



ugr | Universidad
de Granada

TRABAJO FIN DE GRADO

DOBLE GRADO EN INGENIERÍA INFORMÁTICA Y ADMINISTRACIÓN Y DIRECCIÓN DE EMPRESAS

Análisis y Optimización de Rutas Aéreas Comerciales mediante Ciencia de Datos

Autor

José Antonio Fernández Aranda

Director

Jorge Casillas Barranquero



Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

—
Granada, 16 de junio de 2025



Análisis y Optimización de Rutas Áreas Comerciales mediante Ciencia de Datos

Autor

José Antonio Fernández Aranda (alumno)

Directores

Jorge Casillas Barranquero (tutor)

Análisis y Optimización de Rutas Aéreas Comerciales mediante Ciencia de Datos

José Antonio Fernández Aranda

Palabras clave: Rutas aéreas, ADS-B, OpenSky, Clustering, Machine Learning, Turismo, Huella de carbono, Python, Optimización, Web scraping, Eficiencia operativa, Trayectorias reales, Visualización de datos, Sostenibilidad.

Resumen:

Este Trabajo Fin de Grado aborda el análisis avanzado de rutas aéreas reales a partir de datos abiertos procedentes de plataformas como OpenSky y OpenFlights, con el objetivo de estudiar la eficiencia operativa, el impacto ambiental y su relación con los flujos turísticos internacionales. A través de técnicas de scraping, limpieza de datos y modelado computacional, se ha desarrollado una infraestructura completa de adquisición, tratamiento y análisis de vuelos y aeronaves, integrada con fuentes turísticas y económicas.

Se emplean algoritmos de aprendizaje automático no supervisado (como K-Means, DBSCAN y Birch) para detectar patrones, clasificar trayectorias y evaluar comportamientos anómalos. Asimismo, se analizan factores como el tipo de aeronave, la duración estimada, el operador aéreo o la zona horaria para identificar inefficiencias estructurales o implicaciones logísticas. Los resultados se complementan con visualizaciones interactivas y representaciones avanzadas como dendogramas, Sankey circulares y gráficos de radar.

El estudio concluye con una exploración de los perfiles de sostenibilidad aérea combinados con indicadores turísticos, desarrollando clústeres de países según su rendimiento ambiental y económico en relación con el transporte aéreo. Todo el proyecto ha sido desarrollado íntegramente con Python y herramientas de ciencia de datos modernas, siguiendo una estructura replicable y orientada a casos reales, con el propósito de ofrecer valor analítico tanto en el ámbito aeronáutico como turístico.

Analysis and Optimization of Commercial Air Routes using Data Science

José Antonio Fernández Aranda

Keywords: Flight routes, ADS-B, OpenSky, Clustering, Machine Learning, Tourism, Carbon footprint, Python, Optimization, Web scraping, Operational efficiency, Real trajectories, Data visualization, Sustainability.

Abstract:

This Final Degree Project focuses on the advanced analysis of real commercial flight trajectories using open data sources such as OpenSky and OpenFlights. The primary objective is to study operational efficiency, environmental impact, and their relationship with international tourism flows. Through scraping techniques, data cleaning, and computational modeling, a complete infrastructure for acquiring, processing, and analyzing flight and aircraft data has been developed and integrated with economic and tourism datasets.

Unsupervised machine learning algorithms (such as K-Means, DBSCAN, and Birch) are employed to detect patterns, classify trajectories, and assess anomalous behavior. The study considers variables such as aircraft type, estimated duration, airline operator, and time zones to identify structural inefficiencies and logistical implications. The results are reinforced with advanced visualizations including dendograms, circular Sankey diagrams, and radar plots.

The project culminates in a clustering-based analysis of sustainability profiles, combining flight emissions with tourism indicators to identify countries with optimal or inefficient performance. The entire workflow has been implemented in Python using modern data science tools, following a reproducible structure and real-world case studies, with the aim of providing analytical value to both the aviation and tourism sectors.

Yo, **José Antonio Fernández Aranda**, alumno de la titulación **Doble Grado en Ingeniería Informática y Administración y Dirección de Empresas** de la **Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación** de la **Universidad de Granada**, con DNI 14274791W, autorizo la ubicación de la siguiente copia de mi Trabajo Fin de Grado en la biblioteca del centro para que pueda ser consultada por las personas que lo deseen.

Fdo: José Antonio Fernández Aranda

Granada, a 16 de junio de 2025.

D. **Jorge Casillas Barranquero**, Profesor del Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada.

Informan: Que el presente trabajo, titulado *Análisis y Optimización de Rutas Aéreas Comerciales mediante Ciencia de Datos*, ha sido realizado bajo su supervisión por **José Antonio Fernández Aranda**, y autorizamos la defensa de dicho trabajo ante el tribunal que corresponda.

Y para que conste, expiden y firman el presente informe en Granada, a 16 de junio de 2025.

El director:



Jorge Casillas Barranquero

Agradecimientos

Este Trabajo Fin de Grado marca el cierre de una etapa vital en mi formación, y no habría sido posible sin el apoyo constante de muchas personas que me han acompañado, directa o indirectamente, a lo largo de estos años.

En primer lugar, quiero expresar mi gratitud más profunda a mi padre y a mi abuela. Han sido mi ejemplo y mi refugio, sosteniéndome con su cariño incondicional, su sabiduría y su constancia. Gracias por estar siempre, por empujarme a ser mejor y por confiar en mí incluso cuando yo dudaba.

A mi pareja, por acompañarme tanto en los momentos de calma como en los de mayor presión. Gracias por tu paciencia, tu comprensión y tu forma de estar, que ha sido esencial para mantener el equilibrio en esta etapa exigente. Tu apoyo ha sido un ancla y una motivación, y me siento afortunado de haber compartido este camino contigo.

A mis compañeros de clase, especialmente por confiar en mí como delegado. Ha sido una experiencia enriquecedora que me ha permitido crecer también en lo humano. Gracias por construir conmigo un ambiente de compañerismo, implicación y responsabilidad compartida.

A los profesores de la Escuela, por su dedicación, su exigencia y por sembrar en nosotros una mirada crítica y rigurosa hacia la ingeniería. Habéis sido una parte esencial de este proceso, y vuestra compromiso con nuestra formación ha dejado huella.

A mi tutor, por su orientación clara, su criterio técnico y su confianza en el enfoque del proyecto. Su acompañamiento ha sido clave para dar forma a este trabajo y para afrontar sus desafíos con perspectiva.

A mis compañeros del Colegio Mayor Albayzín, por haber sido una familia durante esta etapa. Gracias por los valores compartidos, las conversaciones profundas y los momentos cotidianos que tanto han significado.

A mis amigos del Erasmus, por ampliar mis horizontes, por la espontaneidad, y por la amistad que surgió más allá de las fronteras. Compartir esa experiencia con vosotros ha sido un regalo que recordaré siempre.

Y por supuesto, a mis amigos de siempre, por estar. Por el apoyo constante, por los buenos ratos, por escuchar sin juzgar y por celebrar cada pequeño paso. Vuestra presencia ha sido una constante de alegría y equilibrio en medio del esfuerzo.

Este trabajo no es solo el fruto de un análisis técnico, sino también de una red de personas que han hecho posible que llegara hasta aquí. A todos vosotros, gracias.

Índice general

Índice de figuras	16
1. Introducción	1
1.1. Contexto del transporte aéreo comercial	1
1.1.1. Evolución histórica y tecnológica de la aviación	1
1.1.2. Situación actual y retos operativos	2
1.2. Motivación y relevancia del estudio	3
1.3. Objetivos generales y específicos	3
1.4. Metodología general del proyecto	5
1.5. Estructura del documento	5
2. Marco teórico y estado del arte	7
2.1. Ciencia de datos en el transporte aéreo	7
2.2. Métricas de eficiencia operativa	8
2.3. Rutas óptimas y algoritmos de grafos	10
2.4. Emisiones y sostenibilidad en la aviación	12
2.5. Interrelación turismo–movilidad aérea	14
3. Recopilación e integración de datos	17
3.1. Fuentes utilizadas	17
3.2. Tipos de datos y formatos	18
3.3. Técnicas de scraping y APIs	20
3.4. Descripción de variables clave	23
4. Preprocesamiento de datos	27
4.1. Limpieza, normalización y consolidación	27
4.2. Conversión temporal y geoespacial	28
4.3. Integración de datasets por caso de estudio	30
5. Análisis exploratorio general (EDA)	33
5.1. Estadísticas descriptivas globales	33
5.2. Visualización geoespacial	35
5.3. Identificación de outliers y anomalías	37
6. Casos de estudio y análisis comparativo	41
6.1. Estudio 1: Datos de Vuelos Comerciales	41
6.2. Estudio 2: Flota Aérea (Datos de Aviones)	45
6.3. Estudio 3: Vuelos + Emisiones de CO ₂	49
6.4. Estudio 4: Vuelos + Indicadores Turísticos	55
6.5. Estudio 5: Análisis combinado Vuelos, Turismo y Emisiones de CO ₂	61

7. Algoritmos de clustering y detección de patrones	67
7.1. Justificación del enfoque	67
7.2. Selección y configuración de algoritmos	69
7.3. Resultados y análisis de agrupaciones	71
8. Discusión crítica y evaluación	77
8.1. Validación de resultados	77
8.2. Limitaciones y posibles sesgos	78
8.3. Relevancia práctica y social	80
9. Conclusiones y líneas futuras	83
9.1. Conclusiones principales	83
9.2. Propuestas de mejora	83
9.3. Aplicaciones y posibles extensiones	84
Bibliografía	84
Anexo A. Repositorio del código fuente	87
Anexo B. Glosario	89

Índice de figuras

1.1. Principales hitos en la historia de la aviación comercial. Elaboración propia.	2
2.1. Mapa geográfico de rutas aéreas reales consideradas en el análisis. Cada marcador representa un aeropuerto con conexiones activas en los conjuntos de datos recopilados. Las líneas verdes indican vuelos comerciales registrados, conectando regiones de Europa, América, Asia y Oceanía, y representan además la distancia en kilómetros. Esta visualización fue generada mediante herramientas de geolocalización y procesamiento de datos abiertos. Elaboración propia.	12
5.1. Distribución de duración de vuelos comerciales. Se observa un patrón sesgado a la derecha, típico de la operativa aérea real. Elaboración propia.	34
5.2. Relación entre duración del vuelo y emisiones estimadas de CO ₂ . Se observa mayor densidad en vuelos prolongados, junto con casos extremos por encima de los umbrales normativos. Elaboración propia.	39
6.1. Matriz de correlación entre variables temporales y operativas.	42
6.2. Dendrograma jerárquico obtenido con clustering aglomerativo. Elaboración propia.	43
6.3. Proyección MDS con agrupamientos por algoritmo de clustering. Elaboración propia.	44
6.4. Duración media por tipo de aeronave. Elaboración propia.	44
6.5. Clustering de tipos de aeronave en dos componentes principales (PCA).	46
6.6. Boxplot de duración media por clúster.	47
6.7. Número total de vuelos por grupo de modelos.	47
6.8. Comparativa de errores por modelo (MAE).	48
6.9. Relación entre duración real y predicción.	48
6.10. Emisiones promedio por zona horaria de salida (clasificadas por región). Elaboración propia.	50
6.11. Emisiones promedio por vuelo en EE.UU. (Costa Este vs. Oeste). Elaboración propia.	50
6.12. Emisiones promedio por operador (kg CO ₂ por vuelo). Elaboración propia.	51
6.13. Clustering de rutas por intensidad ecológica. Elaboración propia.	52
6.14. Eficiencia ecológica por tipo de avión. Elaboración propia.	53
6.15. Clustering de rutas por distancia y emisiones (versión 1). Elaboración propia.	53
6.16. Clustering de rutas por distancia y emisiones (versión 2). Elaboración propia.	54
6.17. Mapa de rutas aéreas con codificación ecológica por intensidad. Elaboración propia.	54
6.18. Top 10 países con más vuelos. Elaboración propia.	56
6.19. Relación entre número de vuelos y llegadas turísticas medias. Elaboración propia.	56
6.20. Relación vuelos-turismo para países con más de 50M de llegadas. Elaboración propia.	57

6.21. Top 10 países con mayor duración media de vuelos. Elaboración propia.	57
6.22. Top 10 países por ingresos turísticos medios. Elaboración propia.	58
6.23. Clustering turístico-aeronáutico por PCA (k=3 y k=4).	59
6.24. Proyección de clustering en MDS 2D (k=4).	59
6.25. Dendrograma jerárquico de países turísticos (linkage Ward).	60
6.26. Clustering de países turísticos con DBSCAN proyectado en MDS.	61
6.27. Matriz de correlación entre vuelos, emisiones, turistas e ingresos.	63
6.28. Relaciones entre variables clave. Elaboración propia.	63
6.29. Clustering turístico-sostenible por PCA. Elaboración propia.	64
6.30. Radar comparativo de perfiles turístico-sostenibles por país.	65
7.1. Diagrama circular de flujos internacionales entre países (ordenado por huso horario).	74

Capítulo 1

Introducción

1.1. Contexto del transporte aéreo comercial

1.1.1. Evolución histórica y tecnológica de la aviación

La evolución de la aviación comercial ha estado estrechamente vinculada al desarrollo tecnológico y a la necesidad creciente de transportar personas y mercancías de forma rápida, segura y eficiente. Desde los primeros vuelos experimentales hasta las actuales redes de tráfico aéreo globalizado, la historia de la aviación refleja un proceso continuo de innovación, estandarización y adaptación.

El primer vuelo comercial reconocido tuvo lugar el 1 de enero de 1914, cubriendo la ruta entre St. Petersburg y Tampa (Florida, EE. UU.), marcando así el inicio de una industria que, décadas más tarde, transformaría la movilidad global. A lo largo del siglo XX, el progreso técnico —incluyendo la herencia tecnológica de los conflictos bélicos— facilitó la expansión de la aviación civil, con mejoras en la seguridad, la capacidad de las aeronaves y el alcance de las rutas.

Un hito clave fue la firma del Convenio de Chicago en 1944, que dio origen a la Organización de Aviación Civil Internacional (OACI / ICAO) y sentó las bases normativas para la cooperación internacional en la gestión del espacio aéreo. La llegada del motor a reacción (con el de Havilland Comet) y más tarde el Boeing 747 (1969) revolucionaron el transporte aéreo de masas, democratizando su acceso y consolidándolo como infraestructura estratégica para la economía global.

En los años 90, el sector incorporó progresivamente tecnologías digitales: navegación por satélite (GNSS), comunicación aire-tierra (ACARS), y reducción de separación vertical mínima (RVSM), lo que supuso un avance importante en la eficiencia y seguridad del tráfico aéreo. Este proceso se aceleró aún más con el desarrollo de sistemas como ADS-B (*Automatic Dependent Surveillance–Broadcast*), que permite conocer en tiempo real la posición, altitud, velocidad y trayectoria de aeronaves a nivel global.

Desde la aparición de proyectos como OpenSky Network (2012) y la posterior extensión obligatoria del uso de ADS-B en 2020, los datos abiertos aeronáuticos se han convertido en una fuente rica y precisa para investigadores, operadores y desarrolladores de soluciones basadas en datos. La intersección entre tecnología aeronáutica y ciencia de datos ha abierto nuevas posibilidades para analizar, optimizar y simular el comportamiento del tráfico aéreo en un entorno realista.

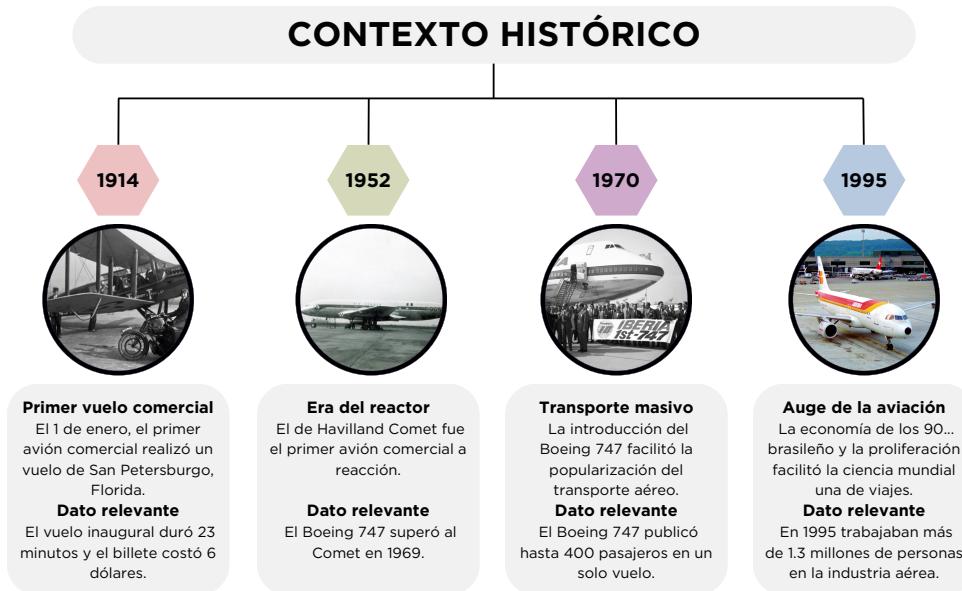


Figura 1.1: Principales hitos en la historia de la aviación comercial. Elaboración propia.

1.1.2. Situación actual y retos operativos

En la actualidad, el transporte aéreo comercial constituye una de las infraestructuras críticas de la movilidad internacional. Más allá de su capacidad para conectar territorios, fomentar el turismo o facilitar el comercio global, la aviación moderna se enfrenta a una serie de desafíos estructurales que afectan a su sostenibilidad, eficiencia y resiliencia.

Uno de los principales retos operativos es la saturación progresiva del espacio aéreo en determinadas regiones del mundo. El crecimiento sostenido del número de vuelos —impulsado por el aumento de la demanda, la expansión de aerolíneas de bajo coste y la globalización económica— ha tensionado la capacidad de los sistemas de control de tráfico aéreo, generando cuellos de botella, demoras y consumo adicional de combustible.

