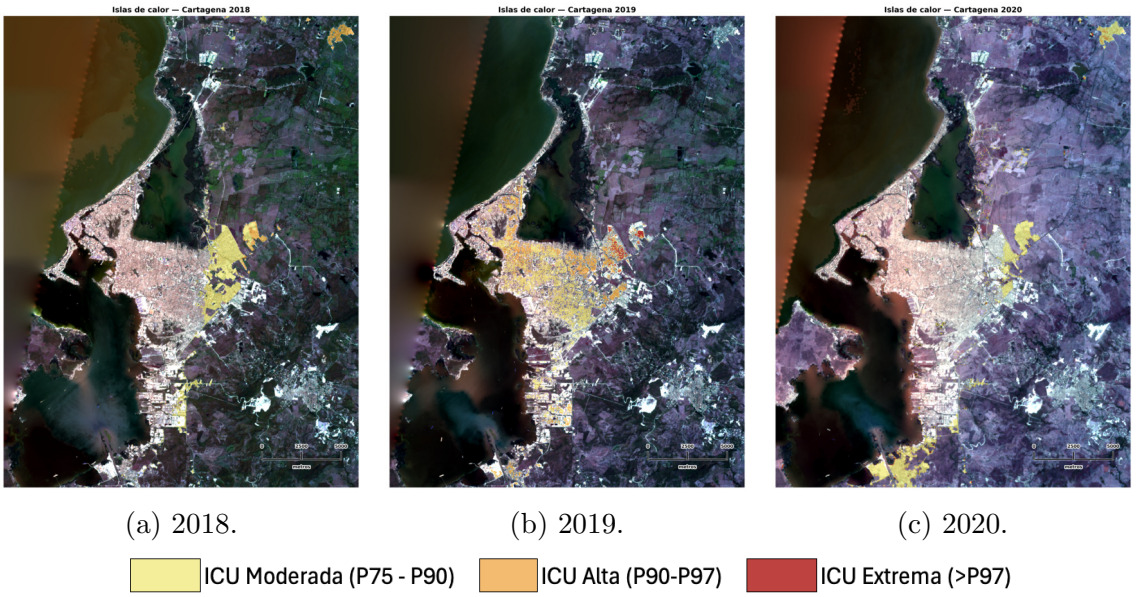
Cartagena. A diferencia de Barranquilla y Santa Marta, Cartagena no exhibió un patrón espacial plenamente persistente a lo largo del periodo. Aunque la mayor parte de las ICU se concentró reiteradamente en el *casco urbano principal*, la localización de las clases *alta/extrema* cambió entre años, sin un eje direccional estable. En la periferia nororiental se observaron episodios de nivel 2–3 hacia Bayunca (especialmente en 2015, 2018 y 2020), tal como se apreciaron en las Figuras 49 y 50,

mientras que en el corredor sur aparecieron de forma intermitente focos vinculados a actividades industriales y portuarias en el eje *Mamonal–Barú*. Dentro del continuo urbano, varios años mostraron una intensificación marcada en sectores del oriente y suroriente, con mayor densidad de ICU en *El Pozón*, *La Esperanza* y el entorno de la *Zona Franca Parque Central*.



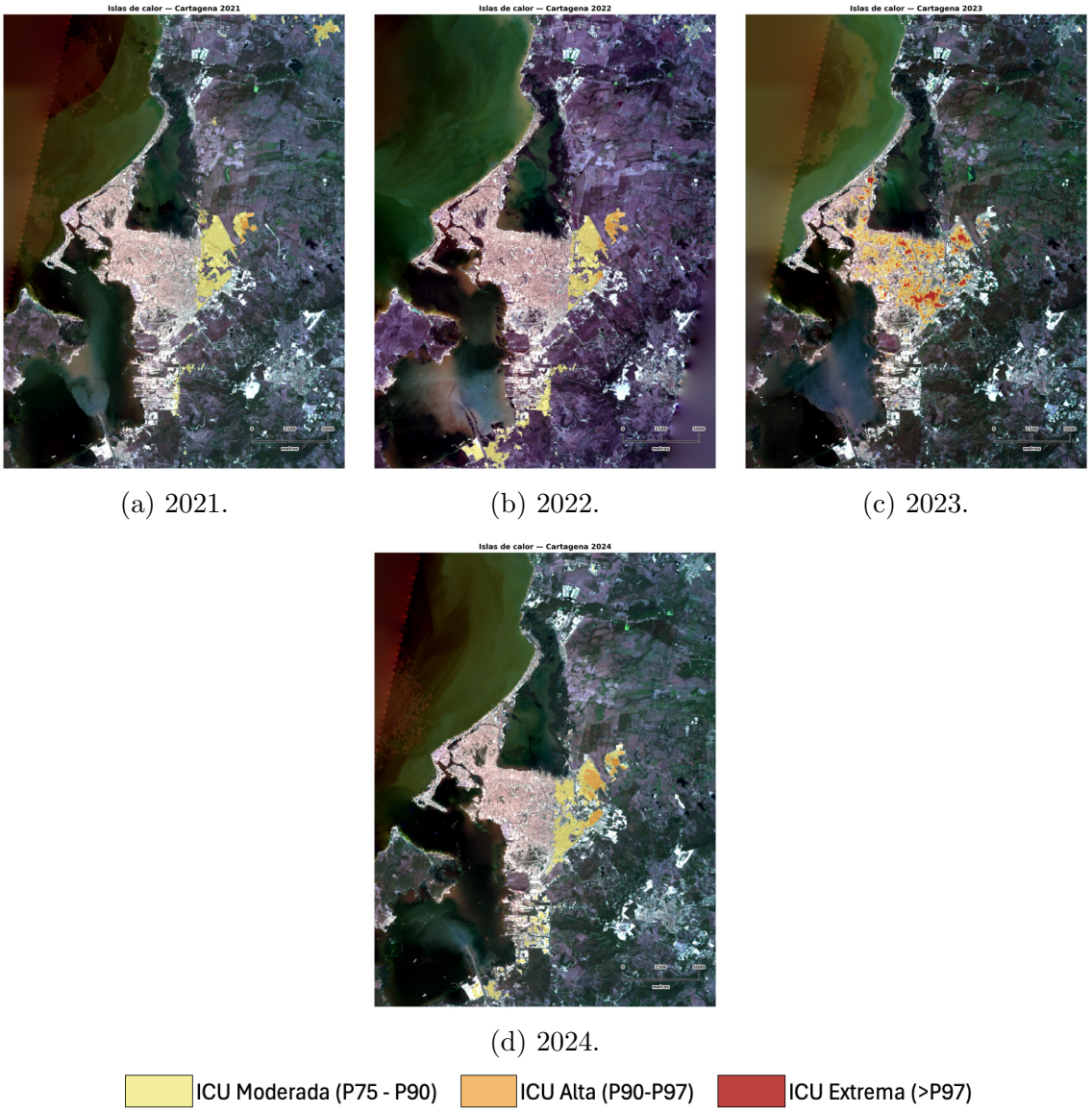
(a) 2015. (b) 2016. (c) 2017.