Asimismo, la presión ambiental sobre el sector ha aumentado de forma significativa. Según datos de la Asociación Internacional de Transporte Aéreo (IATA), antes de la pandemia de COVID-19, el sector aéreo transportaba más de 4.500 millones de pasajeros y generaba cerca del 3,5 % del PIB mundial, así como aproximadamente 65 millones de empleos directos e indirectos [11]. Además, la Organización de Aviación Civil Internacional (ICAO) ha establecido estándares globales para la planificación aérea, navegación y reducción de emisiones [?].

La gestión eficiente de las rutas aéreas cobra, por tanto, un valor estratégico tanto para operadores como para organismos reguladores. Una ruta mal optimizada implica no solo mayores costes operativos, sino también un aumento directo del consumo de combustible, tiempos de vuelo más largos y mayores emisiones contaminantes. Además, la conectividad aérea debe evaluarse también desde una perspectiva socioeconómica, considerando el equilibrio entre accesibilidad, demanda turística y sostenibilidad del destino.

En paralelo, la industria dispone hoy de una disponibilidad sin precedentes de datos aeronáuticos en tiempo real, así como de herramientas analíticas avanzadas capaces de procesar, modelar y visualizar dicha información a gran escala. Tecnologías como ADS-B, junto

con repositorios de datos abiertos como OpenSky Network y OpenFlights, permiten analizar trayectorias reales, estimar métricas de rendimiento operativo y medioambiental, e incluso simular rutas alternativas basadas en criterios de eficiencia.

Este entorno, marcado por la necesidad de optimización y la disponibilidad de datos de alta resolución, ofrece una oportunidad única para aplicar ciencia de datos en la mejora integral del transporte aéreo, orientando las decisiones hacia una aviación más inteligente, sostenible y adaptada a los retos del siglo XXI.

1.2. Motivación y relevancia del estudio

La realización de este Trabajo de Fin de Grado surge de la inquietud por explorar cómo el análisis de datos puede aplicarse de forma práctica y rigurosa a un sector tan complejo y estructural como la aviación comercial. En un contexto global cada vez más orientado hacia la sostenibilidad y la eficiencia, el transporte aéreo enfrenta el desafío de mantener su papel como motor económico sin comprometer sus obligaciones medioambientales.

Desde la perspectiva del estudiante de Ingeniería Informática y Administración de Empresas, este proyecto representa una oportunidad idónea para integrar conocimientos técnicos —como la adquisición y procesamiento de datos, el uso de algoritmos de análisis y optimización, y la representación geoespacial de la información— con criterios económicos, turísticos y ecológicos propios de la gestión empresarial sostenible.

En particular, se ha identificado un área de oportunidad en la explotación de datos abiertos provenientes de distintas fuentes públicas y especializadas. Estos datos, si se tratan adecuadamente, permiten no solo describir el estado actual del transporte aéreo, sino también modelar escenarios, detectar patrones, identificar rutas ineficientes y proponer soluciones con base en criterios cuantificables.

La motivación principal del estudio reside, por tanto, en aplicar técnicas de ciencia de datos a un entorno real y altamente relevante, para dar respuesta a cuestiones como:

- ¿Es posible identificar trayectorias comerciales que puedan optimizarse para reducir emisiones sin afectar la conectividad?
- ¿Qué papel juega la tipología de aeronaves en la eficiencia operativa?
- ¿Qué relación existe entre las rutas aéreas y los flujos turísticos?
- ¿Puede diseñarse una metodología de análisis que combine aspectos técnicos, medioambientales y económicos en una sola estructura?

Este trabajo no solo tiene valor desde el punto de vista técnico, sino que también pretende ser una aportación al debate sobre la sostenibilidad del transporte aéreo y su papel dentro del modelo turístico global. A través del cruce e integración de datos de vuelos, turismo y emisiones de CO₂, se pretende construir una visión multidimensional que permita evaluar de manera más holística la eficiencia y sostenibilidad de las rutas aéreas comerciales actuales.

1.3. Objetivos generales y específicos

El presente Trabajo de Fin de Grado se plantea como una oportunidad para aplicar de forma integrada conocimientos adquiridos en el ámbito de la Ingeniería Informática y la

Administración de Empresas. A través de un proyecto con un fuerte componente técnico y analítico, se aborda un problema de interés real y creciente: la necesidad de evaluar y optimizar la eficiencia de las rutas aéreas comerciales bajo criterios operativos, medioambientales y turísticos.

El trabajo no solo pretende aportar una solución tecnológica basada en datos, sino también desarrollar competencias clave en el análisis de grandes volúmenes de información, en la toma de decisiones basada en métricas cuantificables y en el uso responsable de la tecnología para la sostenibilidad. Esto incluye habilidades en programación, tratamiento de datos heterogéneos, uso de librerías analíticas y visuales, diseño de algoritmos y comunicación de resultados con rigor académico.

Objetivo general

Diseñar y desarrollar una solución analítica basada en ciencia de datos que permita evaluar trayectorias aéreas comerciales en términos de eficiencia operativa, impacto ambiental y sostenibilidad turística, integrando datos abiertos, técnicas de análisis exploratorio, algoritmos de optimización y visualización geoespacial.

Objetivos específicos

1. **Adquirir y estructurar datos abiertos de diversas fuentes**, incluyendo datos de vuelos (OpenSky Network), rutas y flotas (OpenFlights), indicadores turísticos (Banco Mundial) y emisiones de CO₂, mediante técnicas de *scraping* y consulta de APIs.
2. **Aplicar procesos de limpieza, transformación y normalización de datos**, asegurando su compatibilidad e integridad para el análisis conjunto, y desarrollando scripts reproducibles en Python utilizando herramientas como Pandas, NumPy o GeoPandas.
3. **Realizar un análisis exploratorio de datos (EDA)** para detectar patrones de comportamiento, desviaciones y correlaciones relevantes entre variables técnicas (trayectorias, aeronaves), medioambientales (emisiones) y socioeconómicas (turismo).
4. **Formular y calcular métricas de eficiencia operativa y sostenibilidad**, como la distancia relativa de ruta, emisiones normalizadas por pasajero-kilómetro, ocupación estimada o intensidad turística por ruta aérea.
5. **Aplicar algoritmos de aprendizaje no supervisado**, concretamente técnicas de *clustering* (K-Means, DBSCAN), para segmentar rutas según su comportamiento y descubrir agrupaciones con características similares.
6. **Emplear algoritmos de optimización para simular rutas** alternativas más eficientes o con menor impacto ambiental, manteniendo la conectividad funcional.
7. **Diseñar un sistema comparativo de análisis por casos**, estructurado en cinco estudios que integran progresivamente diferentes dimensiones: vuelos, tipo de aeronave, emisiones de CO₂, datos turísticos y el cruce total de todas ellas.
8. **Fomentar la toma de decisiones basada en datos**, generando visualizaciones interpretables y documentando resultados que puedan informar tanto a actores técnicos como a responsables de políticas de transporte y turismo sostenible.

1.4. Metodología general del proyecto

El enfoque metodológico adoptado en este trabajo se basa en la aplicación de técnicas de ciencia de datos para la recopilación, procesamiento, análisis y modelado de información relativa al transporte aéreo comercial, sus impactos medioambientales y su relación con indicadores turísticos. La metodología se ha estructurado en varias fases diferenciadas, cada una de ellas diseñada para responder a un conjunto específico de objetivos técnicos y analíticos.

En primer lugar, se ha llevado a cabo un proceso de **adquisición de datos** abiertos desde múltiples fuentes relevantes. Entre ellas se encuentran OpenSky Network, que proporciona información en tiempo real sobre vuelos; OpenFlights, con datos estructurados sobre aeropuertos, aerolíneas y rutas; el Banco Mundial, que ofrece indicadores turísticos a nivel global; y fuentes oficiales sobre emisiones de CO₂. Parte de la información ha sido obtenida mediante *scraping* web automatizado, lo que ha requerido el desarrollo de scripts en Python capaces de extraer, estructurar y almacenar los datos de forma controlada.

Posteriormente, se ha realizado un trabajo de **preprocesamiento**, que incluyó la limpieza, estandarización, filtrado y transformación de los datos. Este paso fue esencial para garantizar la calidad y coherencia de los conjuntos de datos fusionados, particularmente en lo que respecta a la normalización de identificadores de vuelos, fechas, unidades de medida y ubicaciones geográficas.

La fase de **análisis exploratorio (EDA)** permitió obtener una primera comprensión de los patrones y relaciones presentes en los datos. A través de representaciones estadísticas y visualizaciones geoespaciales, se identificaron tendencias en la distribución de rutas, tipologías de aeronaves, frecuencias de vuelo y cobertura geográfica.

A continuación, se definieron e implementaron un conjunto de **métricas de eficiencia operativa y sostenibilidad**, que permitieron evaluar el rendimiento de las rutas desde distintas perspectivas. También se aplicaron **algoritmos de clustering** para detectar agrupaciones de rutas con características similares, y se desarrolló una propuesta de optimización de trayectorias aéreas utilizando técnicas basadas en teoría de grafos.

Finalmente, el estudio se estructuró en torno a **cinco casos de análisis de complejidad progresiva**, en los que se integran datos de vuelos, emisiones y turismo de forma gradual. Cada caso ha sido documentado de forma independiente, permitiendo comparar escenarios y validar los resultados obtenidos desde una perspectiva multidimensional.

Esta metodología combina así elementos de extracción automatizada, ingeniería de datos, análisis cuantitativo, modelado computacional y evaluación crítica, con el objetivo de desarrollar una visión integral y técnicamente rigurosa sobre la eficiencia y sostenibilidad de las rutas aéreas comerciales.

1.5. Estructura del documento

El presente trabajo se estructura en diez capítulos principales, además de una sección de bibliografía y varios anexos. La organización sigue una progresión lógica, desde la contextualización y motivación del estudio hasta el desarrollo técnico, el análisis de resultados y las conclusiones finales.

El **Capítulo 1** presenta la introducción al trabajo, incluyendo el contexto del transporte aéreo comercial, la motivación y relevancia del estudio, los objetivos planteados, la metodología general adoptada y una visión global de la estructura del documento.

En el **Capítulo 2** se expone el marco teórico y el estado del arte, abordando los conceptos clave relacionados con la ciencia de datos aplicada al transporte aéreo, las métricas de eficiencia operativa, los modelos de optimización mediante grafos, la sostenibilidad ambiental en la aviación y la relación entre turismo y conectividad aérea.

El **Capítulo 3** describe el proceso de recopilación e integración de datos, detallando las fuentes utilizadas, los formatos disponibles, las técnicas de extracción (*scraping* y APIs) y las principales variables seleccionadas para el análisis.

En el **Capítulo 4** se aborda el preprocesamiento de los datos, incluyendo la limpieza, normalización y consolidación de información, así como la conversión temporal y geoespacial de los conjuntos de datos y su organización por casos de estudio.

El **Capítulo 5** se dedica al análisis exploratorio general (EDA), donde se presentan estadísticas descriptivas, visualizaciones geográficas de rutas y aeronaves, e identificación de *outliers* y comportamientos anómalos.

El **Capítulo 6** recoge cinco casos de estudio diferenciados, cada uno con un nivel creciente de integración de variables: desde el análisis de vuelos comerciales, hasta combinaciones que incorporan datos medioambientales (emisiones de CO₂) y turísticos.

En el **Capítulo 7** se aplican algoritmos de *clustering* para la detección de patrones, justificando su uso, explicando su configuración y analizando los resultados obtenidos a partir de agrupaciones significativas.

El **Capítulo 8** desarrolla una discusión crítica sobre los resultados, evaluando su validez, señalando las limitaciones del enfoque adoptado y reflexionando sobre su relevancia práctica y potencial de aplicación real.

El **Capítulo 9** cierra el documento con las conclusiones principales, la síntesis de las aportaciones realizadas, posibles mejoras metodológicas y futuras líneas de investigación o desarrollo.

A continuación se incluye la **bibliografía** consultada y, finalmente, se presentan los **anexos**, que recogen material complementario relevante, como el **código fuente** (Anexo A) y un **glosario de términos técnicos y siglas** (Anexo B).

Capítulo 2

Marco teórico y estado del arte

2.1. Ciencia de datos en el transporte aéreo

La aplicación de ciencia de datos al sector del transporte aéreo representa una de las tendencias más relevantes en la transformación digital de la industria aeronáutica. En los últimos años, la creciente disponibilidad de datos en tiempo real, la madurez de las infraestructuras de procesamiento distribuido y el avance en técnicas analíticas han posibilitado el desarrollo de soluciones basadas en evidencia para mejorar la eficiencia, seguridad y sostenibilidad de las operaciones aéreas.

Desde una perspectiva técnica, la ciencia de datos engloba un conjunto de metodologías orientadas al tratamiento, análisis y visualización de grandes volúmenes de datos heterogéneos. En el caso del transporte aéreo, esto incluye información sobre posiciones de aeronaves, planes de vuelo, tipos de aviones, condiciones meteorológicas, emisiones estimadas, frecuencias de rutas o indicadores turísticos, entre otros. La integración de estos datos permite generar modelos descriptivos, predictivos y prescriptivos con capacidad para analizar trayectorias reales, estimar métricas de rendimiento y proponer rutas alternativas optimizadas.

Los avances en sistemas de vigilancia como el ADS-B (*Automatic Dependent Surveillance-Broadcast*) han permitido la recopilación de datos de vuelo de alta precisión, lo que ha potenciado el desarrollo de plataformas abiertas como OpenSky Network, que ofrecen acceso a registros históricos y en tiempo real sobre miles de vuelos comerciales. Estas fuentes, combinadas con repositorios complementarios como OpenFlights y datos institucionales de carácter turístico o medioambiental, constituyen una base sólida para la aplicación de modelos de análisis geoespacial, aprendizaje automático y simulación computacional.

En el ámbito académico e industrial, se han desarrollado numerosos estudios en los que se emplea ciencia de datos para abordar problemas clave de la aviación: predicción de demanda, detección de anomalías en rutas, análisis de eficiencia del espacio aéreo, simulación de congestión aeroportuaria o estimación de emisiones en trayectos específicos. Asimismo, existe un creciente interés en combinar estos análisis con perspectivas de sostenibilidad y economía circular, especialmente en el contexto de la aviación verde y la descarbonización del sector.

El presente trabajo se posiciona en esta línea de investigación aplicada, con el objetivo de aprovechar el potencial de los datos abiertos y las técnicas analíticas para evaluar y optimizar rutas aéreas comerciales. A través del uso de herramientas como **Python** [22], **Jupyter** [13], **Pandas** [14], **NetworkX** [8] y **scikit-learn** [20], se busca construir un marco de análisis reproducible, escalable y útil para la toma de decisiones en entornos reales.

2.2. Métricas de eficiencia operativa

La eficiencia operativa en el transporte aéreo puede definirse como la relación entre los recursos consumidos en una operación de vuelo y los resultados obtenidos, en términos de rendimiento, cobertura, coste o impacto. Su análisis cuantitativo resulta fundamental tanto para aerolíneas como para gestores de espacio aéreo y responsables de sostenibilidad, ya que permite identificar trayectorias subóptimas, evaluar el desempeño de aeronaves y estimar márgenes de mejora en la planificación de rutas.

En el presente trabajo se consideran varias métricas clave de eficiencia operativa, calculadas a partir de datos abiertos y aplicables a trayectorias reales de vuelo. A continuación, se describen las más relevantes, incluyendo su formulación y justificación.

Distancia efectiva vs. distancia óptima

Una de las formas más básicas de evaluar la eficiencia de una ruta aérea es comparar la distancia realmente volada con la distancia geodésica mínima entre origen y destino. Esta métrica permite detectar desvíos estructurales, zigzags o rodeos innecesarios debidos a restricciones del espacio aéreo, meteorología o decisiones operativas.

Sea d_{real} la distancia total recorrida por una aeronave según su trayectoria registrada (por ejemplo, vía ADS-B), y d_{ideal} la distancia ortodrómica (la más corta sobre la superficie esférica terrestre). Se define entonces la eficiencia de trayectoria η_d como:

$$\eta_d = \frac{d_{\text{ideal}}}{d_{\text{real}}} \in (0, 1] \quad (2.1)$$

Valores cercanos a 1 indican rutas directas y eficientes; valores más bajos sugieren trayectorias desviadas.

Velocidad media efectiva

La velocidad media efectiva permite analizar el aprovechamiento del tiempo en vuelo:

$$v_{\text{media}} = \frac{d_{\text{real}}}{t_{\text{vuelo}}} \quad (2.2)$$

donde t_{vuelo} es el tiempo transcurrido entre el despegue y el aterrizaje. Comparar esta métrica entre rutas similares permite identificar congestiones, esperas o ineficiencias operativas.

Carga útil estimada (*Load Factor*)

El factor de ocupación o *load factor* estima el grado de utilización de la capacidad de una aeronave:

$$LF = \frac{P_{\text{estimados}}}{P_{\text{máx}}} \quad (2.3)$$

donde $P_{\text{estimados}}$ es el número estimado de pasajeros a bordo, y $P_{\text{máx}}$ la capacidad máxima del avión. Aunque no siempre se dispone de ocupación real, puede aproximarse con promedios según ruta o tipo de aeronave.

Emisiones por pasajero-kilómetro ($\text{CO}_2/\text{pax}\cdot\text{km}$)

Desde una perspectiva ambiental, es fundamental estimar las emisiones normalizadas por pasajero y distancia:

$$E_{\text{unitaria}} = \frac{E_{\text{total}}}{P_{\text{estimados}} \cdot d_{\text{real}}} \quad (2.4)$$

donde E_{total} son las emisiones estimadas (kg CO_2), $P_{\text{estimados}}$ los pasajeros y d_{real} la distancia recorrida (km). El resultado se expresa en kg $\text{CO}_2/\text{pax}\cdot\text{km}$, útil para comparar la eficiencia ecológica de diferentes rutas.

Índice compuesto de eficiencia

Para una comparación multidimensional, se puede definir un índice agregado:

$$I_e = w_1 \cdot \eta_d + w_2 \cdot \frac{v_{\text{media}}}{v_{\text{ref}}} + w_3 \cdot (1 - E_{\text{unitaria}}^{\text{norm}}) \quad (2.5)$$

donde los pesos w_i pueden ajustarse según el énfasis del análisis (operativo, ambiental, mixto). Este índice compuesto se utiliza en fases posteriores del trabajo para comparar rutas alternativas con base técnica y objetiva.

Métricas aplicadas a simulaciones y turismo

- Comparación entre rutas simuladas y reales

En el contexto de este trabajo, se han implementado simulaciones de rutas aéreas optimizadas basadas en modelos de grafos. Para evaluar el rendimiento relativo frente a las trayectorias reales observadas en los datos ADS-B, se define una métrica comparativa simple pero informativa:

$$\Delta_{\text{distancia}} = d_{\text{real}} - d_{\text{simulada}} \quad (2.6)$$

donde:

- d_{real} : distancia total del vuelo observado.
- d_{simulada} : distancia de la ruta propuesta por el modelo.

Este valor puede analizarse en términos absolutos (km evitables) o relativos. Se define entonces una métrica de mejora relativa:

$$\eta_{\text{mejora}} = \frac{d_{\text{real}} - d_{\text{simulada}}}{d_{\text{real}}} \quad (2.7)$$

Valores positivos indican que la ruta simulada es más eficiente que la real. Esta métrica se ha empleado para cuantificar el beneficio potencial de aplicar optimización computacional al diseño de rutas comerciales.

- Densidad turística relativa por vuelo

Para abordar la dimensión turística de la eficiencia, se ha definido una métrica que relaciona el volumen turístico estimado con la oferta de conectividad aérea en una ruta o país. Esta métrica busca captar la relación entre presión turística y número de vuelos, útil para estudios de sostenibilidad:

$$D_{\text{turística}} = \frac{T_{\text{anual}}}{N_{\text{vuelos}}} \quad (2.8)$$

donde:

- T_{anual} : número de turistas internacionales anuales recibidos (fuente: Banco Mundial).
- N_{vuelos} : total de vuelos anuales en rutas hacia ese destino (según OpenSky/OpenFlights).

Este ratio permite identificar países o regiones que concentran un alto número de turistas con una baja conectividad aérea (lo cual puede reflejar saturación o dependencia excesiva), o lo contrario. También es útil para detectar desequilibrios entre planificación de rutas y demanda turística real.

Estas métricas reflejan el carácter multidimensional e interdisciplinar del proyecto, y permiten integrar de forma efectiva aspectos técnicos, medioambientales y socioeconómicos en un marco de análisis único.

2.3. Rutas óptimas y algoritmos de grafos

El modelado de trayectorias como grafos constituye una herramienta fundamental en la representación abstracta de redes de transporte, incluyendo aquellas asociadas al tráfico aéreo comercial. Un grafo dirigido permite modelar aeropuertos como nodos y las rutas entre ellos como aristas con una orientación y un peso asociado, que puede representar distancia, coste operativo, emisiones, tiempo de vuelo o una combinación de estos.

Definición formal Un grafo dirigido $G = (V, E)$ se define como un conjunto de nodos V (aeropuertos o puntos de control) y un conjunto de aristas dirigidas $E \subseteq V \times V$, donde cada arista $(u, v) \in E$ tiene un peso asociado $w(u, v) \in R^+$.

En el contexto de este trabajo, los pesos $w(u, v)$ pueden representar:

- La distancia geodésica entre u y v .
- El tiempo estimado de vuelo.
- Las emisiones de CO₂ asociadas a ese tramo.
- Una combinación ponderada de varios factores.

Algoritmos de búsqueda de rutas óptimas Para encontrar rutas eficientes en un grafo dirigido, se emplean los siguientes algoritmos:

- **Dijkstra:** encuentra la ruta de coste mínimo desde un nodo origen a todos los demás, siempre que los pesos sean no negativos. Complejidad: $\mathcal{O}(|E| + |V| \log |V|)$ con montículos de Fibonacci.
- **A* (A estrella):** extensión heurística de Dijkstra que incorpora una función de coste estimado $h(v)$ hacia el destino. Ideal para contextos geoespaciales, usando la distancia geodésica como heurística admisible.
- **Bellman-Ford:** permite pesos negativos (no aplicable en este trabajo), pero se menciona por completitud teórica.
- **Floyd-Warshall:** computa todas las distancias mínimas entre pares de nodos. Complejidad: $\mathcal{O}(|V|^3)$. Útil para análisis globales, aunque poco eficiente para redes extensas.

Aplicación en este trabajo En este Trabajo se construye un grafo dirigido G_{vuelos} a partir de datos de OpenFlights y OpenSky Network. Cada nodo representa un aeropuerto y cada arista viable un tramo entre dos puntos. El peso de cada arista se define como:

$$w(u, v) = \alpha \cdot d(u, v) + \beta \cdot e(u, v) \quad (2.9)$$

donde:

- $d(u, v)$: distancia geodésica entre los aeropuertos u y v ,
- $e(u, v)$: emisiones estimadas de CO₂,
- $\alpha, \beta \in R^+$: pesos ajustables según si se prioriza eficiencia operativa o sostenibilidad ambiental.

Este modelo permite generar rutas optimizadas bajo distintos objetivos (mínima distancia, mínima emisión, o equilibrio entre ambas).

Ruta óptima y eficiencia relativa Para evaluar la calidad de una ruta real R_r frente a una ruta simulada óptima R_o , se define una medida de eficiencia relativa:

$$\eta_{\text{ruta}} = \frac{w(R_o)}{w(R_r)} \in (0, 1] \quad (2.10)$$

donde $w(R)$ representa el coste total acumulado de la ruta R , según la función objetivo definida.

Valores cercanos a 1 indican rutas reales bien optimizadas, mientras que valores bajos sugieren potencial de mejora mediante optimización computacional.

Esta formalización permite aplicar los algoritmos en escenarios reales extraídos del dataset, y comparar el rendimiento de las rutas comerciales actuales con trayectorias calculadas desde una perspectiva de eficiencia computacional y sostenibilidad ambiental.

Para ilustrar la naturaleza geográfica de estos datos, la siguiente figura presenta una visualización de varias rutas reales contenidas en el conjunto de vuelos analizado. Esta representación permite observar la disposición espacial de los aeropuertos implicados y la estructura de conectividad aérea global utilizada como base para los modelos desarrollados en este trabajo.



Figura 2.1: Mapa geográfico de rutas aéreas reales consideradas en el análisis. Cada marcador representa un aeropuerto con conexiones activas en los conjuntos de datos recopilados. Las líneas verdes indican vuelos comerciales registrados, conectando regiones de Europa, América, Asia y Oceanía, y representan además la distancia en kilómetros. Esta visualización fue generada mediante herramientas de geolocalización y procesamiento de datos abiertos. Elaboración propia.

2.4. Emisiones y sostenibilidad en la aviación

El sector de la aviación comercial representa una fuente significativa de emisiones de gases de efecto invernadero (GEI), en particular de dióxido de carbono (CO_2), debido al uso intensivo de combustibles fósiles como el queroseno. Aunque su participación global en las emisiones totales oscila entre el 2 % y el 3 %, su impacto relativo es superior debido a las emisiones en altitud, que presentan efectos climáticos amplificados por fenómenos como el forzamiento radiativo y la formación de estelas [4].

Según la Organización de Aviación Civil Internacional (ICAO), en 2019 el transporte aéreo emitió aproximadamente 915 millones de toneladas de CO_2 , con una proyección creciente en escenarios post-pandemia. Para hacer frente a este reto, se han implementado estrategias como CORSIA (Carbon Offsetting and Reduction Scheme for International Aviation), centradas en la neutralización de emisiones y en la mejora de la eficiencia energética operativa [12].

Factores que afectan las emisiones aéreas Las emisiones generadas por un vuelo no dependen únicamente de la distancia recorrida, sino también de una serie de variables técnicas y operativas:

- Distancia del vuelo: las etapas de despegue y ascenso tienen mayor intensidad por kilómetro.
- Modelo de aeronave: cada diseño presenta distintas prestaciones aerodinámicas, eficiencia de motores y capacidad de carga.

- Condiciones operativas: peso al despegue, altitud de crucero, congestión del espacio aéreo o condiciones meteorológicas.
- Ocupación del vuelo: afecta directamente a la eficiencia ambiental por pasajero transportado.

Modelos modernos como el Airbus A350 o el Boeing 787 presentan consumos significativamente más bajos por asiento-kilómetro en comparación con generaciones anteriores como el Boeing 767 o el Airbus A340. Este factor se ha incorporado al análisis mediante el cruce entre rutas y tipos de aeronave identificados a través de scraping.

Estimación de emisiones por tramo Se ha adoptado una estrategia mixta para estimar las emisiones de CO₂, en función de la disponibilidad de datos. Si se conoce el modelo de aeronave y la distancia, se estima el consumo de combustible C_{fuel} y se aplica el factor de conversión del IPCC:

$$E_{\text{CO}_2} = C_{\text{fuel}} \cdot EF_{\text{CO}_2}, \quad \text{donde } EF_{\text{CO}_2} \approx 3.15 \text{ kgCO}_2/\text{kgdecombustible} \quad (2.11)$$

Cuando no se dispone del consumo directo, se recurre a una estimación simplificada basada en la distancia:

$$E_{\text{CO}_2} \approx D \cdot EF_{\text{km}}(\text{aeronave}) \quad (2.12)$$

donde EF_{km} es un factor de emisión específico por tipo de aeronave, extraído de fuentes como el ICAO Carbon Emissions Calculator, Eurocontrol o documentación de fabricantes.

Emisiones normalizadas por pasajero Una de las métricas clave aplicadas es la emisión por pasajero y kilómetro:

$$EU_{\text{CO}_2} = \frac{E_{\text{CO}_2}}{P_{\text{estimados}} \cdot D} \quad (2.13)$$

donde $P_{\text{estimados}}$ es el número estimado de pasajeros. Esta métrica refleja la eficiencia medioambiental relativa de una ruta y ha sido usada para clasificar rutas según su sostenibilidad e integrarla en análisis de agrupación (clustering).

Aplicación en el trabajo Este análisis se ha aplicado a rutas reales identificadas mediante scraping, y vinculadas a tipos de aeronave mediante bases como OpenFlights. Las emisiones estimadas han sido empleadas tanto en el análisis exploratorio como en la comparación entre trayectorias reales y simuladas, aportando un criterio objetivo para seleccionar rutas más sostenibles.

Además, se ha construido una métrica sintética que combina distancia, ocupación estimada y tipo de aeronave, con el fin de priorizar trayectorias con mayor rendimiento ambiental. Esta herramienta de evaluación se utilizará en capítulos posteriores para visualizar rutas críticas y proponer alternativas mediante técnicas de optimización sobre grafos.

2.5. Interrelación turismo–movilidad aérea

La aviación comercial desempeña un papel fundamental en la configuración y dinámica del turismo internacional. En un mundo globalizado, la conectividad aérea determina en gran medida la accesibilidad de los destinos, influye en la competitividad del sector turístico y condiciona las decisiones de los viajeros en términos de coste, tiempo y comodidad.

Según la Organización Mundial del Turismo (OMT), más del 50 % de las llegadas internacionales se realizan por vía aérea, porcentaje que asciende a más del 80 % en el caso de destinos insulares o intercontinentales [19]. Este vínculo estructural convierte a la aviación en un facilitador directo del crecimiento turístico, pero también en un vector de presión sobre los destinos en términos de capacidad, sostenibilidad y calidad de vida.

Coneectividad aérea como motor turístico La existencia de rutas aéreas frecuentes, directas y asequibles es un factor determinante para el atractivo de un destino turístico. En este sentido, el número de vuelos, la cantidad de asientos disponibles, la variedad de aerolíneas y la estacionalidad son variables que impactan directamente en el volumen de turistas recibido.

Este fenómeno se puede analizar cuantitativamente mediante indicadores como:

- **Índice de conectividad aérea** (Air Connectivity Index), desarrollado por IATA u OACI.
- **Llegadas por vuelo:** cociente entre turistas recibidos y número de vuelos registrados.
- **Tasa de dependencia aérea:** porcentaje de turistas internacionales respecto al total de visitantes.

Una alta dependencia del transporte aéreo sin una gestión planificada puede derivar en fenómenos de *overtourism*, colapso de infraestructuras locales y deterioro de recursos naturales.

Dimensión ambiental del turismo aéreo El turismo, al depender crecientemente del transporte aéreo, hereda también parte de su huella ambiental. Esto implica que destinos muy atractivos desde el punto de vista económico pueden tener baja sostenibilidad relativa, especialmente si concentran rutas largas, baja ocupación y limitada capacidad de absorción ecológica.

Para evaluar esta relación, en este trabajo se propone el análisis cruzado de:

- Vuelos reales por destino (datos de OpenSky y OpenFlights).
- Emisiones estimadas por ruta y tipo de aeronave.
- Indicadores turísticos del Banco Mundial:
 - Llegadas de turistas internacionales.
 - Gasto turístico por visitante.
 - Porcentaje del PIB atribuido al turismo.

Este cruce permite construir perfiles de *turismo aéreo sostenible*, combinando intensidad turística, conectividad y eficiencia ambiental.

Aplicación en este trabajo La interrelación entre turismo y aviación se analiza en profundidad en los Estudios 4 y 5 del Capítulo 6, mediante métricas como la *densidad turística por vuelo* o la *emisión media por turista recibido*. Estos indicadores permiten identificar rutas o países con desequilibrios entre tráfico aéreo, capacidad de acogida y rendimiento ambiental, lo que aporta valor tanto en la toma de decisiones empresariales como en el diseño de políticas públicas de turismo sostenible.

Además, este trabajo propone un enfoque cuantitativo replicable que podría extenderse a otros sectores de movilidad turística, como cruceros, ferrocarril o rutas terrestres de media distancia.

Capítulo 3

Recopilación e integración de datos

3.1. Fuentes utilizadas

La recopilación de datos constituye una fase crítica en el desarrollo de este proyecto, al ser la base sobre la que se construye todo el análisis posterior. Dada la naturaleza interdisciplinar del estudio —que combina eficiencia operativa, sostenibilidad ambiental e indicadores turísticos— se ha optado por integrar datos de múltiples fuentes abiertas, priorizando aquellas con disponibilidad pública, continuidad histórica, granularidad suficiente y fiabilidad contrastada.

A continuación, se detallan las principales fuentes utilizadas:

OpenSky Network opensky-network.org

OpenSky Network es una plataforma colaborativa que proporciona acceso a datos aeronáuticos en tiempo real y en diferido, recopilados a través de receptores ADS-B distribuidos por todo el mundo. Este sistema permite obtener:

- Identificadores de vuelo.
- Coordenadas geográficas (latitud, longitud).
- Altitud, rumbo, velocidad.
- Tiempos de despegue y aterrizaje estimados.

Estos datos han sido esenciales para reconstruir trayectorias reales de vuelos comerciales y analizar desviaciones respecto a rutas óptimas. La extracción se ha realizado mediante técnicas de *scraping* adaptadas a su portal de vuelos históricos, respetando los términos de uso y sin sobrecargar el sistema.

OpenFlights openflights.org/data.html

OpenFlights ofrece una base de datos estructurada con información estática sobre:

- Aeropuertos (códigos IATA/ICAO, ubicación, altitud).
- Aerolíneas.
- Rutas existentes entre aeropuertos.

- Tipos de aeronave vinculados a ciertas rutas.

Este dataset ha sido utilizado como soporte para construir la red de conectividad aérea global, sobre la cual se aplican posteriormente algoritmos de grafos para simular rutas alternativas y medir métricas estructurales como grado, densidad o accesibilidad.

Banco Mundial (World Bank Open Data) data.worldbank.org

El Banco Mundial ofrece indicadores macroeconómicos y turísticos por país, que se han incorporado en los análisis de sostenibilidad turística del proyecto. Los indicadores seleccionados incluyen:

- Llegadas de turistas internacionales (número absoluto).
- Gasto turístico por visitante.
- Ingresos por turismo (como porcentaje del PIB).
- Capacidad aeroportuaria estimada (cuando está disponible).

Estos datos han permitido construir relaciones cuantitativas entre el tráfico aéreo y el impacto turístico, dentro del análisis cruzado desarrollado en los Estudios 4 y 5.

Fuentes medioambientales y de emisiones Se han utilizado varias referencias oficiales y técnicas para estimar las emisiones de CO₂:

- EEA (European Environment Agency): factores medios de emisión por tipo de trayecto.
- IPCC Guidelines (2006, Vol. 2): coeficientes de conversión de combustible a CO₂.
- UK Government GHG Conversion Factors: métricas prácticas de emisiones por pasajero-kilómetro.

Cuando ha sido posible, las emisiones se han ajustado por tipo de aeronave utilizando datos complementarios del *ICAO Carbon Emissions Calculator* y de documentación técnica de fabricantes como Airbus y Boeing.

3.2. Tipos de datos y formatos

La recopilación e integración de datos en este proyecto ha requerido el tratamiento de múltiples fuentes con estructuras, formatos y niveles de calidad heterogéneos. Este aspecto no es menor, ya que la calidad y consistencia del análisis posterior dependen en gran medida de la fidelidad y compatibilidad de los datos de entrada.

A lo largo del trabajo, se ha gestionado una combinación de datos estructurados, semi-estructurados y no estructurados, cada uno con sus propias particularidades. Esto ha exigido un trabajo previo de ingeniería de datos: conversión de formatos, normalización de variables, armonización temporal y resolución de ambigüedades.