Figura 49: Evolución cartográfica de ICU — Cartagena (2015–2017).



(a) 2018. (b) 2019. (c) 2020.

Figura 50: Evolución cartográfica de ICU — Cartagena (2018–2020).



(a) 2021. (b) 2022. (c) 2023.

(d) 2024.

Figura 51: Evolución cartográfica de ICU — Cartagena (2021–2024).

6.2.2. Cuantificación anual y composición por clases.

Esta subsección presenta, para cada ciudad, la cuantificación anual (2015–2024) del área afectada por ICU desagregada en *moderada* (1), *alta* (2) y *extrema* (3). Los valores provienen de la zonificación cartográfica descrita previamente y se disponen, en cada caso, como (i) una tabla anual para dimensionar magnitudes y composición entre clases y (ii) su serie temporal, útil para reconocer variaciones interanuales, inflexiones y episodios atípicos. Los años con posibles limitaciones de observación (p. ej., nubosidad residual) se interpretan con cautela y no comprometen las conclusiones sobre la evolución general.

Barranquilla. En Barranquilla, la serie de tiempo del 2015 al 2024 (Figura 52) mostró un patrón consistente de predominio del área con *ICU moderada (1)*, que osciló alrededor de 1.36 mil ha y presentó la menor variabilidad (CV≈10.6 %): cayó en 2016 y 2017, presentó un repunte en 2018, alcanzó su máximo en 2020 (1598 ha) y luego descendió ligeramente, manteniéndose alta en 2021–2024. Las clases más intensas fueron más volátiles: *ICU alta (2)* registró su mínimo en 2019 (354 ha) y un salto marcado en 2020 (823 ha, máximo), sostuvo valores relativamente elevados en 2021 y se estabilizó en 2022–2024; *ICU extrema (3)* descendió hasta su mínimo en 2019 (128 ha) y también pico en 2020 (337 ha, máximo), con niveles posteriores que se mantuvieron por encima (o cercanos) a 2015. En conjunto, la evolución anual tabulada (Tabla 14) y la curva temporal evidenciaron un pico sincronizado en 2020 para las tres categorías y una recuperación parcial posterior. Además, la tabla de estadísticas descriptivas (Tabla 15) confirmó medias de 1361.9 ha (moderada), 572.9 ha (alta) y 254.8 ha (extrema), con mayor dispersión relativa en las clases alta (CV≈27.5 %) y extrema (CV≈25.4 %) que en la moderada, y con años extremos coherentes con los mínimos de 2019 y máximos de 2020 observados en la serie.

Tabla 14: Barranquilla: área (ha) de ICU por clase y año.

Año	ICU moderada (1)	ICU alta (2)	ICU extrema (3)
2015	1480.23	618.30	292.23
2016	1193.58	393.75	192.42
2017	1153.17	390.42	242.91
2018	1404.99	568.26	228.96
2019	1235.70	354.33	128.34
2020	1598.13	822.87	336.51
2021	1473.48	754.20	325.35
2022	1439.01	542.07	221.22
2023	1272.87	607.05	276.21
2024	1368.09	677.97	303.75

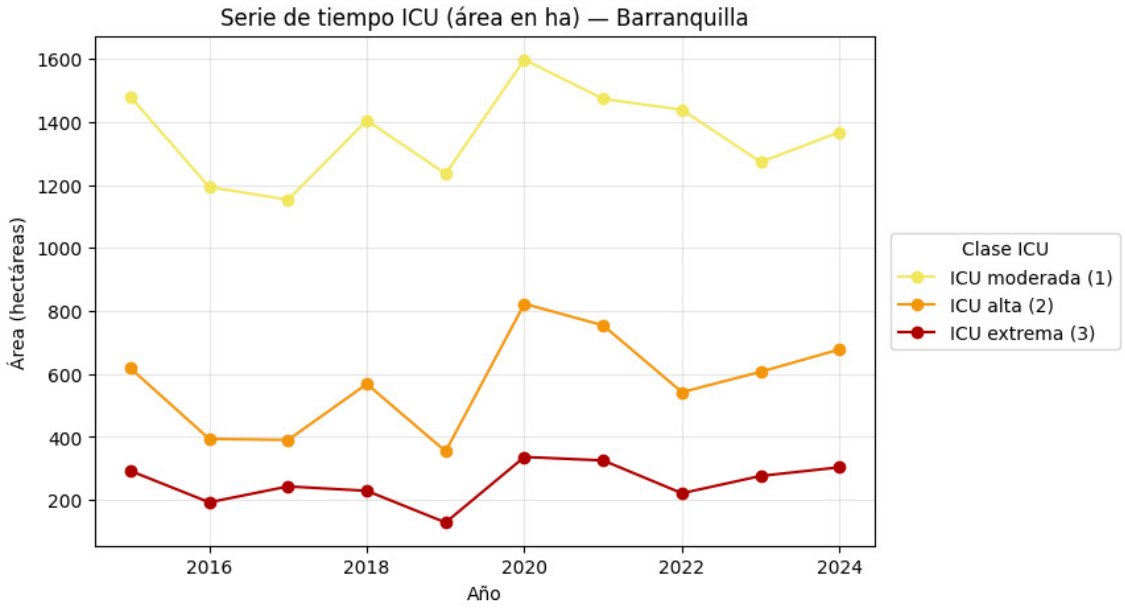


Figura 52: Serie de tiempo del área (ha) de ICU por clase — Barranquilla, 2015–2024.

Tabla 15: Barranquilla — estadísticas descriptivas del área (ha) por clase (2015–2024).

Clase	n	**Media (ha)**	**Desv. (ha)**	**CV (%)**	**Mín (ha) [año]**	**Máx (ha) [año]**
ICU moderada (1)	10	1361.93	143.70	10.55	1153.17 [2017]	1598.13 [2020]
ICU alta (2)	10	572.92	157.68	27.52	354.33 [2019]	822.87 [2020]
ICU extrema (3)	10	254.79	64.75	25.41	128.34 [2019]	336.51 [2020]

Santa Marta. En Santa Marta, la serie de tiempo del 2015 al 2024 (Figura 53) mostró que predominó el área con *ICU moderada (1)*, con una media cercana a 625 ha y variabilidad moderada (CV cercano a 20 %). La serie presentó un descenso inicial entre 2015 y 2018, un pico claro en 2020 (886 ha) y luego una caída hasta el mínimo en 2023 (473 ha), con recuperación parcial en 2024 (557 ha). *La ICU alta (2)* acompañó ese comportamiento: valores contenidos hasta 2019, máximo en 2020 (413 ha), descenso en 2021, leve rebote en 2022, mínimo en 2023 (221 ha) y repunte en 2024 (260 ha); su dispersión fue similar (CV cercano a 20 %). *La ICU extrema (3)* mantuvo los menores volúmenes (media cercana a 125 ha) y también sincronizó su máximo en 2020 (177 ha) y mínimo en 2023 (95 ha), con CV cercano

a 20 %. En conjunto, la tabla anual (Tabla 16) y la curva temporal evidenciaron un comportamiento pulsante con máximos simultáneos en 2020 para las tres categorías, seguidos de una contracción en 2023 y una recuperación parcial en 2024, coherente con las estadísticas descriptivas (Tabla 17), que mostraron medias, desviaciones estándar, coeficientes de variación y los años extremos para cada clase de ICU.

Tabla 16: Área en hectáreas (ha) de ICU por nivel de severidad en Santa Marta (2015–2024).

Año	ICU Moderada (ha)	ICU Alta (ha)	ICU Extrema (ha)
2015	779.40	363.69	155.97
2016	593.28	276.84	118.71
2017	553.14	257.67	111.15
2018	532.71	248.67	106.56
2019	619.65	288.99	124.11
2020	885.78	413.37	177.21
2021	567.18	264.69	113.49
2022	686.79	320.49	137.43
2023	473.04	220.68	94.77
2024	557.10	259.92	111.51