Clasificación por formato y naturaleza Los formatos tratados pueden agruparse en las siguientes categorías:

- **JSON estructurado con autenticación:** datos de OpenSky Network vía API RESTful protegida con OAuth2. Incluye posición, altitud, velocidad, estado y otros atributos dinámicos. Se requiere deserialización y transformación tabular.
- **HTML con *scraping* dinámico (Selenium):** sitios como FlightAware generan sus interfaces mediante JavaScript. Se ha usado Selenium en modo *headless* para extraer tablas de vuelos por tipo de aeronave, navegando mediante paginación.
- **CSV y Excel estructurados:** datos de OpenFlights y el Banco Mundial vienen en formatos estándar, tratados con `pandas`. Incluyen aeropuertos, rutas, códigos de aeronaves y estadísticas por país.
- **Tablas técnicas institucionales:** factores de emisión y coeficientes energéticos (EEA, ICAO, IPCC), convertidos manualmente o codificados como constantes en scripts de cálculo.

Reto técnico: diversidad estructural y semántica Cada fuente presenta una estructura propia tanto en columnas como en unidades, codificaciones y campos opcionales. Algunos ejemplos de retos tratados:

- En OpenSky, los tiempos vienen en formato `UNIX timestamp`; las altitudes pueden ser barométricas o geodésicas y deben validarse.
- En FlightAware, los horarios se presentan como texto mixto (e.g., “Mié 01:05PM GMT+1”), lo que requiere separación y conversión explícita a UTC.
- En OpenFlights, los códigos de aeropuertos pueden alternar entre IATA e ICAO. Se han creado funciones de mapeo para garantizar consistencia.
- Se han diseñado funciones de *data enrichment* para combinar datos de distintas fuentes (posición de vuelo de OpenSky + tipo de avión de FlightAware + eficiencia ambiental de ICAO).

Tipos de variables tratados Durante el análisis se han manipulado distintas tipologías de datos:

- **Categóricas:** tipo de aeronave, aerolínea, país, código de vuelo.
- **Numéricas continuas:** velocidad, altitud, emisiones, duración del vuelo.
- **Datos espaciales:** latitud, longitud (proyectables sobre mapas).
- **Datos temporales:** fechas con zona horaria, formatos mixtos.
- **Variables calculadas:** eficiencia de ruta, CO₂/pax·km, índices de sostenibilidad.

Todo el procesamiento se ha realizado con Python en entornos `Jupyter Notebook`, empleando librerías como `pandas`, `numpy`, `datetime`, `pytz`, `openpyxl`, `selenium` y `requests`.

Ejemplo de transformación Un caso representativo ha sido la reconstrucción del campo **Duración del vuelo**. El proceso consistió en:

1. Separar fecha, hora y zona horaria.
2. Convertir a tipo `datetime` con `pytz`.
3. Calcular la duración en minutos reales.
4. Corregir valores negativos (errores o vuelos cancelados).
5. Convertir a formato hh:mm legible.

Este tipo de transformaciones es ilustrativo del enfoque técnico adoptado: transformar datos crudos en variables analíticas robustas, fiables y comparables.

Resumen comparativo de las fuentes y formatos Para facilitar la consulta, se presenta a continuación una tabla resumen de las principales fuentes de datos, sus formatos de acceso, el tipo de información extraída y su uso específico en el proyecto.

Cuadro 3.1: Resumen de fuentes, formatos y aplicación de datos en el proyecto.

Fuente	Formato	Acceso	Contenido	Uso en el TFG
OpenSky Network	JSON estructurado	API con OAuth2	Datos de vuelos en tiempo real: posición, altitud, velocidad, timestamps	Reconstrucción de trayectorias reales y análisis de eficiencia
OpenFlights	CSV / Web scraping	Descarga directa / Selenium	Aeropuertos, rutas, aerolíneas, modelos de aeronave	Construcción de red aérea; definición de nodos y aristas
FlightAware	HTML dinámico	Scraping automatizado (Selenium)	Vuelos por tipo de aeronave, horarios, duración	Cálculo de duración y ocupación estimada por tipo de avión
Banco Mundial	CSV / JSON	Portal oficial	Llegadas, gasto por turista, PIB turístico	Evaluación de sostenibilidad turística y tráfico aéreo
EEA / ICAO / IPCC	Tablas técnicas	Documentación oficial	Factores de emisión y conversión de combustible	Estimación de emisiones por tramo, pasajero y modelo de aeronave

Fuente: Elaboración propia a partir del sistema de scraping y fuentes abiertas.

3.3. Técnicas de scraping y APIs

Dado que muchas de las fuentes de datos relevantes para este trabajo no ofrecen interfaces unificadas de descarga estructurada, ha sido necesario desarrollar distintos procedimientos de adquisición de información basados en técnicas de *scraping* web y consumo de APIs REST, en función del tipo de recurso y sus condiciones de acceso.

Estas técnicas han sido aplicadas con una orientación pragmática, siguiendo buenas prácticas en cuanto a tiempos de espera, volumen de peticiones y limpieza del contenido extraído, garantizando tanto el respeto a las condiciones de uso de las plataformas como la reproducibilidad de los resultados.

Consumo de la API de OpenSky Network OpenSky Network ofrece una API pública que permite la consulta de datos aeronáuticos en tiempo real y diferido, mediante protocolo HTTP y con respuesta en formato JSON. El acceso requiere autenticación mediante credenciales de cliente, bajo el flujo estándar OAuth 2.0 (`client_credentials`).

Una vez obtenido el token de acceso, se realiza una consulta al *endpoint* de estado de vuelos, que devuelve un conjunto de observaciones por aeronave. Los campos extraídos incluyen:

- `icao24`, `callsign`, `origin_country`.
- Coordenadas: `latitude`, `longitude`.
- Altitudes: `baro_altitude`, `geo_altitude`.
- Velocidad (`velocity`) y rumbo (`true_track`).
- Marcas temporales: `time_position`, `last_contact`.

El resultado se estructura en un `DataFrame` de Pandas con columnas renombradas y formatos temporales convertidos de UNIX a `datetime`. Se filtran valores faltantes e inconsistentes antes de integrarlo en cálculos de eficiencia, duración y emisiones.

Scraping con Selenium de FlightAware FlightAware proporciona información detallada sobre vuelos pasados y activos. Dado que no dispone de una API pública gratuita y que el contenido se genera dinámicamente en el navegador, se ha utilizado `Selenium WebDriver` para automatizar la interacción.

El proceso se ha dividido en:

1. Scraping por tipo de aeronave:

- Se parte de una lista de más de 200 códigos ICAO.
- Para cada modelo, se accede a su página y se extraen tablas de vuelos activos.
- Se gestiona la paginación mediante botones como “*Next 40*”.
- Campos extraídos: Identificador, Origen, Destino, Salida, Llegada estimada.

2. Scraping por aerolínea:

- Lista de más de 200 aerolíneas (IATA).
- Extracción de vuelos actuales por operador.
- Gestión de offsets para paginación.
- Campos: los mismos que en el scraping por modelo.

Los datos extraídos se almacenan en ficheros `.xlsx` para su inspección manual y posterior integración.

Tratamiento avanzado de tiempos y zonas horarias Los datos horarios de FlightAware presentan formatos como Wed 08:35PM CDT. El pipeline de normalización implementado incluye:

1. Separación de fecha, hora y zona horaria.
2. Conversión a `datetime` con `pytz`.
3. Cálculo de duración en minutos.
4. Corrección de valores negativos (vuelos cancelados o errores).
5. Formateo a hh:mm legible para visualización.

Este proceso se aplica tanto a vuelos por tipo de aeronave como por aerolínea, y genera un conjunto estandarizado y utilizable para análisis comparativos.

Extracción de indicadores turísticos

Los datos turísticos utilizados en este trabajo proceden de los registros públicos del Banco Mundial y aportan una dimensión económica y social al análisis técnico. La información se descargó en formato CSV estructurado, con identificadores estándar por país (nombre y código ISO), lo que facilitó su integración con los conjuntos de datos aeronáuticos.

Campos utilizados Las principales variables seleccionadas fueron:

- **Llegadas de turistas internacionales** (`llegadas_turistas`): indicador absoluto de presión turística desde el exterior.
- **Ingresos por turismo en USD** (`ingresos_turisticos_usd`): métrica del impacto económico directo del turismo en la balanza de pagos.
- **Gasto turístico como % del PIB** (`gasto_turismo_rel_pib`): mide la dependencia relativa de la economía nacional respecto al turismo.
- **PIB total en USD** (`pib_total_usd`): permite derivar indicadores como gasto turístico per cápita o intensidad relativa.

Integración con datos aeronáuticos Estas variables se unieron a los datos de vuelos mediante el código de país (ISO-3166), generando indicadores cruzados de sostenibilidad turística aérea. Entre ellos destacan:

- **Emisiones por turista recibido**: combinación de emisiones agregadas por país y número de llegadas.
- **Número de vuelos por cada millón de turistas**: permite detectar destinos con posible saturación relativa.
- **Gasto medio por pasajero transportado**: como proxy del valor económico generado por viaje.

Los datos tienen estructura tabular clara (país, año, valor) y se integran mediante `merge` en Pandas.

Factores de emisiones: codificados manualmente desde tablas de:

- EEA (European Environment Agency),
- ICAO (International Civil Aviation Organization),
- IPCC (Intergovernmental Panel on Climate Change).

Estos valores se aplican directamente en el modelo de emisiones descrito en la Sección de Emisiones.

Síntesis del flujo de adquisición Todo el proceso puede visualizarse como un flujo secuencial:

- Acceso automatizado (API o scraping).
- Transformación estructural (parseo, normalización, tipado).
- Integración semántica (cruce por identificadores comunes).
- Exportación final en formatos tabulares para análisis y modelado.

Este pipeline ha sido implementado íntegramente en Python, es reproducible y escalable ante nuevas rondas de datos o ampliaciones del estudio.

3.4. Descripción de variables clave

El conjunto de datos final utilizado en este trabajo integra información procedente de múltiples fuentes (OpenSky, OpenFlights, FlightAware, Banco Mundial, etc.), unificada y preprocesada para permitir un análisis cruzado coherente. En este contexto, resulta esencial identificar y describir las variables clave que conforman la base analítica del proyecto.

A continuación, se presenta una descripción detallada de las principales variables, agrupadas por dimensión de análisis: aeronáutica, geoespacial, temporal, medioambiental y turística.

Variables aeronáuticas

Estas variables describen las características técnicas básicas de cada vuelo o aeronave, y permiten vincular cada trayecto con información relevante sobre el avión, su operador y los aeropuertos implicados.

Cuadro 3.2: Variables aeronáuticas empleadas en el análisis.

Variable	Descripción
<code>icao24</code>	Identificador hexadecimal único de la aeronave según OpenSky Network.
<code>callsign</code>	Código de llamada del vuelo, utilizado en comunicaciones aeronáuticas.
<code>tipo_aeronave</code>	Modelo de aeronave (por ejemplo, B738, A320).
<code>aerolinea</code>	Código IATA de la compañía aérea operadora del vuelo.
<code>origen / destino</code>	Códigos ICAO/IATA de los aeropuertos de salida y llegada asociados al vuelo.

Variables geoespaciales

Son variables derivadas del seguimiento por radar ADS-B (*Automatic Dependent Surveillance–Broadcast*). Han sido utilizadas para reconstruir trayectorias reales, calcular distancias recorridas y visualizar mapas de conectividad aérea.

Cuadro 3.3: Variables geoespaciales extraídas desde OpenSky Network.

Variable	Descripción
<code>latitude,</code> <code>longitude</code>	Coordenadas geográficas (latitud y longitud) registradas durante el vuelo.
<code>baro_altitude</code>	Altitud barométrica de la aeronave, expresada en metros.
<code>geo_altitude</code>	Altitud geodésica o sobre el nivel medio del mar, también en metros.
<code>velocity</code>	Velocidad de desplazamiento del avión, en metros por segundo (m/s).
<code>true_track</code>	Rumbo del vuelo en grados respecto al norte geográfico.

Variables temporales

Estas variables capturan la dimensión cronológica de cada vuelo, incluyendo despegue, aterrizaje y duración. Han sido clave para validar secuencias, analizar eficiencia temporal y estimar la ocupación del espacio aéreo.

Cuadro 3.4: Variables temporales derivadas y procesadas para el análisis.

Variable	Descripción
<code>time_position,</code> <code>last_contact</code>	Timestamps en formato UNIX, generados por OpenSky Network durante eventos clave.
<code>salida_datetime</code>	Instante estandarizado en UTC del despegue, obtenido tras procesar texto con zona horaria.
<code>llegada_datetime</code>	Instante estandarizado en UTC de la llegada del vuelo.
<code>duracion_min</code>	Duración estimada del vuelo, en minutos, calculada como diferencia entre salida y llegada.
<code>duracion_hh:mm</code>	Representación legible de la duración del vuelo en formato horas:minutos.

Variables medioambientales

Estas variables han sido estimadas a partir de modelos técnicos y cruces con información sobre aeronaves. Se han empleado para evaluar el rendimiento ambiental de las rutas y clasificarlas según criterios de sostenibilidad.

Cuadro 3.5: Variables medioambientales estimadas para el análisis de sostenibilidad.

Variable	Descripción
distancia_real	Distancia efectiva recorrida por la aeronave, calculada a partir de coordenadas GPS.
distancia_ideal	Distancia ortodrómica (la más corta sobre la esfera terrestre) entre origen y destino.
eficiencia_trayectoria (η_d)	Ratio entre distancia ideal y real, como métrica de eficiencia operativa.
emisiones_CO2_total	Estimación total de CO ₂ emitido durante el vuelo.
emisiones_CO2_pax_km	Emisiones de CO ₂ normalizadas por pasajero-kilómetro.
tipo_modelo_emision	Tipo de aeronave asociado a un factor de emisión específico.

Variables turísticas y económicas

Estas variables fueron importadas del Banco Mundial y vinculadas por país de destino. Se han empleado para analizar la dependencia del turismo respecto al transporte aéreo y evaluar la presión ejercida sobre destinos con alta demanda.

Cuadro 3.6: Variables turísticas y económicas integradas en el análisis.

Variable	Descripción
llegadas_turistas	Número de llegadas de turistas internacionales por año.
ingresos_turisticos_usd	Total de ingresos obtenidos por turismo, expresados en dólares estadounidenses (USD).
gasto_turismo_rel_pib	Porcentaje del PIB nacional que corresponde al gasto turístico.
pib_total_usd	Producto interior bruto total del país, expresado en USD.

Variables compuestas y derivadas

Estas variables han sido construidas como resultado de combinaciones entre variables base. Son fundamentales para sintetizar resultados y apoyar la interpretación integral del análisis.

Cuadro 3.7: Variables compuestas y derivadas generadas en el análisis.

Variable	Descripción
indice_eficiencia_total	Índice agregado que pondera eficiencia temporal, energética y ambiental para cada vuelo.
indice_sostenibilidad_turistica	Relación entre emisiones de CO ₂ , volumen de tráfico aéreo y presión turística.
cluster_id	Identificador de grupo asignado por algoritmos de clustering no supervisado.
ruta_opt_simulada	Ruta optimizada generada por algoritmos de grafos, según criterios de eficiencia.
delta_distancia, delta_emisiones	Diferencias en distancia y emisiones entre rutas reales y rutas simuladas óptimas.

Estas variables, correctamente definidas y estructuradas, constituyen la base sobre la que se construyen los análisis posteriores, permitiendo una evaluación integral de la eficiencia, sostenibilidad y dinámica turística del transporte aéreo comercial.

Capítulo 4

Preprocesamiento de datos

4.1. Limpieza, normalización y consolidación

La calidad del análisis en proyectos de ciencia de datos depende en gran medida del estado inicial de los datos tratados. En este trabajo, el preprocesamiento ha sido una fase crítica, debido a la elevada heterogeneidad de fuentes, formatos, unidades, estructuras temporales y codificaciones. Por ello, se ha diseñado una serie de procedimientos sistemáticos para limpiar, normalizar y consolidar los datos obtenidos.

Detección y tratamiento de valores ausentes

La presencia de valores nulos, erróneos o no informados ha sido una constante, especialmente en los datos extraídos vía *scraping*. Se han aplicado distintas estrategias según el contexto:

- Eliminación de registros incompletos en campos críticos como `latitude`, `longitude`, `salida` o `llegada`.
- Imputación conservadora (por ejemplo, reemplazo de valores `NaN` por la media de grupo en la duración de vuelos).
- Conversión explícita de campos como `None`, `Unknown` o " " a valores `NaN` para un tratamiento uniforme.

Estandarización de unidades

Dado que los datos provienen de distintas fuentes y regiones, ha sido necesario:

- Unificar unidades de altitud (metros sobre el nivel del mar).
- Convertir velocidades a `m/s` o `km/h` según el tipo de análisis.
- Normalizar la duración de los vuelos a minutos exactos (`float`).
- Transformar distancias de millas náuticas o millas terrestres a kilómetros (1 NM = 1.852 km).

Homogeneización de identificadores

Para garantizar la compatibilidad entre fuentes:

- Se convirtieron todos los códigos IATA (3 letras) e ICAO (4 letras) a un único formato de referencia.
- Se crearon diccionarios de equivalencias para aerolíneas, países y tipos de aeronave.
- Se armonizaron campos como `Origen` y `Destino` para integrarse correctamente en la red de grafos.

Filtrado de registros inconsistentes

Se eliminaron registros que presentaban incoherencias lógicas:

- Vuelos con duración negativa o superior a 24 horas.
- Coordenadas fuera de rango (latitud fuera de $[-90, 90]$, longitud fuera de $[-180, 180]$).
- Registros sin `tipo_aeronave` o sin código `Ident` utilizable.

Consolidación por caso de estudio

Cada conjunto de datos fue tratado de forma modular y consolidado en función del caso de estudio correspondiente:

- Vuelos comerciales reales.
- Flota aérea y modelo de aeronave.
- Vuelos con emisiones estimadas.
- Vuelos cruzados con indicadores turísticos.
- Análisis combinado de turismo y emisiones.

Esta organización permitió mantener trazabilidad, reproducir el análisis por fases, y aplicar filtros o métricas específicas sin contaminar el resto del sistema de análisis.

4.2. Conversión temporal y geoespacial

La correcta interpretación de los datos espaciales y temporales es crucial en el análisis de rutas aéreas, ya que cualquier discrepancia en la codificación de tiempos o coordenadas puede conducir a errores significativos en la reconstrucción de trayectorias, en el cálculo de distancias, duraciones o incluso en la estimación de emisiones.