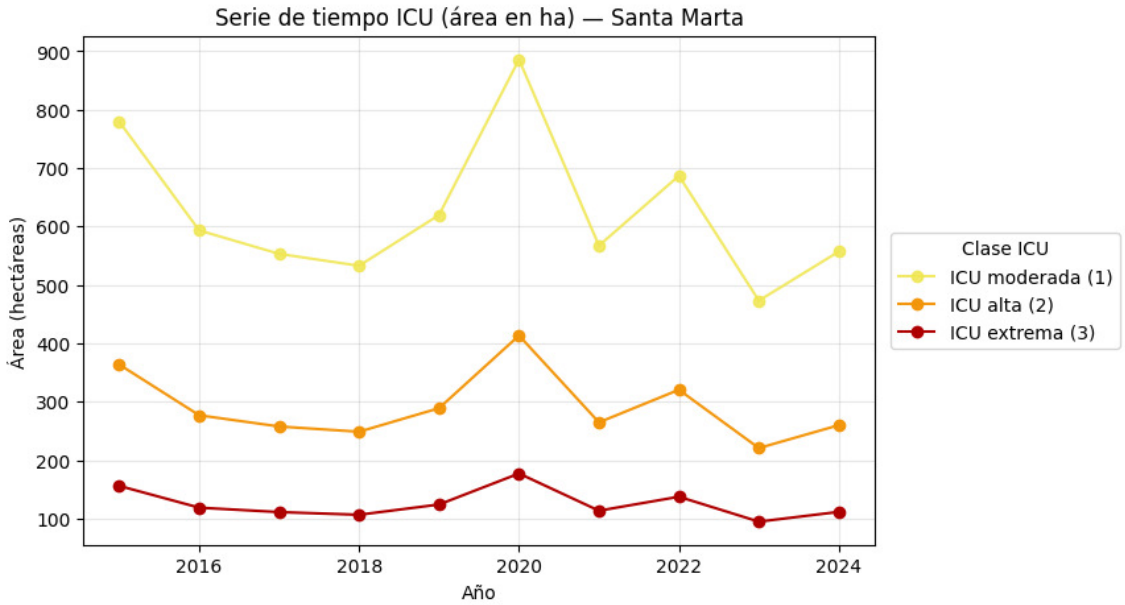


Figura 53: Serie de tiempo del área (ha) de ICU por clase — Santa Marta, 2015–2024.

Tabla 17: Santa Marta — estadísticas descriptivas del área (ha) por clase (2015–2024).

Clase	n	**Media (ha)**	**Desv. (ha)**	**CV (%)**	**Mín (ha) [año]**	**Máx (ha) [año]**
ICU moderada (1)	10	624.81	125.35	20.06	473.04 [2023]	885.78 [2020]
ICU alta (2)	10	291.50	58.53	20.08	220.68 [2023]	413.37 [2020]
ICU extrema (3)	10	125.09	25.03	20.01	94.77 [2023]	177.21 [2020]

Cartagena. En Cartagena, la serie de tiempo del 2015 al 2024 (Figura 54) mostró que predominó el área con *ICU moderada (1)*, con una media de 1176 ha y baja variabilidad (CV≈8.5 %). Esta presentó una trayectoria descendente suave desde su máximo en 2016 (1364 ha) hasta el mínimo en 2024 (1046 ha), con oscilaciones intermedias. Las clases más severas fueron más inestables: la *ICU alta (2)* alcanzó un pico en 2016 (658 ha), cayó hasta su mínimo en 2020 (140 ha), experimentó un repunte en 2023 (518 ha) y se moderó nuevamente en 2024 (271 ha), registrando una dispersión elevada (CV≈57 %). Por su parte, la *ICU extrema (3)* fue intermitente y de baja magnitud (media de 78 ha; CV≈144 %), con ausencia total o casi total en 2015, 2018, 2021–2022 y 2024, y máximos puntuales en 2016 (280 ha), 2017 (239 ha) y 2023 (189 ha). En conjunto, la tabla anual (Tabla 18) y la tabla de estadísticas descriptivas (Tabla 19) reflejaron una estabilidad relativa en la cobertura moderada y una alta volatilidad en las clases alta y extrema, con concentraciones máximas en 2016, un valle pronunciado en 2020 y un repunte no sostenido en 2023.

Tabla 18: Área en hectáreas (ha) de ICU por nivel de severidad en Cartagena (2015–2024).

Año	ICU Moderada (ha)	ICU Alta (ha)	ICU Extrema (ha)
2015	1297.98	236.52	0.00
2016	1364.13	657.99	279.72
2017	1204.11	571.32	238.95
2018	1053.45	204.39	0.00
2019	1215.18	521.91	60.30
2020	1160.73	139.86	13.50
2021	1112.94	143.19	0.00
2022	1153.80	188.46	0.00
2023	1152.72	517.59	188.64
2024	1046.07	271.26	0.81

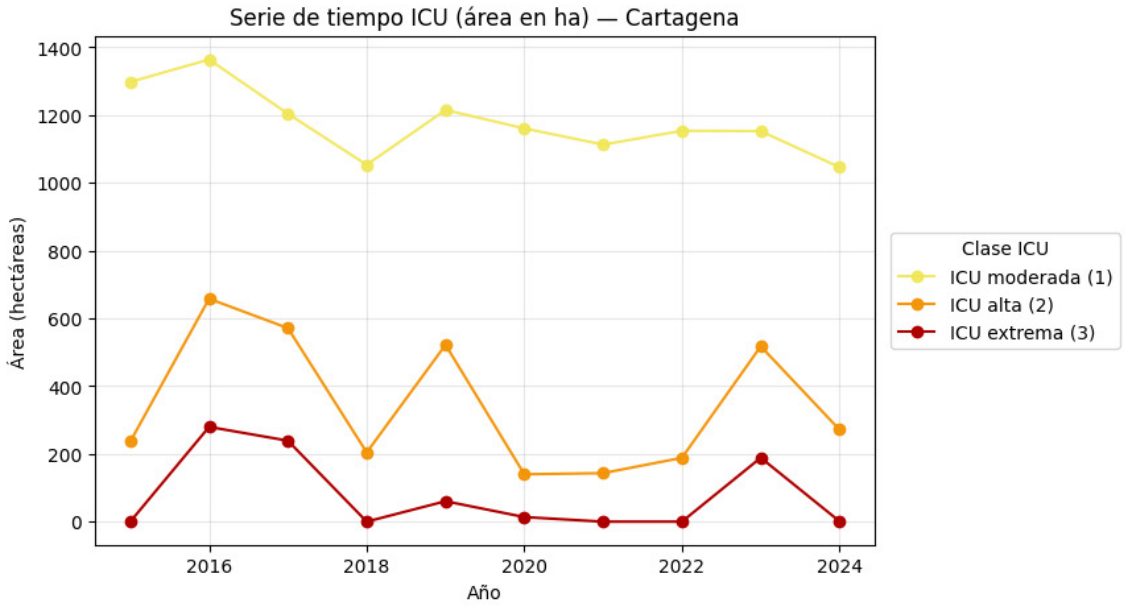


Figura 54: Serie de tiempo del área (ha) de ICU por clase — Cartagena, 2015–2024.

Tabla 19: Cartagena — estadísticas descriptivas del área (ha) por clase (2015–2024).

Clase	n	**Media (ha)**	**Desv. (ha)**	**CV (%)**	**Mín (ha) [año]**	**Máx (ha) [año]**
ICU moderada (1)	10	1176.11	99.86	8.49	1046.07 [2024]	1364.13 [2016]
ICU alta (2)	10	345.25	198.48	57.49	139.86 [2020]	657.99 [2016]
ICU extrema (3)	10	78.19	112.35	143.69	0.00 [2015]	279.72 [2016]

7. Conclusiones y trabajos futuros.

7.1. Conclusiones.

Las imágenes satelitales multiespectrales de la misión *Landsat* constituyen una fuente confiable y accesible para el estudio del comportamiento térmico superficial. Su resolución espacial y radiométrica resulta adecuada para la estimación de la temperatura terrestre, dado que el modelamiento de este parámetro no requiere de alta resolución espacial. Esto se confirma en la literatura científica, donde numerosos estudios utilizan este tipo de sensores para el análisis de la temperatura de la superficie terrestre (LST) y la identificación de islas de calor urbanas.

El uso de herramientas de *código abierto* permitió desarrollar un flujo de trabajo automatizado y reproducible de alto nivel. *PostgreSQL* y *QGIS*, en conjunto con el módulo *PyQGIS*, fueron esenciales para estructurar y ejecutar los procesos de preprocesamiento, muestreo y análisis de imágenes satelitales. Esta integración posibilitó la generación de insumos consistentes y la automatización de tareas críticas como la reproyección, limpieza y cálculo de índices espectrales. En estudios que requieren procesar grandes volúmenes de datos ráster, este tipo de estrategias reduce el tiempo de ejecución y aumenta la fiabilidad de los resultados.

Las bandas multiespectrales del infrarrojo demostraron ser variables de gran valor para la generación de índices espectrales como el *NDVI*, *NDBI* y *NDWI*, los cuales resultaron esenciales para la caracterización de coberturas y el modelamiento de la temperatura. Estos índices permiten realzar coberturas específicas combinando las bandas del espectro visible e infrarrojo cercano, mejorando la capacidad de los modelos para discriminar entre áreas vegetadas, construidas y cuerpos de agua. No obstante, el *NDWI* no fue finalmente incorporado en los modelos de clasificación ni en la estimación de temperatura. Su inclusión generaba inconsistencias en la representación de coberturas hídricas y aumentaba los falsos positivos, afectando la precisión y estabilidad del modelo. Esta decisión metodológica, aunque limitante, permitió conservar la coherencia de los resultados obtenidos y optimizar el desempeño general de los modelos.

En cuanto al desempeño de los algoritmos empleados, se observó que los modelos de *machine learning* (particularmente *Extra Trees* y *Gradient Boosting*) presentaron un mejor equilibrio entre precisión, estabilidad y tiempo de cómputo frente a las redes neuronales. Esto se explica porque, en este caso de estudio, las variables derivadas de imágenes satelitales presentan baja variabilidad intrínseca, lo que favorece los

métodos basados en árboles de decisión. Estos modelos mostraron alta capacidad de generalización y tiempos de entrenamiento sustancialmente menores, manteniendo un rendimiento comparable con arquitecturas de aprendizaje profundo.