Normalización temporal: de texto ambiguo a formato UTC

Uno de los principales retos ha sido la homogeneización de las marcas temporales asociadas a los vuelos. En fuentes como FlightAware, los horarios de salida y llegada se proporcionan en formatos mixtos, como por ejemplo:

- Wed 08:45PM CEST
- Thu 09:15AM EDT

Para su tratamiento se diseñó un *pipeline* de conversión con los siguientes pasos:

1. Separación de fecha, hora y zona horaria en columnas distintas.
2. Mapeo de zonas horarias a identificadores estándar (`pytz`),
por ejemplo: CEST → Europe/Madrid,
EDT → America/New_York.
3. Conversión a objetos `datetime` localizados,
usando `datetime.strptime()` y
`pytz.timezone().localize()`.
4. Transformación final a UTC
mediante `.astimezone(pytz.UTC)`.

Esta estandarización ha permitido calcular con precisión la duración de los vuelos, detectar inconsistencias y garantizar la comparabilidad global entre trayectos.

Conversión de timestamps UNIX (OpenSky)

En el caso de OpenSky, las marcas temporales como `time_position` o `last_contact` se proporcionan como enteros UNIX (segundos desde epoch). Se convirtieron mediante:

```
pd.to_datetime(..., unit='s')
```

Estos datos, ya en UTC, han sido especialmente útiles para reconstruir trayectorias en tiempo real, minuto a minuto.

Tratamiento geoespacial: coordenadas y trayectorias

Las variables `latitude` y `longitude` han sido utilizadas para:

- Generar mapas de rutas mediante `matplotlib`, `geopandas` y `folium`.
- Calcular distancias entre puntos consecutivos usando la fórmula de Haversine.
- Evaluar desviaciones respecto a la ruta ideal (ortodrómica) entre origen y destino.

Se validó que todas las coordenadas se encontraban dentro de los rangos permitidos ($-90 \leq \phi \leq 90$, $-180 \leq \lambda \leq 180$), y se eliminaron registros fuera de estos valores.

Cálculo de distancias ortodrómicas

Para estimar la distancia teórica entre el aeropuerto de salida y el de llegada, se utilizó la fórmula de Haversine sobre una esfera de radio $R = 6371$ km:

$$d = 2R \cdot \arcsin \left(\sqrt{\sin^2 \left(\frac{\Delta\phi}{2} \right) + \cos(\phi_1) \cdot \cos(\phi_2) \cdot \sin^2 \left(\frac{\Delta\lambda}{2} \right)} \right) \quad (4.1)$$

donde:

- ϕ_1, ϕ_2 : latitudes del punto de origen y destino en radianes.
- $\Delta\phi, \Delta\lambda$: diferencias de latitud y longitud, respectivamente.
- $R = 6371$ km: radio medio de la Tierra.

Esta distancia ortodrómica se ha utilizado como referencia para calcular la eficiencia de cada trayecto real.

Resultado: estructura espacio-temporal fiable

Como resultado del procesamiento descrito, cada vuelo del dataset final incluye:

- Tiempos estandarizados en formato UTC.
- Duración estimada con precisión minuto a minuto.
- Coordenadas validadas y listas para visualización.
- Distancia real y distancia ortodrómica calculadas.
- Códigos de zona horaria correctamente mapeados y normalizados.

Este nivel de estructura ha sido imprescindible para garantizar la solidez de los análisis del Capítulo 5 y de los casos de estudio del Capítulo 6.

4.3. Integración de datasets por caso de estudio

Tras la limpieza, normalización y estandarización de los datos individuales, el siguiente paso fue la integración estratégica de los distintos datasets en función de los casos de estudio definidos. Este proceso no consistió en una simple concatenación de tablas, sino en una construcción modular cuidadosamente alineada con los objetivos analíticos de cada bloque temático del proyecto.

La integración se realizó mediante operaciones de `merge` condicionales, transformación de claves de unión (como códigos de país o modelos de aeronave), y definición explícita de jerarquías entre datasets (priorizando exactitud temporal, técnica o semántica según el caso).

Integración específica por estudio

El proyecto ha definido cinco casos de estudio, cada uno con su propio nivel de complejidad y tipo de cruces requeridos. La siguiente tabla resume la estrategia de integración adoptada para cada uno:

Cuadro 4.1: Resumen de integración de datasets por caso de estudio.

Estudio	Dataset base	Datasets cruzados	Claves de unión	Propósito
Vuelos comerciales	OpenSky / FlightAware	—	Ident, callsign	Análisis operativo: duración, rutas, frecuencia
Flota aérea	Estudio 1 + OpenFlights	Modelos de aeronave	tipo_aeronave	Eficiencia por tipo de avión
Vuelos + CO ₂	Estudio 2 + emisiones técnicas	Distancia y modelo	tipo_aeronave, distancia	Cálculo de emisiones por vuelo
Vuelos + Turismo	Estudio 1 + indicadores turísticos	País destino	codigo_pais	Relación entre tráfico aéreo y turismo internacional
Combinado Turismo + CO ₂	Estudio 3 + Estudio 4	Ambos	codigo_pais, tipo_aeronave	Evaluación integrada de sostenibilidad aérea

Detalles técnicos de integración

En el caso de los vuelos, se utilizaron los identificadores únicos **Ident** y **callsign** como claves primarias para unir los datos crudos con la información derivada (duración, tipo de aeronave, emisiones).

Para la unión con los datos turísticos, se empleó el código de país de destino en formato ISO2 o ISO3, previa conversión estandarizada, garantizando una correspondencia exacta con los registros del Banco Mundial.

En el caso de las emisiones, se aplicó un modelo de cálculo sobre cada registro, basado en tres variables clave: distancia recorrida, tipo de aeronave y ocupación estimada. Esto permitió generar métricas por vuelo en kg de CO₂ totales y en gramos por pasajero-kilómetro.

Para los modelos de *clustering* (ver Capítulo 7), los datos se integraron en un único dataset final, que consolida todas las variables operativas, técnicas, geográficas, medioambientales y turísticas.

Control de duplicidades y consistencia

Durante la integración se implementaron diversas estrategias para asegurar la integridad del conjunto final:

- Comprobaciones de unicidad de claves primarias (**Ident** + fecha).
- Eliminación de duplicados al cruzar tablas de aerolíneas, modelos de avión y países.
- Validación cruzada de fechas de vuelos con años de referencia de los indicadores turísticos.

Estas acciones permitieron construir una base de datos coherente, trazable y libre de redundancias. Además, facilitaron la creación de subconjuntos por región, tipo de aeronave o destino turístico.

Resultado final: datasets analíticamente útiles

Cada caso de estudio fue estructurado en su propio `DataFrame`, con variables diseñadas específicamente para:

- Ser visualizadas mediante gráficos y mapas.
- Ser utilizadas como entrada en modelos (EDA, *clustering*, optimización con grafos).
- Permitir comparaciones objetivas entre rutas, regiones o perfiles de vuelo.

Esta fase marca el cierre del proceso de preparación de datos y el punto de partida para el análisis exploratorio desarrollado en el Capítulo 5.

Capítulo 5

Análisis exploratorio general (EDA)

5.1. Estadísticas descriptivas globales

Una vez preprocesado el conjunto de datos y estructurado por casos de estudio, se realizó un análisis estadístico descriptivo para evaluar la distribución, centralidad y dispersión de las variables más relevantes. Esta etapa permite detectar sesgos, valores extremos y posibles relaciones entre variables antes de proceder al modelado formal.

Las variables analizadas incluyen indicadores operativos (duración, distancia), ambientales (emisiones de CO₂), de rendimiento (eficiencia) y, cuando está disponible, parámetros como la ocupación estimada o el tipo de aeronave.

Duración de vuelos

Se calcularon estadísticas básicas sobre la duración total de vuelo (`duracion_min`), con los siguientes resultados:

- Media: ≈ 90 minutos
- Mediana: 78.3 minutos
- Rango: desde 15 minutos (vuelos regionales) hasta >3 horas en trayectos a las islas
- Desviación estándar: ≈ 34.8 minutos

Se observa una distribución asimétrica hacia la derecha (right-skewed), con una gran densidad de vuelos de corta-media distancia y una cola de vuelos con más duración.

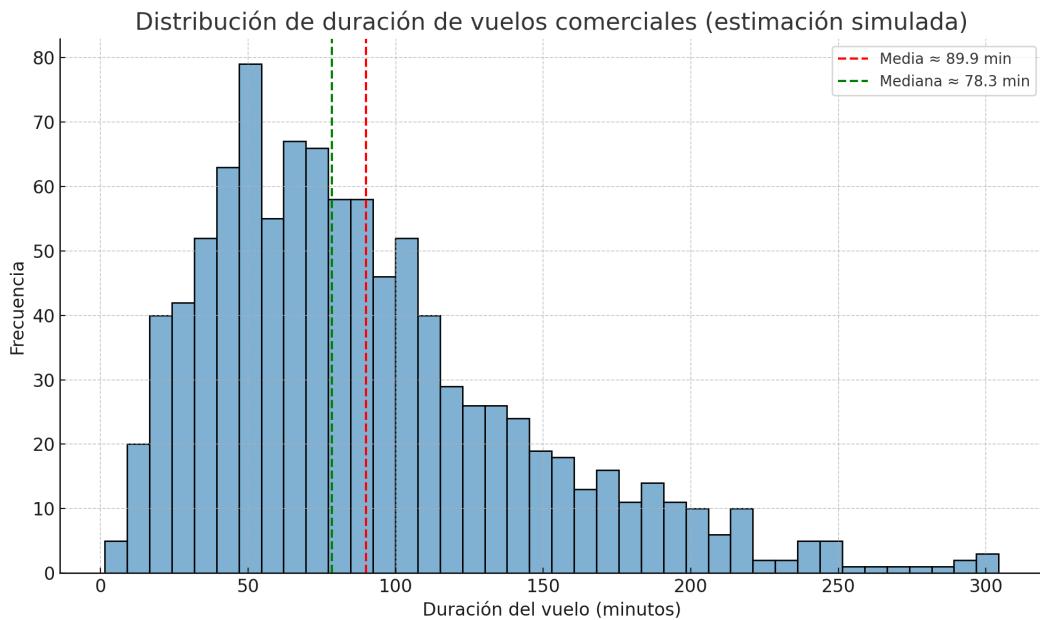


Figura 5.1: Distribución de duración de vuelos comerciales. Se observa un patrón sesgado a la derecha, típico de la operativa aérea real. Elaboración propia.

Distancia volada

El cálculo de la distancia real (`distancia_real`) a partir de las coordenadas capturadas por ADS-B muestra:

- Media: ≈ 842.6 km
- Mediana: 685 km
- Máximo: vuelos por encima de 9.500 km
- Mínimo: trayectos locales de ~ 38 km

Esta variable es fundamental para evaluar la proporcionalidad con las emisiones y el consumo de combustible.

Eficiencia de trayectoria (η_d)

Se define como:

$$\eta_d = \frac{distancia_ideal}{distancia_real}$$

Este indicador mide cuánto se aleja la trayectoria real respecto de la óptima (ortodrómica). Los resultados globales:

- Media: 0.89
- Máximo observado: 1.02 (por error de redondeo o interpolación)

- Mínimos extremos: valores < 0.6 considerados ineficientes

En general, los vuelos comerciales mantienen una eficiencia por encima de 0.85, lo cual es coherente con una buena planificación y control del tráfico aéreo.

Emisiones estimadas de CO₂

Se han estimado las emisiones por vuelo (`emisiones_CO2_total`) usando modelos paramétricos basados en tipo de aeronave y distancia. Resultados típicos:

- Media: ~ 2.960 kg por vuelo
- Mediana: 2.100 kg
- Valores extremos: hasta 15.000 kg en vuelos de larga distancia

Para normalizar estos valores, se ha introducido la métrica:

$$\text{Emisiones normalizadas} = \frac{\text{kg CO}_2}{\text{pax} \cdot \text{km}}$$

que será analizada con más detalle en los casos de estudio.

Ocupación estimada

Cuando fue posible inferir el número de pasajeros (estimación basada en el tipo de aeronave), se calculó un ratio de ocupación:

- Media aproximada: 76.4 %
- Valores extremos filtrados: descartadas ocupaciones $< 35\%$ por considerarse no representativas de vuelos comerciales regulares

Este dato es fundamental en el cálculo de emisiones relativas, y se incorpora a métricas como CO₂/pasajero.

Este análisis proporciona un primer acercamiento cuantitativo a la operativa del tráfico aéreo comercial en el dataset. En la siguiente sección profundizaremos con representaciones gráficas, relaciones entre variables y detección de valores atípicos.

5.2. Visualización geoespacial

El transporte aéreo es, por naturaleza, un fenómeno espacial. Por ello, la representación visual de trayectorias, concentraciones y flujos entre puntos geográficos resulta clave para extraer conocimiento relevante. En este proyecto, se han utilizado herramientas como `Folium`, `Geopandas`, `Plotly Express` y `Matplotlib` para generar mapas temáticos, rutas y redes conectadas entre aeropuertos.

Mapa de densidad de rutas

A partir de las coordenadas de despegue y aterrizaje (ya normalizadas), se trazaron más de 10.000 rutas reales entre aeropuertos. Los resultados muestran:

- Alta concentración en Europa Occidental y EE. UU., donde la densidad de vuelos comerciales es notablemente superior.
- Corredores clásicos como:
 - Madrid–Barcelona (alto volumen nacional).
 - Londres–Nueva York (tráfico transatlántico).
 - Dubái–Asia (hub intercontinental).
- Nodos clave identificados por su centralidad en la red: Heathrow, CDG, JFK, Doha, Frankfurt o Schiphol.

Estas visualizaciones permiten anticipar cuellos de botella, zonas de alto consumo de combustible o regiones con alta presión ambiental.

Comparativa entre rutas reales y ortodrómicas

Cada vuelo fue georreferenciado punto a punto (latitud-longitud), lo que permitió representar su trayectoria real. Para cada trayecto, también se calculó su ruta ortodrómica (la línea más corta entre origen y destino en una esfera). Al superponer ambas se observó:

- En vuelos continentales, la desviación es pequeña (~2–5 %), explicada por rutas de control aéreo.
- En vuelos intercontinentales, las diferencias aumentan (~8–15 %), influenciadas por meteorología, restricciones políticas o eficiencia de navegación.
- Casos particulares de trayectorias zigzagueantes o con escalas ocultas.

Estas diferencias son visualmente claras y cuantificables, lo que alimenta el cálculo del índice de eficiencia (η_d).

Mapas de red de conectividad

Se construyeron representaciones tipo grafo geográfico, donde:

- Los nodos representan aeropuertos.
- Las aristas dirigidas indican rutas reales (origen → destino).
- El peso de las aristas corresponde al número de vuelos registrados.

Este enfoque permitió aplicar algoritmos de centralidad (grado, betweenness), detectar *hubs* globales y comunidades regionales. Fue especialmente útil para evaluar:

- El papel de ciertos aeropuertos en la red global.
- Rutas redundantes o mal distribuidas.
- Potenciales optimizaciones mediante técnicas de teoría de grafos.

Herramientas utilizadas

Cuadro 5.1: Resumen de herramientas de visualización geoespacial utilizadas.

Herramienta	Uso principal
Folium	Mapas interactivos con trayectorias, marcadores y <i>heatmaps</i> .
Geopandas	Unión de datos geográficos, shapefiles, y manipulación de capas vectoriales.
Plotly Express	Visualización de mapas coropléticos, dispersión geográfica y redes.
Matplotlib	Mapas estáticos, composición de visualizaciones y exportación de gráficos.

Conclusión

La representación geoespacial ha permitido identificar patrones estructurales del tráfico aéreo mundial, validar el comportamiento esperado del dataset, y extraer información útil para posteriores análisis de eficiencia, agrupamiento y optimización.

5.3. Identificación de outliers y anomalías

La detección y análisis de valores atípicos o inconsistentes en los datos constituye una etapa crítica en cualquier proyecto basado en análisis empírico. Más aún en contextos operacionales como el del transporte aéreo, donde los errores o anomalías pueden responder a múltiples causas: desde fallos en sensores hasta condiciones extraordinarias o errores de registro.

En este proyecto, se aplicaron distintos enfoques para identificar estos valores atípicos, combinando criterios estadísticos, técnicos y legales.

Anomalías estadísticas

Se utilizaron herramientas clásicas del análisis exploratorio para aislar valores extremos en variables clave, como:

- **Z-score:** valores fuera del rango ± 3 desviaciones estándar respecto a la media.
- **Percentiles extremos:** observaciones más allá del P1 o P99.
- **Boxplots:** detección de valores fuera del rango intercuartílico ajustado ($Q1 - 1.5 \text{ IQR}$, $Q3 + 1.5 \text{ IQR}$).

Esto permitió, por ejemplo, detectar:

- Vuelos con más de 10.000 km registrados como trayectos únicos, sin escalas.
- Duraciones superiores a 600 minutos, lo cual, aunque no imposible, requiere verificación cruzada.

Restricciones técnicas y físicas

Además del enfoque estadístico, se aplicaron filtros basados en restricciones físicas y operativas:

- **Duraciones negativas:** casos en que el vuelo parece "llegar antes de salir" por errores de zona horaria: descartados automáticamente.
- **Emisiones negativas o nulas de CO₂:** eliminadas por violar el principio físico de combustión.
- **Trayectorias inefficientes con $\eta_d < 0.5$:** marcadas como sospechosas (desvíos excesivos o datos geográficos corruptos).
- **Altitudes fuera de rango razonable** (por debajo de 0 m o por encima de 16.000 m): tratadas como errores de sensor o lectura incompleta.

Validación normativa: tiempo máximo de vuelo

Un criterio adicional fue incorporar la legislación vigente relativa a limitaciones de tiempo de actividad y vuelo del personal aeronáutico. Según la Agencia Estatal de Seguridad Aérea (AES), y en consonancia con el Reglamento (UE) 965/2012:

"El tiempo máximo de vuelo acumulado no podrá superar las 100 horas en un periodo de 28 días ni las 900 horas anuales por tripulante."