Durante la construcción y validación del modelo de temperatura, se identificó una dependencia directa del número de estaciones meteorológicas disponibles. En zonas con mayor densidad de registros, como *Barranquilla* y *Santa Marta*, se logró una mejor representación espacial del fenómeno térmico. Sin embargo, la limitada cantidad de estaciones terrestres en algunas áreas motivó la inclusión de fuentes complementarias como *ERA5*, que permitió ampliar la cobertura temporal y espacial del análisis. Aun así, el contraste entre los datos reales y los derivados evidenció la importancia de fortalecer la red de observación terrestre, ya que la ausencia de mediciones en territorio reduce la capacidad predictiva de los modelos meteorológicos y térmicos.

El análisis multitemporal respaldó una relación directa entre las islas de calor y las zonas de mayor densificación urbana. Las ciudades costeras analizadas Barranquilla, Cartagena y Santa Marta mostraron comportamientos diferenciados pero coherentes con sus dinámicas urbanas:

- En **Barranquilla**, el fenómeno se concentró hacia el oriente, especialmente en los sectores adyacentes al río Magdalena.
- En **Santa Marta**, los focos térmicos persistieron en el núcleo central y el distrito norte.
- En **Cartagena**, aunque el patrón espacial fue menos estable por discontinuidades de datos, se identificaron zonas recurrentes de calor en el suroriente y áreas portuarias.

Finalmente, este tipo de estudios constituye un insumo estratégico para la gestión urbana y ambiental. La información generada permite a las entidades territoriales orientar políticas públicas hacia una planificación más equitativa y sostenible del territorio. La pérdida de zonas verdes y el incremento de superficies construidas intensifican el fenómeno de isla de calor, con implicaciones directas sobre la salud pública, el confort térmico y la calidad de vida urbana. Por tanto, el monitoreo continuo y el modelamiento predictivo de las islas de calor deben considerarse herramientas esenciales para la gestión del cambio climático y la adaptación de las ciudades del Caribe colombiano.

7.2. Trabajos futuros.

Los resultados obtenidos en este estudio permiten la exploración de diversas líneas de investigación complementarias. Una de las principales oportunidades de mejora reside en el fortalecimiento y la expansión de la red de estaciones meteorológicas terrestres, particularmente en las zonas urbanas y expansión urbana de las ciudades de Barranquilla, Cartagena y Santa Marta. Un incremento en el número de registros en sitio facilitaría la optimización de la calibración y validación de los modelos de estimación de temperatura, así como el desarrollo de sistemas de predicción continua que integren series temporales actualizadas y procesos de aprendizaje automático diarios.

Adicionalmente, los productos derivados de este trabajo, tales como los mapas térmicos, las clasificaciones de coberturas y los análisis multitemporales, constituyen una base sólida para la realización de estudios orientados a la formulación de estrategias de intervención y mitigación del fenómeno de isla de calor urbana. En ciudades costeras con alta densidad poblacional y relevancia turística, como las analizadas, estos resultados pueden contribuir a la elaboración de políticas de ordenamiento territorial, arborización urbana y diseño de infraestructura verde, promoviendo entornos más sostenibles frente a los impactos del cambio climático en la región Caribe colombiana.

Referencias

[1] I. G. A. Codazzi, "Colombia en mapas," https://www.colombiaenmapas.gov.co/, 2025, accessed: 2025-02-14.

[2] H. Tran, D. Uchihama, S. Ochi, and Y. Yasuoka, "Assessment with satellite data of the urban heat island effects in asian mega cities," *International Journal of Applied Earth Observation and Geoinformation*, vol. 8, no. 1, pp. 34–48, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0303243405000565

[3] D. Zhou, J. Xiao, S. Bonafoni, C. Berger, K. Deilami, Y. Zhou, S. Frolking, R. Yao, Z. Qiao, and J. A. Sobrino, "Satellite remote sensing of surface urban heat islands: Progress, challenges, and perspectives," *Remote Sensing*, vol. 11, no. 1, p. 48, 2018.

[4] C. S. Pérez and D. F. C. Castro, "Análisis espacial de islas de calor en la ciudad de bogotá: los efectos de la urbanización, un estudio desde la teledetección," Master's thesis, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia, 2019.

[5] G. Tanoori, A. Soltani, and A. Modiri, "Machine learning for urban heat island (uhi) analysis: Predicting land surface temperature (lst) in urban environments," *Urban Climate*, vol. 55, p. 101962, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212095524001585

[6] J. A. Orozco Triana, "Emprendimiento con alto potencial de crecimiento: recomendaciones para el caribe colombiano," *Universidad Tecnológica de Bolívar*, 2013.

[7] U.S. Environmental Protection Agency, "What are heat islands?" n.d., accessed: November 10, 2024. [Online]. Available: https://www.epa.gov/heatislands/learn-about-heat-islands#_ftn1

[8] Z.-L. Li, B.-H. Tang, H. Wu, H. Ren, G. Yan, Z. Wan, I. F. Trigo, and J. A. Sobrino, "Satellite-derived land surface temperature: Current status and perspectives," *Remote sensing of environment*, vol. 131, pp. 14–37, 2013.

[9] C. Kuenzer and S. Dech, "Thermal infrared remote sensing," *Remote Sensing and Digital Image Processing. doi*, vol. 10, no. 1007, pp. 978–94, 2013.

[10] E. Chuvieco, *Fundamentos de teledetección espacial*. Rialp, 1990.