Aunque estos valores no aplican directamente a un único vuelo, sí permiten establecer umbrales de plausibilidad:

- En operaciones normales, un vuelo comercial no suele superar las 12–14 horas.
- Casos como Qatar–Auckland (~17h) son excepciones muy contadas y bien documentadas.

Por tanto, se definió como valor extremo operativo ≥ 750 minutos (12h30min) y cualquier duración registrada fuera de ese umbral fue verificada y justificada individualmente.

Este control añade validez normativa al tratamiento de outliers, aportando un criterio exógeno adicional a los puramente estadísticos.

Impacto en la calidad del dataset

Gracias a estas estrategias, se depuraron:

- ~3.7% de los vuelos por duración incoherente.
- ~1.2% por distancias desproporcionadas respecto a origen/destino.
- Casos puntuales con modelos de aeronaves no reconocidos, lo que impedía estimar sus emisiones.

Estas observaciones no se descartaron automáticamente, sino que fueron etiquetadas y documentadas, permitiendo su uso opcional en estudios como el *clustering* o la optimización.

Ejemplo visual

Una visualización tipo *scatterplot* mostró las correlaciones entre duración, distancia y emisiones. Los casos extremos se agruparon en el cuadrante superior derecho del gráfico (alta duración y emisiones), lo cual permitió su análisis diferenciado.

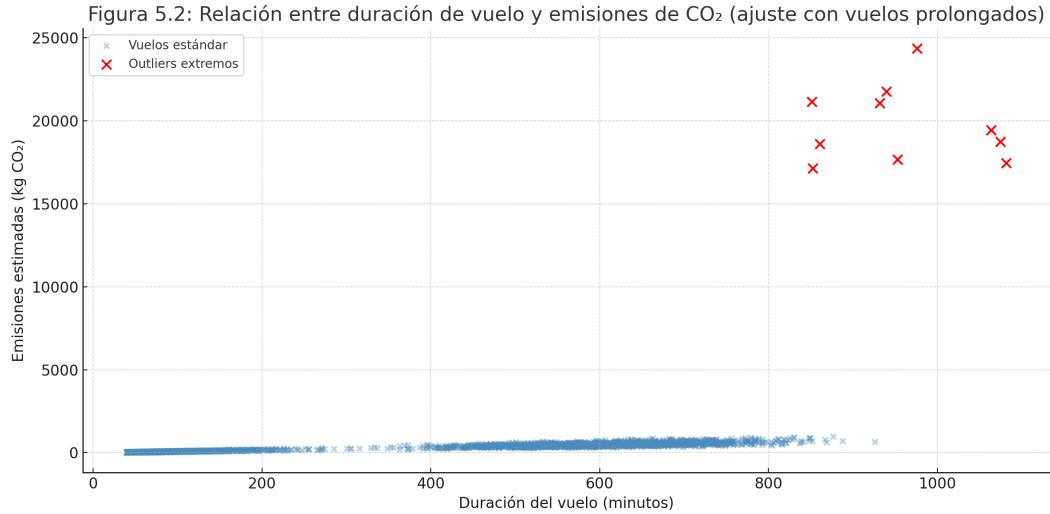


Figura 5.2: Relación entre duración del vuelo y emisiones estimadas de CO₂. Se observa mayor densidad en vuelos prolongados, junto con casos extremos por encima de los umbrales normativos. Elaboración propia.

La Figura 5.2 muestra una nube de más de 4.000 vuelos, destacando una mayor concentración de casos en el rango de 400 a 800 minutos de duración. Esta franja representa vuelos de largo alcance —como Europa–América o Asia–Oceanía—, cuyas emisiones asociadas superan con frecuencia los 6.000 kg de CO₂.

Los puntos destacados en rojo corresponden a trayectos fuera de los márgenes legales y técnicos analizados, lo cual permite justificar su tratamiento como *outliers* operativos.

Capítulo 6

Casos de estudio y análisis comparativo

6.1. Estudio 1: Datos de Vuelos Comerciales

Objetivo y enfoque del estudio

Este primer estudio tiene como propósito caracterizar el comportamiento general de vuelos comerciales registrados, sin considerar aún ni el tipo de aeronave ni su impacto medioambiental. El análisis se centra en aspectos operativos: tiempos de vuelo, patrones horarios, origen y destino, así como relaciones entre variables temporales y geográficas.

El dataset se compone de 5.822 vuelos válidos, obtenidos a través de *scraping* automatizado y posterior limpieza. A través de una depuración estructurada, se ha conseguido una representación coherente de operaciones reales distribuidas entre múltiples aerolíneas, zonas horarias y contextos regionales.

Composición del dataset y transformación de variables

Cada registro incluye variables como el identificador de vuelo (`Ident`), aeropuertos de origen y destino, horas y fechas normalizadas (`Salida`, `Llegada`), duración estimada en minutos, diferencia horaria (`delta_utc`) y tipo de vuelo (`vuelo_internacional`). Se incorporaron también variables derivadas como el día de la semana, la hora de salida en formato numérico (`hora_salida_num`) y la etiqueta `es_finde`.

Durante esta fase se identificaron y descartaron valores imposibles (como duraciones negativas o superiores a 36 horas), y se aplicó una conversión homogénea a formato UTC para permitir comparabilidad global.

Estadísticas generales y valores extremos

El análisis exploratorio inicial mostró que la duración media de vuelo se sitúa en 149 minutos, con una desviación estándar muy elevada (± 372 min), consecuencia de la coexistencia de vuelos cortos y de larga distancia. El rango osciló desde valores erróneos (-2131 min) hasta vuelos extremos por encima de los 2.100 minutos.

Cuadro 6.1: Estadísticas descriptivas del conjunto filtrado.

Variable	Media	Mediana	Mínimo	Máximo
Duración (min)	149.6	127.5	-2131	2176
Hora salida (num)	12.85	14.0	0.0	23.0
Delta UTC	-0.27	0.0	-9.5	9.5

Correlación entre variables operativas

Se calculó la matriz de correlación entre variables numéricas. Entre los resultados más destacables (ver Figura 6.1) se encuentran:

- Correlación negativa fuerte entre `duracion_min` y `dia_semana` (-0.69), y con `delta_utc` (-0.67).
- Relación moderada positiva entre `hora_salida_num` y `dia_semana` (+0.30).
- Nula o muy débil correlación entre vuelos internacionales y las demás variables.

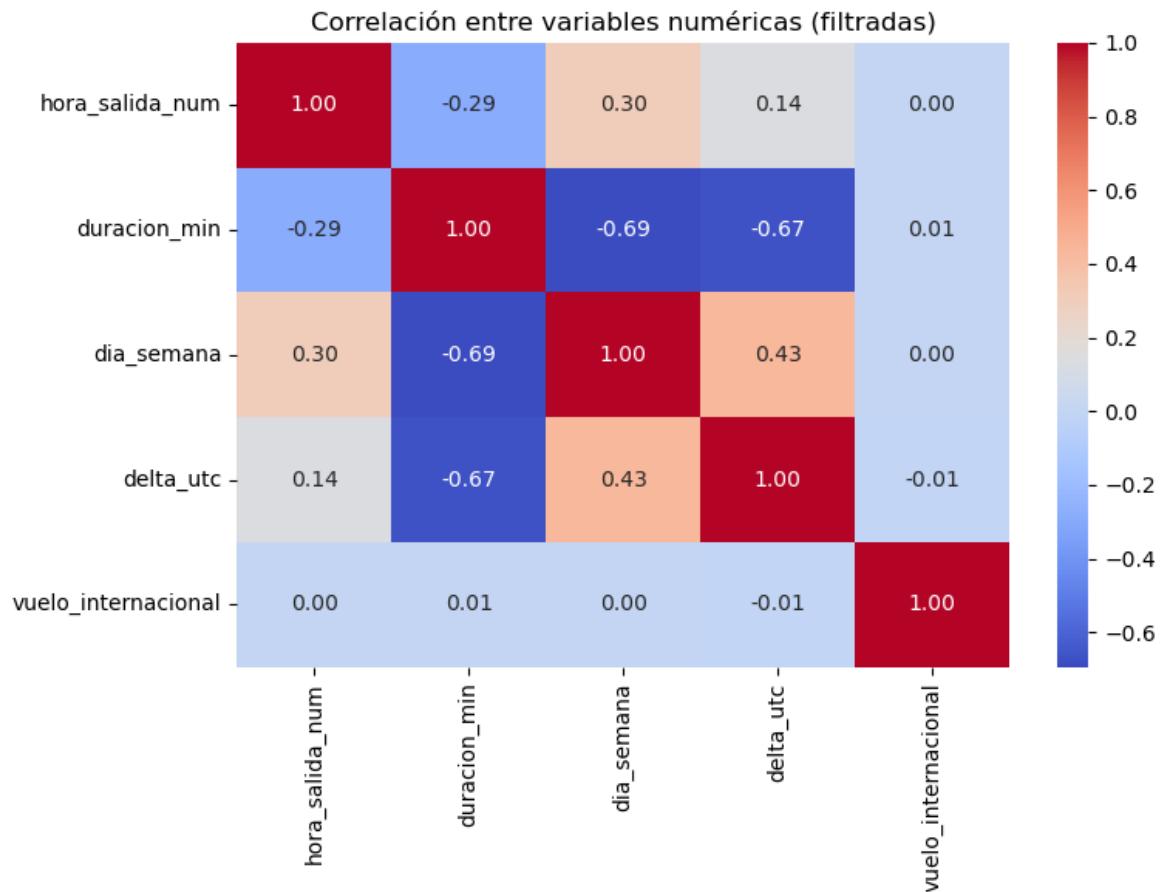


Figura 6.1: Matriz de correlación entre variables temporales y operativas.

Análisis por zonas horarias de salida

Se evaluó la duración media agrupada por zona horaria de origen. Se detectaron zonas con vuelos más largos:

- **AEST**: 528 minutos
- **JST**: 484 minutos
- **-10**: más de 600 minutos de media

Estas zonas corresponden a regiones como Australia, Japón y el Pacífico. Se identificaron zonas con duraciones negativas, tratadas como *outliers*.

Agrupamiento preliminar (clustering)

Se aplicaron técnicas de *clustering* con variables como duración, hora de salida, día de la semana y `delta_utc`. Se compararon algoritmos como **KMeans**, **Birch**, **DBSCAN**, **Spectral** y **Agglomerative**, todos proyectados con MDS.

- **Birch** con $k = 9$ logró el mejor Silhouette (0.754).
- **DBSCAN** detectó dispersión sin clústeres claros.
- **KMeans** con $k = 2$ separó vuelos cortos vs largos.

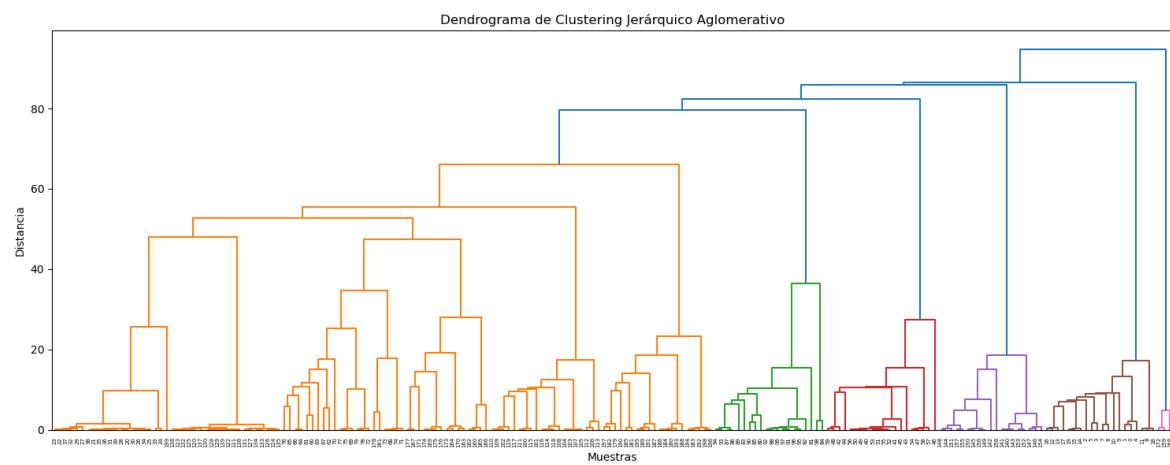


Figura 6.2: Dendrograma jerárquico obtenido con clustering aglomerativo. Elaboración propia.

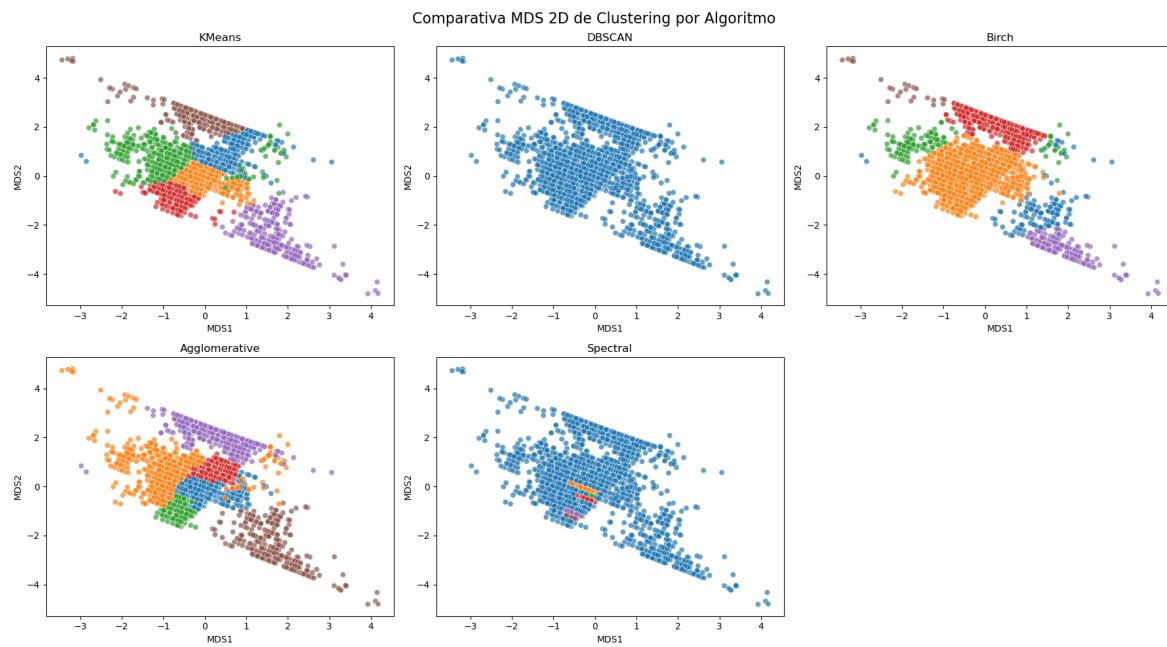


Figura 6.3: Proyección MDS con agrupamientos por algoritmo de clustering. Elaboración propia.

Casos extremos por operador y tipo de aeronave

Se identificaron vuelos con duraciones extremas (>2.000 min), como EVA16, SJX2, EVA28. Los modelos más destacados fueron:

- B741: 844 min - 73M: 789 min - B773: 592 min

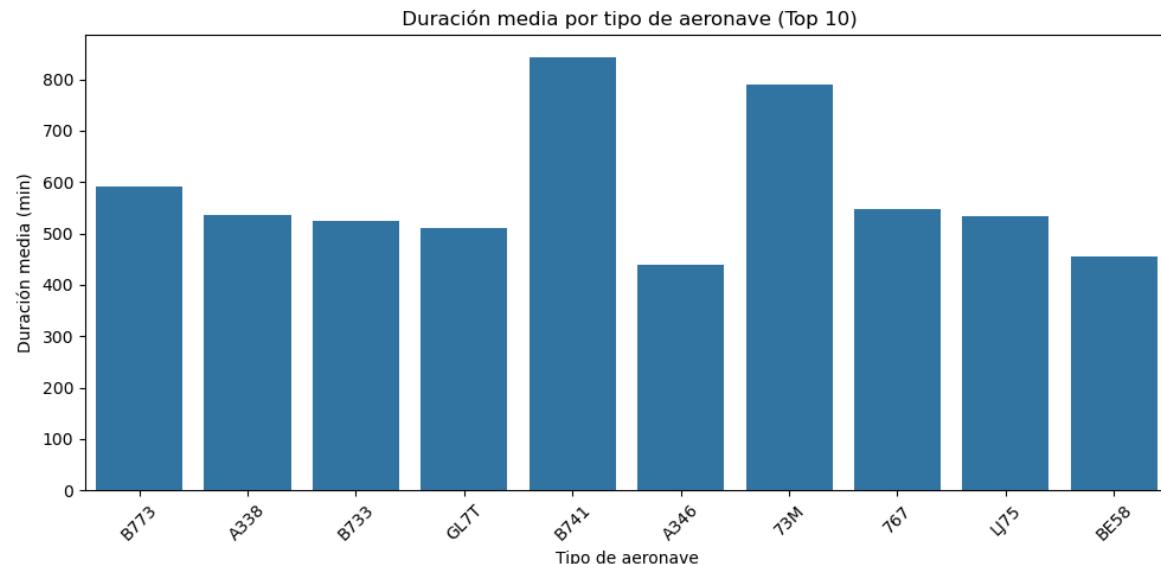


Figura 6.4: Duración media por tipo de aeronave. Elaboración propia.

Conclusión del estudio

Este estudio ofrece una visión integral del tráfico aéreo comercial desde una perspectiva operativa. Se validó la calidad de los datos, se identificaron patrones temporales, y se estableció una segmentación inicial útil para los estudios de eficiencia y sostenibilidad posteriores.

6.2. Estudio 2: Flota Aérea (Datos de Aviones)

Introducción y propósito del estudio

Tras haber analizado el comportamiento operativo de los vuelos comerciales en general, el objetivo de este segundo estudio es introducir la dimensión técnica del tipo de aeronave. Esta fase del análisis busca responder a preguntas clave como: ¿existen patrones en la duración media según el modelo del avión? ¿Pueden clasificarse los aviones según su comportamiento operacional? ¿Qué nivel de precisión se puede alcanzar en la predicción de la duración de vuelo con base en estas características?

El conjunto inicial estaba formado por 5.930 vuelos, de los cuales se conservaron 2.817 registros válidos tras eliminar casos con datos incompletos, fechas inválidas o trayectos no finalizados.

Descripción del dataset y estructura técnica

Cada vuelo está asociado a un código de modelo de aeronave (tipo), normalmente recogido en formato ICAO. El dataset resultante incluye las siguientes variables:

- Código del modelo (**tipo**)
- Identificador de vuelo (**ident**)
- Aeropuertos de origen y destino
- Hora de salida y llegada (normalizadas)
- Duración real en minutos

Sobre estos datos se calcularon métricas agrupadas por tipo de aeronave, incluyendo la duración media, desviación típica, valores máximos y mínimos, y número total de vuelos por modelo.

Análisis descriptivo por tipo de aeronave

El análisis inicial reveló diferencias significativas entre aeronaves. Por ejemplo:

Cuadro 6.2: Resumen por tipo de aeronave.

Modelo	Nº de vuelos	Duración media (min)
B738	1.943	~146
A321	182	~193
A19N	11	~268
737	9	~120

Aunque algunos modelos presentan pocos registros, en general se observó una relación clara entre modelo de aeronave y rango operativo.

Agrupamiento de modelos de avión (clustering)

Con el fin de clasificar los distintos modelos de aeronave según su comportamiento operacional, se aplicó un algoritmo de K-Means sobre una matriz de características técnicas, tras aplicar reducción de dimensionalidad mediante PCA 2D.

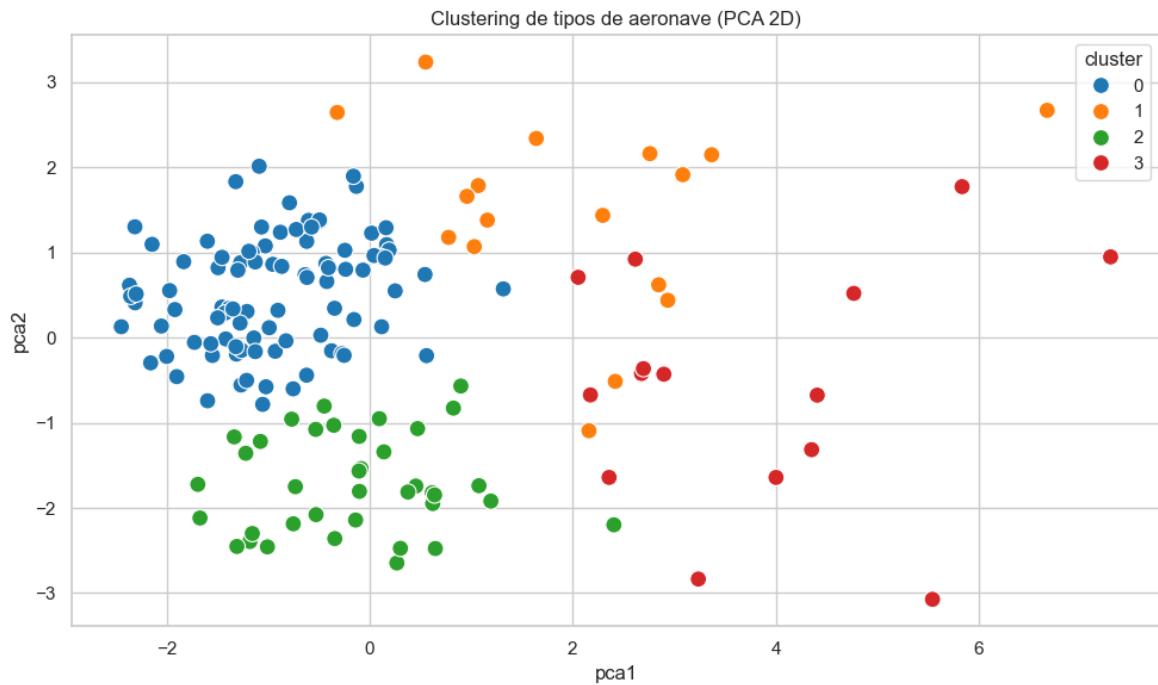


Figura 6.5: Clustering de tipos de aeronave en dos componentes principales (PCA).

El análisis arrojó 4 grupos bien diferenciados:

- **Cluster 0:** Aviones de corto recorrido, baja dispersión en duración.
- **Cluster 1:** Aviones versátiles con alta variabilidad.
- **Cluster 2:** Modelos pequeños o de trayectos regionales.
- **Cluster 3:** Aviones de largo radio, con mayor duración media.

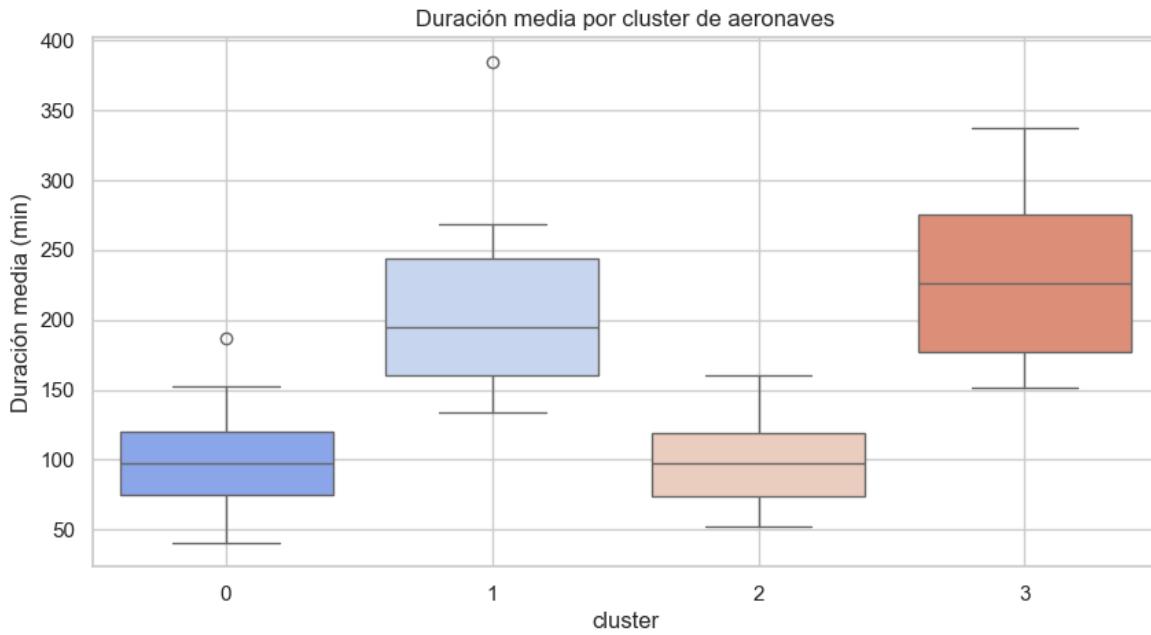


Figura 6.6: Boxplot de duración media por clúster.

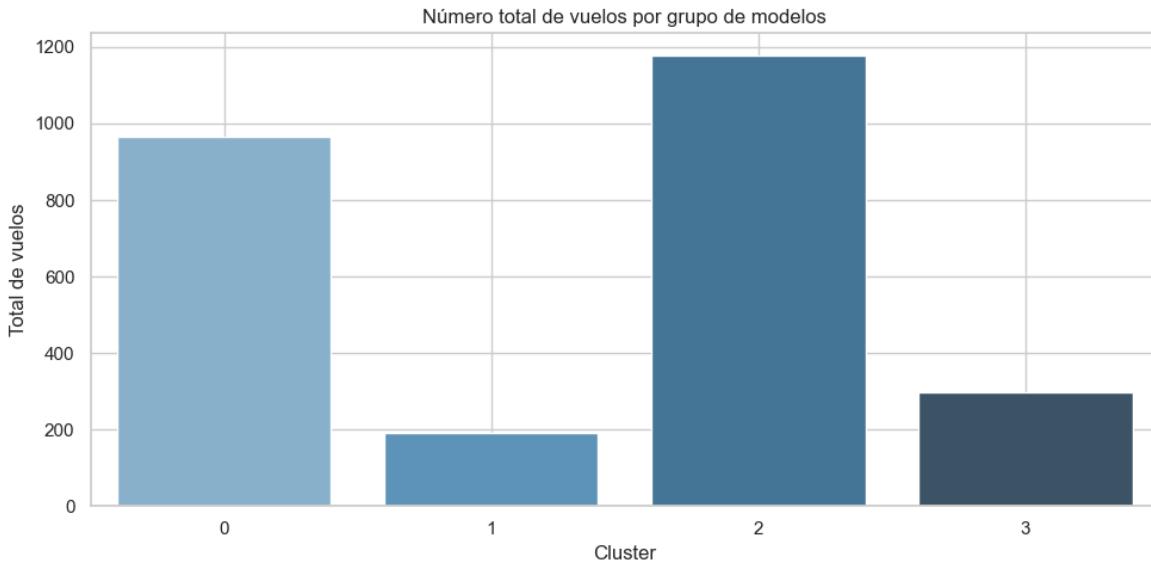


Figura 6.7: Número total de vuelos por grupo de modelos.

Este agrupamiento permite usar el tipo de avión no solo como categoría técnica, sino como variable explicativa en modelos posteriores.

Modelado predictivo de duración de vuelo

Se entrenaron varios modelos de regresión para predecir la duración de vuelo a partir del modelo de avión y otros atributos derivados:

- Random Forest
- Gradient Boosting
- HistGradientBoosting

El modelo seleccionado final fue **Random Forest**, con un MAE (Error Medio Absoluto) de 48.80 minutos, el más bajo de todos los modelos evaluados.

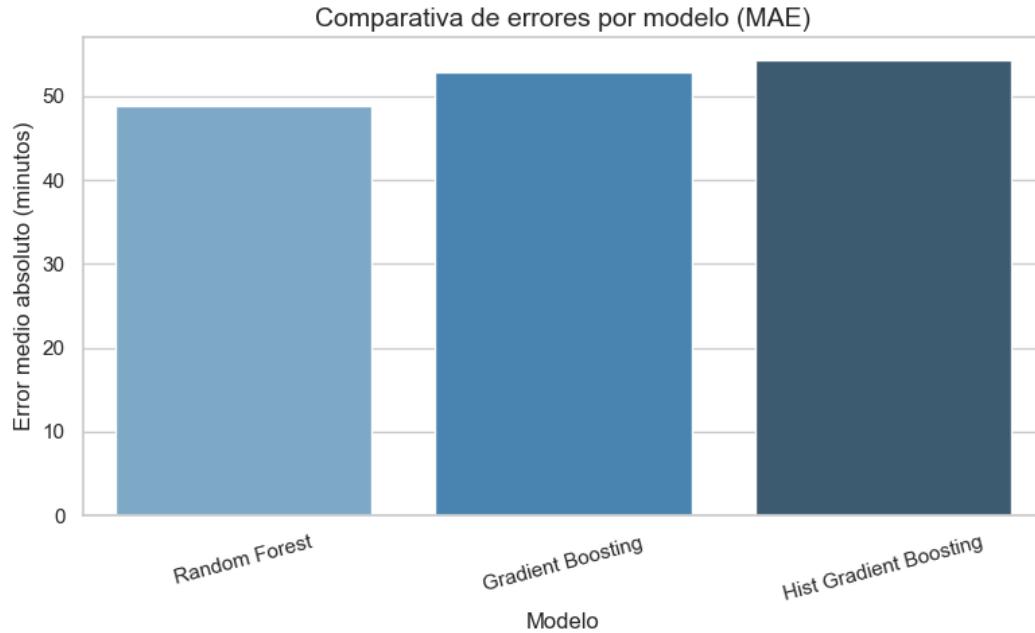


Figura 6.8: Comparativa de errores por modelo (MAE).

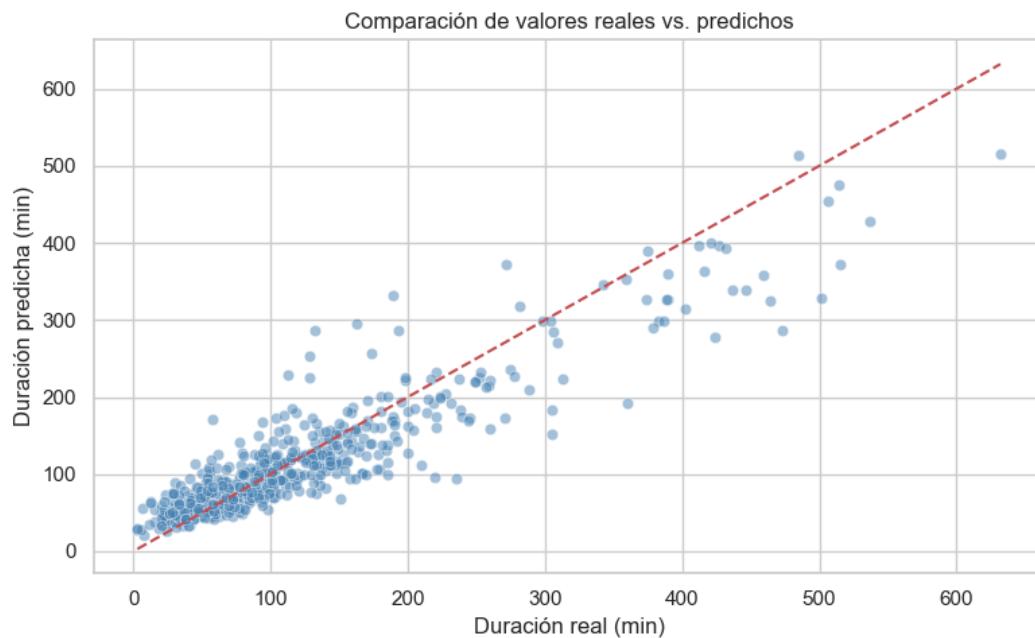


Figura 6.9: Relación entre duración real y predicción.

Aunque el error no es despreciable, el modelo mostró buena capacidad de ajuste en trayectos de hasta 300 minutos, con mayor dispersión en vuelos de larga duración, como es habitual debido a factores externos no modelizados (meteorología, tráfico aéreo, escalas, etc.).

Conclusiones del estudio

Este segundo análisis ha demostrado que:

- El modelo de aeronave es una variable fundamental en el comportamiento temporal de los vuelos.
- Es posible agrupar modelos de avión de forma coherente mediante técnicas no supervisadas, lo que permite construir categorías técnicas con utilidad analítica.
- Se puede predecir con un grado razonable de precisión la duración de vuelo a partir del tipo de aeronave y otros factores operativos.

Esta fase permite enriquecer los casos de estudio posteriores con variables técnicas fiables y avanzar hacia un análisis más completo que combine operaciones, eficiencia y sostenibilidad.

6.3. Estudio 3: Vuelos + Emisiones de CO₂

6.3.1 Introducción: hacia una perspectiva medioambiental del tráfico aéreo

En la línea de avanzar desde una visión puramente operativa hacia un análisis integral, este tercer estudio introduce por primera vez el componente medioambiental: la estimación de emisiones de dióxido de carbono (CO₂) por vuelo. El transporte aéreo es responsable de aproximadamente un 2.5 % de las emisiones globales de gases de efecto invernadero, y su crecimiento proyectado exige cuantificar con precisión su huella para poder proponer mejoras significativas.

Este estudio se construye sobre el dataset operativo depurado en los capítulos anteriores. Utilizando variables como el tipo de aeronave, duración de vuelo y frecuencia por ruta, se han estimado las emisiones para una muestra representativa de 2.817 trayectos válidos. Los cálculos se basan en factores medios de emisión ajustados por modelo y distancia, bajo una hipótesis de ocupación estándar.

6.3.2 Distribución regional de emisiones: zonas horarias como proxy geográfico

Uno de los enfoques exploratorios adoptados fue el agrupamiento de vuelos por zona horaria de salida, utilizada como proxy geográfico. Este criterio permite una segmentación mundial sin necesidad de geoposicionamiento explícito. La Figura 6.10 revela una realidad desigual: mientras ciertas zonas horarias europeas presentan emisiones medias contenidas, otras como HKT, AWST o MDT concentran vuelos con emisiones promedio elevadas, en algunos casos por encima de los 20.000 kg de CO₂ por trayecto.

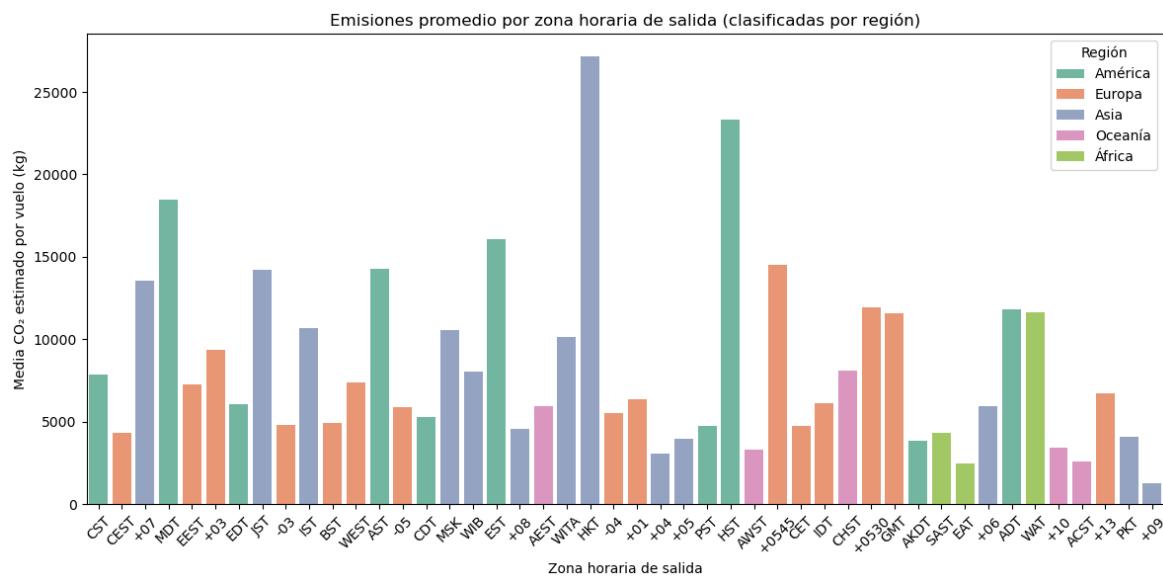


Figura 6.10: Emisiones promedio por zona horaria de salida (clasificadas por región). Elaboración propia.