[11] D. L. Williams, S. Goward, and T. Arvidson, "Landsat," *Photogrammetric Engineering & Remote Sensing*, vol. 72, no. 10, pp. 1171–1178, 2006.

[12] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut, "The era5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020. [Online]. Available: https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.3803

[13] Copernicus Climate Change Service (C3S). (2025) Era5: hourly data on single levels from 1940 to present — documentation. Climate Data Store (CDS) documentation y ficha del dataset. [Online]. Available: https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels

[14] Instituto Nacional de Estadística y Geografía (INEGI), "Índice de vegetación de diferencia normalizada (ndvi)," n.d., accessed: November 10, 2024. [Online]. Available: https://www.inegi.org.mx/investigacion/ndvi/

[15] T. Esch, W. Heldens, A. Hirner, M. Keil, M. Marconcini, A. Roth, J. Zeidler, S. Dech, and E. Strano, "Breaking new ground in mapping human settlements from space – the global urban footprint," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 134, pp. 30–42, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271617301880

[16] M. Perevochtchikova, "La situación actual del sistema de monitoreo ambiental en la zona metropolitana de la ciudad de méxico," *Estudios demográficos y urbanos*, vol. 24, no. 3, pp. 513–547, 2009.

[17] A. Arozarena Villar, I. Otero Pastor, and A. Ezquerra Canalejo, "Sistemas de captura de la información: fotogrametría y teledetección," 2016.

[18] G. A. F. Seber and A. J. Lee, *Linear Regression Analysis*, 2nd ed. John Wiley & Sons, 2012.

[19] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[20] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.

[21] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[22] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[23] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[24] scikit-learn developers, *LabelEncoder — scikit-learn preprocessing module*, scikit-learn, 2025, documentación oficial del módulo de codificación de etiquetas categóricas de scikit-learn. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

[25] ——, *GridSearchCV — scikit-learn model selection module*, scikit-learn, 2025, documentación oficial del módulo de búsqueda de hiperparámetros con validación cruzada en scikit-learn. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[26] ——, *StratifiedKFold scikit-learn model selection module*, scikit-learn, 2025, documentación oficial del módulo de validación cruzada estratificada en scikit-learn. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html

[27] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google earth engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, vol. 202, pp. 18–27, 2017, big Remotely Sensed Data: tools, applications and experiences. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425717302900

[28] The PostgreSQL Global Development Group. (2025) Postgresql 18.0 documentation. Manual oficial, versión actual. [Online]. Available: https://www.postgresql.org/docs/current/

[29] M. Stonebraker and L. A. Rowe, "The design of postgres," *SIGMOD Record*, 1986. [Online]. Available: https://dl.acm.org/doi/10.1145/16856.16888

[30] The PostGIS Development Team. (2025) Postgis documentation. Manual oficial, versión 3.6.x. [Online]. Available: https://postgis.net/docs/

[31] QGIS Development Team, *QGIS Geographic Information System*, QGIS Association, 2025. [Online]. Available: https://www.qgis.org

[32] GDAL Development Team, *GDAL — Geospatial Data Abstraction Library*, Open Source Geospatial Foundation, 2025, documentación oficial de la biblioteca GDAL para procesamiento geoespacial. [Online]. Available: https://gdal.org/

[33] GRASS Development Team, *GRASS GIS — Geographic Resources Analysis Support System*, Open Source Geospatial Foundation (OSGeo), 2025, sistema libre de procesamiento geoespacial y teledetección. [Online]. Available: https://grass.osgeo.org/

[34] QGIS Development Team, *PyQGIS — QGIS Python API*, QGIS Association, 2025, documentación oficial de la API de Python para QGIS. [Online]. Available: https://qgis.org/pyqgis/3.40/

[35] *GeoPackage An Open Format for Geospatial Information*, Open Geospatial Consortium Std., 2024, Última versión aprobada del estándar para el formato GeoPackage (.gpkg). [Online]. Available: https://www.geopackage.org

[36] QGIS Development Team, *QGIS Processing Algorithm: Minimum Bounding Geometry (Bounding Box option)*, QGIS Association, 2025, documento oficial de QGIS, sección "Minimum Bounding Geometry". [Online]. Available: https://docs.qgis.org/latest/en/docs/user_manual/processing_algs/qgis/vectorgeometry.html#minimum-bounding-geometry

[37] G. Kaur, K. S. Saini, D. Singh, and M. Kaur, "A comprehensive study on computational pansharpening techniques for remote sensing images," *Archives of Computational Methods in Engineering*, pp. 1–18, 2021.

[38] QGIS Development Team, *Raster Calculator — QGIS Processing Algorithm*, QGIS Association, 2025, documento oficial de QGIS, sección R̈aster Calculator.̈ [Online]. Available: https://docs.qgis.org/3.40/en/docs/user_manual/processing_algs/gdal/rastermiscellaneous.html

[39] GRASS Development Team, *r.fillnulls — GRASS GIS module for filling null cells in raster maps*, Open Source Geospatial Foundation (OSGeo), 2025, módulo de interpolación de GRASS GIS para rellenar celdas nulas en datos ráster. [Online]. Available: https://grass.osgeo.org/grass-stable/manuals/r.fillnulls.html

[40] GDAL Development Team, *gdal:merge — GDAL Raster Merge Algorithm*, Open Source Geospatial Foundation (OSGeo), 2025, algoritmo oficial de GDAL para combinar múltiples ráster en un único archivo. [Online].

Available: https://docs.qgis.org/latest/en/docs/user_manual/processing_algs/gdal/rastermiscellaneous.html#merge

[41] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th ed. Hoboken, NJ: John Wiley & Sons, 2012.

[42] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.

[43] T. Chai and R. R. Draxler, "Root mean square error (rmse) or mean absolute error (mae)? arguments against avoiding rmse in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014.

[44] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008. [Online]. Available: https://nlp.stanford.edu/IR-book/information-retrieval-book.html

[45] D. M. W. Powers, "Evaluation: From precision, recall and f-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011. [Online]. Available: https://bioinfopublication.org/files/articles/2_1_1_JMLT.pdf

[46] Y. Sasaki, "The truth of the F-measure," School of Computer Science, University of Manchester, Tech. Rep., 2007, technical report. [Online]. Available: https://people.cs.pitt.edu/~litman/courses/cs1671s20/F-measure-YS-26Oct07.pdf

[47] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009. [Online]. Available: https://engold.ui.ac.ir/~h.marateb/SokolovaLapalme-JIPM09.pdf

[48] N. A. S. Kasniza Jumari, A. N. Ahmed, Y. F. Huang, J. L. Ng, C. H. Koo, K. L. Chong, M. Sherif, and A. Elshafie, "Analysis of urban heat islands with landsat satellite images and gis in kuala lumpur metropolitan city," *Heliyon*, vol. 9, no. 8, p. e18424, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405844023056323

[49] U. G. Survey, "Earthexplorer," https://earthexplorer.usgs.gov/, 2025, accessed: 2025-02-14.

[50] Instituto Geográfico Agustín Codazzi, "Resolución No. 370 de 2021: Por medio de la cual se establece el sistema de proyección cartográfica oficial pa-

ra Colombia," Bogotá, Colombia, Jun. 2021, https://redgeodesica.igac.gov.co/documentos/resolucion_370_de_2021.pdf.

[51] U.S. Geological Survey, "Landsat surface reflectance-derived spectral indices and scaling factors," U.S. Geological Survey, Tech. Rep., 2023, product Guide, Landsat Collection 2 Level-2 Science Products. [Online]. Available: https://www.usgs.gov/media/files/landsat-8-9-olitirs-collection-2-level-2-data-format-control-book

[52] J. Xue and B. Su, "Significant remote sensing vegetation indices: A review of developments and applications," *Journal of sensors*, vol. 2017, no. 1, p. 1353691, 2017.

[53] Y. Zha, J. Gao, and S. Ni, "Use of normalized difference built-up index in automatically mapping urban areas from tm imagery," *International journal of remote sensing*, vol. 24, no. 3, pp. 583–594, 2003.

[54] S. K. McFeeters, "The use of the normalized difference water index (ndwi) in the delineation of open water features," *International journal of remote sensing*, vol. 17, no. 7, pp. 1425–1432, 1996.

[55] T. Chakraborty, X. Lee, S. Ermida, and W. Zhan, "On the land emissivity assumption and landsat-derived surface urban heat islands: A global analysis," *Remote Sensing of Environment*, vol. 265, p. 112682, 2021.

[56] Z.-L. Li, H. Wu, S.-B. Duan, W. Zhao, H. Ren, X. Liu, P. Leng, R. Tang, X. Ye, J. Zhu *et al.*, "Satellite remote sensing of global land surface temperature: Definition, methods, products, and applications," *Reviews of Geophysics*, vol. 61, no. 1, 2023.

[57] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[58] P.-F. Hsieh, L. C. Lee, and N.-Y. Chen, "Effect of spatial resolution on classification errors of pure and mixed pixels in remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 12, pp. 2657–2663, 2001.

[59] T. Sun, R. Sun, and L. Chen, "The trend inconsistency between land surface temperature and near surface air temperature in assessing urban heat island effects," *Remote sensing*, vol. 12, no. 8, p. 1271, 2020.

[60] A. N. Ahmed, N. AlDahoul, N. A. Aziz, Y. Huang, M. Sherif, and A. El-Shafie, "The urban heat island effect: A review on predictive approaches using artificial intelligence models," *City and Environment Interactions*, p. 100234, 2025.

[61] J. Voogt and T. Oke, "Thermal remote sensing of urban climates," *Remote Sensing of Environment*, vol. 86, no. 3, pp. 370–384, 2003, urban Remote Sensing. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425703000798

[62] S. Peng, S. Piao, P. Ciais, P. Friedlingstein, C. Ottle, F.-M. Bréon, H. Nan, L. Zhou, and R. B. Myneni, "Surface urban heat island across 419 global big cities," *Environmental science & technology*, vol. 46, no. 2, pp. 696–703, 2012.

[63] N. Clinton and P. Gong, "Modis detected surface urban heat islands and sinks: Global locations and controls," *Remote Sensing of Environment*, vol. 134, pp. 294–304, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425713000874