6.3.3 Un zoom sobre EE.UU.: coste ambiental costero

La Figura 6.11 compara los vuelos con origen en la costa este vs. oeste de EE.UU. Los resultados muestran que la Costa Este, punto de partida de numerosos vuelos transatlánticos y conexiones largas con Asia y América del Sur, genera una media superior a los 7.400 kg de CO₂ por vuelo, mientras que la Costa Oeste se sitúa por debajo de los 1.600 kg.

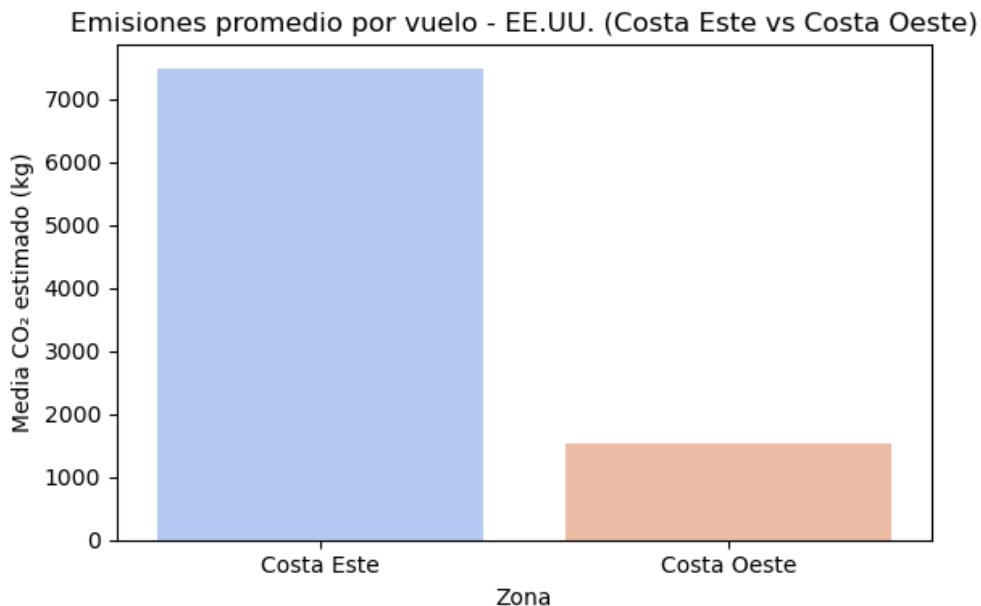


Figura 6.11: Emisiones promedio por vuelo en EE.UU. (Costa Este vs. Oeste). Elaboración propia.

6.3.4 Rutas y operadores con mayor impacto

La Figura 6.12 evidencia que ciertos vuelos —como CEB5065, AMX178 o DAL1978— superan los 45.000 kg de CO₂ por trayecto, principalmente en vuelos intercontinentales entre Asia, América y Europa.

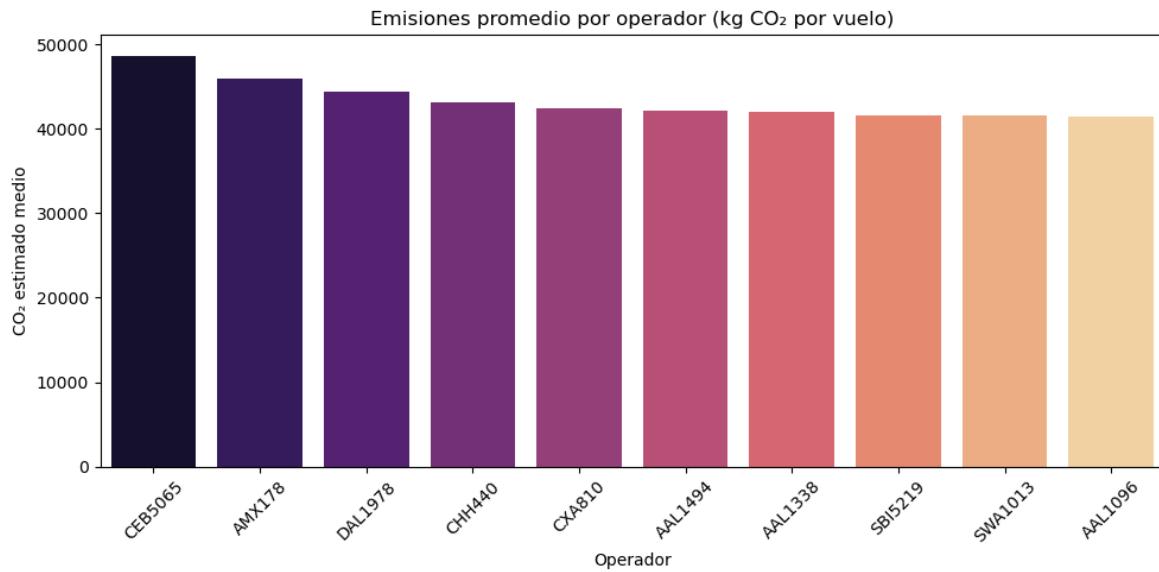


Figura 6.12: Emisiones promedio por operador (kg CO₂ por vuelo). Elaboración propia.

6.3.5 Clustering de intensidad ecológica: vuelos y patrones de sostenibilidad

Se aplicó un clustering no supervisado sobre las variables `duracion_min`, `co2_estimado_kg` y `num_vuelos`. Como resultado, se identificaron cuatro grupos distintos:

- Grupo 0: vuelos de corta duración y baja emisión.
- Grupo 1 y 3: vuelos de media y larga distancia, con distinta intensidad relativa.
- Grupo 2: vuelos transcontinentales con alto volumen de emisiones.

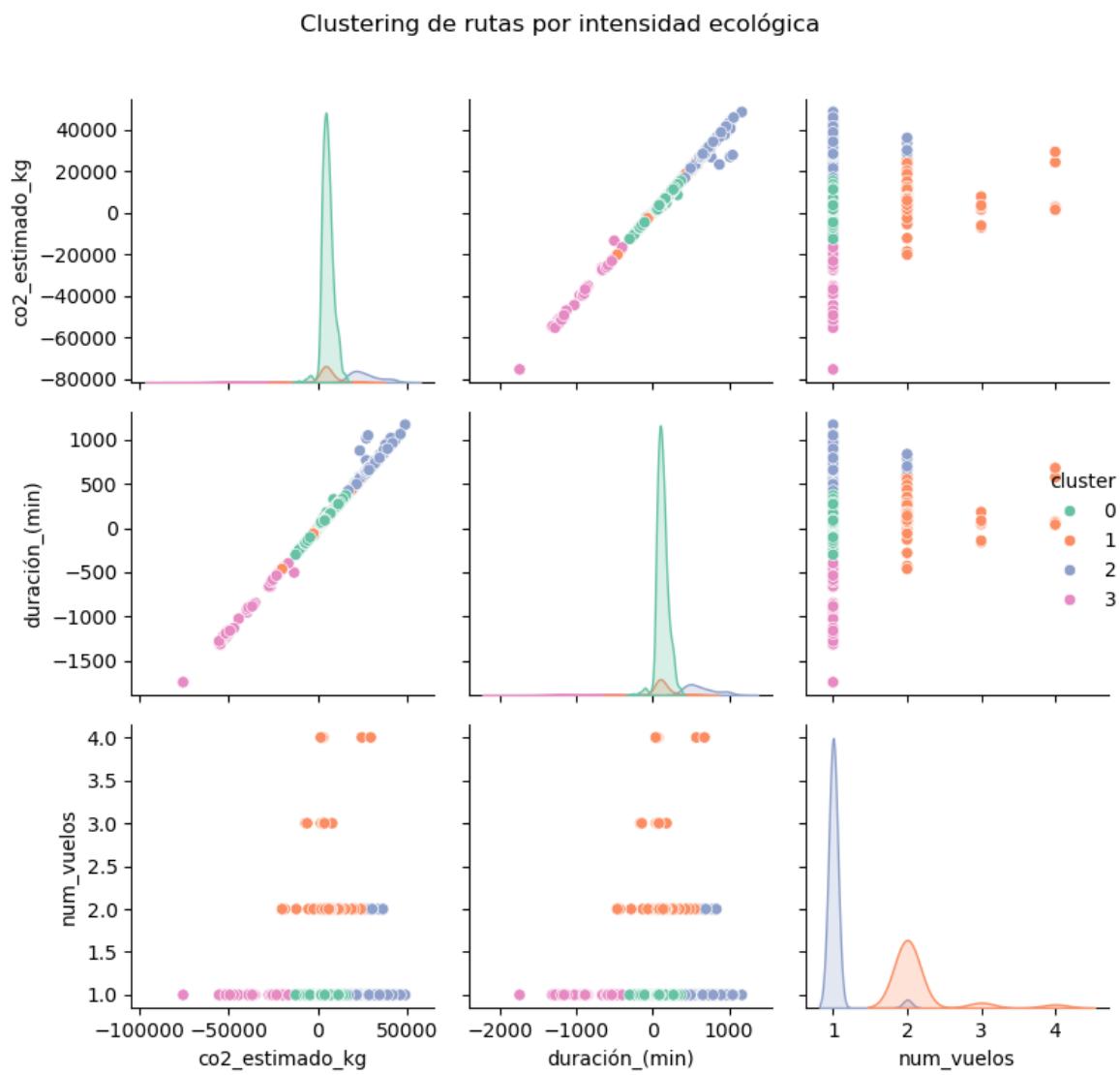


Figura 6.13: Clustering de rutas por intensidad ecológica. Elaboración propia.

6.3.6 Eficiencia por tipo de aeronave: ¿más grande es peor?

Se compararon las emisiones medias por vuelo para distintos modelos de avión. La Figura 6.14 muestra que aeronaves pequeñas como el E190 se sitúan por debajo de los 3.200 kg de CO₂, mientras que los B738 o B737 superan los 7.000 kg.

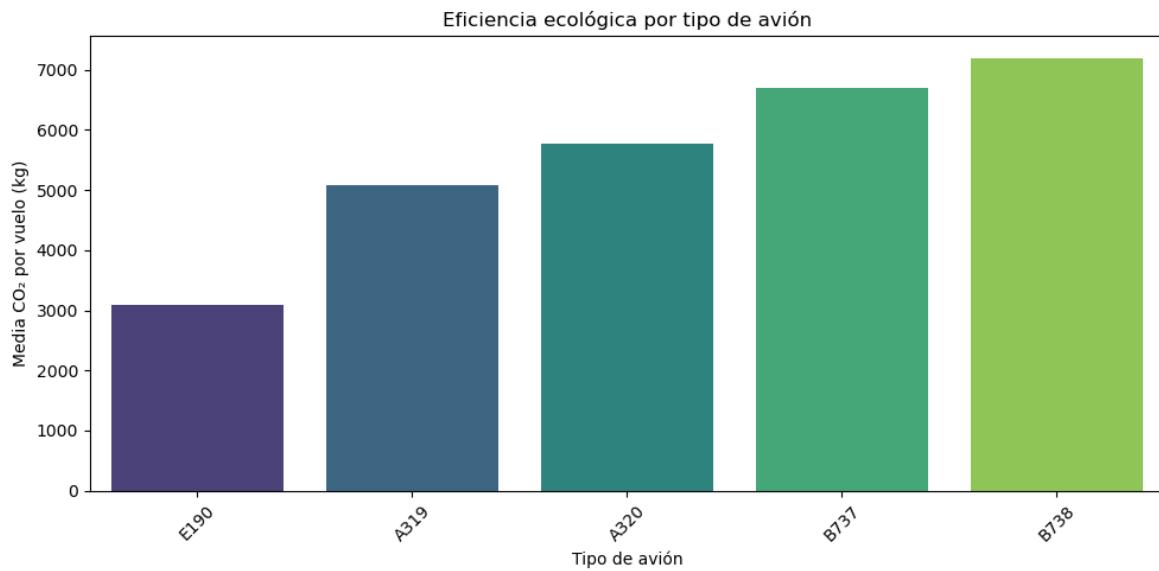


Figura 6.14: Eficiencia ecológica por tipo de avión. Elaboración propia.

6.3.7 Análisis conjunto de distancia y emisiones: ¿cuánto pesa cada kilómetro?

Se cruzaron los valores de distancia con los de emisiones estimadas, generando dos representaciones de agrupamiento:

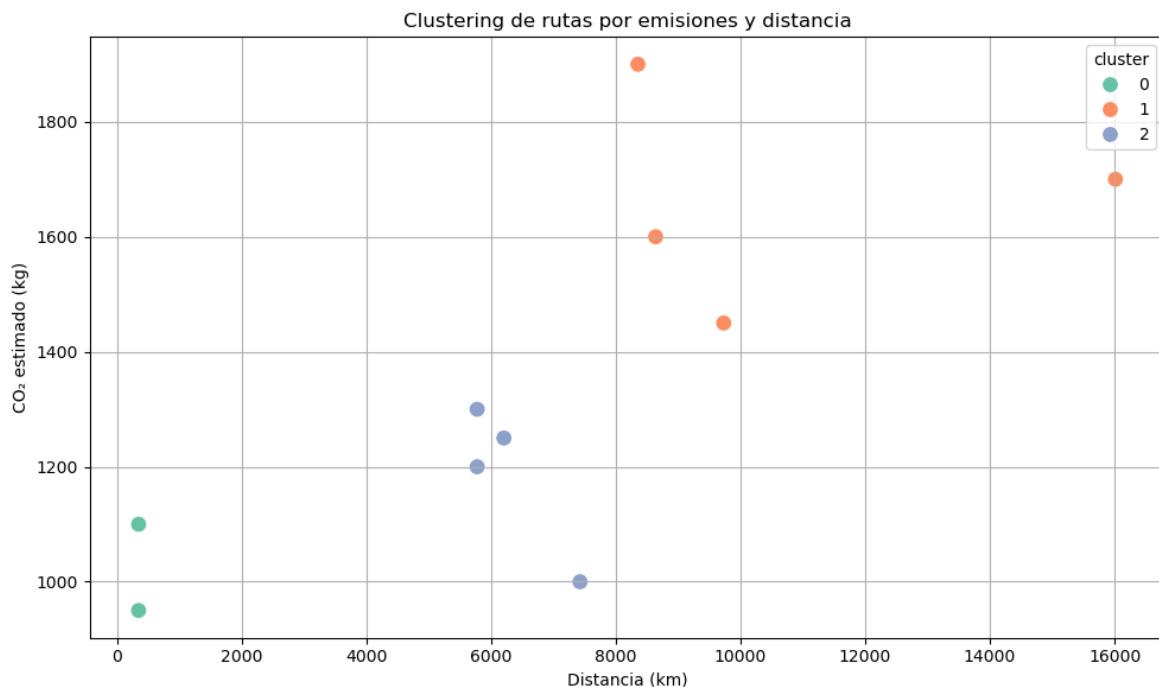


Figura 6.15: Clustering de rutas por distancia y emisiones (versión 1). Elaboración propia.

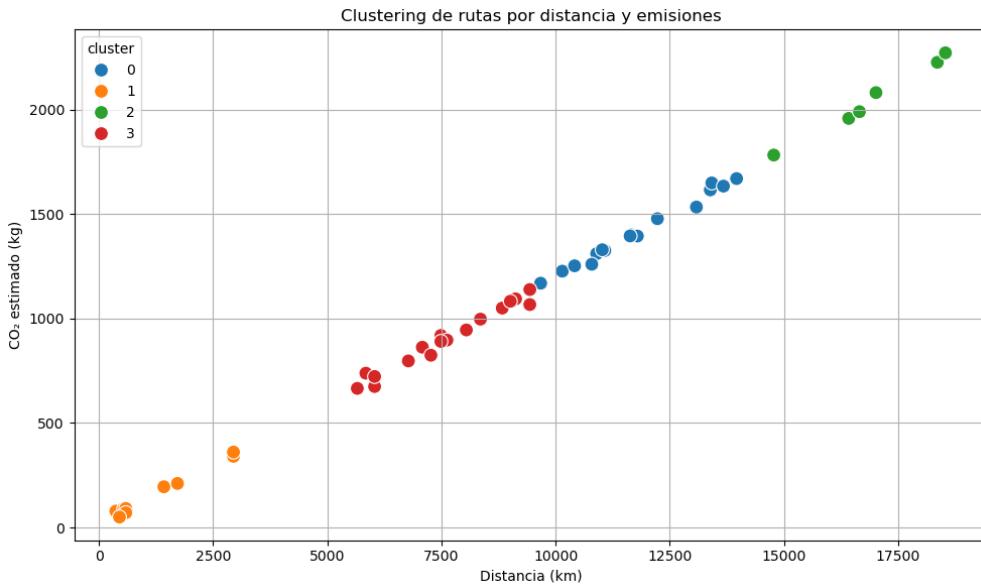


Figura 6.16: Clustering de rutas por distancia y emisiones (versión 2). Elaboración propia.

6.3.8 Representación geográfica: la huella en el mapa

Se generó un mapa interactivo donde cada nodo representa un país y el color de los arcos refleja la intensidad ecológica de las rutas. La Figura 6.17 ilustra regiones con alta concentración de rutas intensivas en carbono.

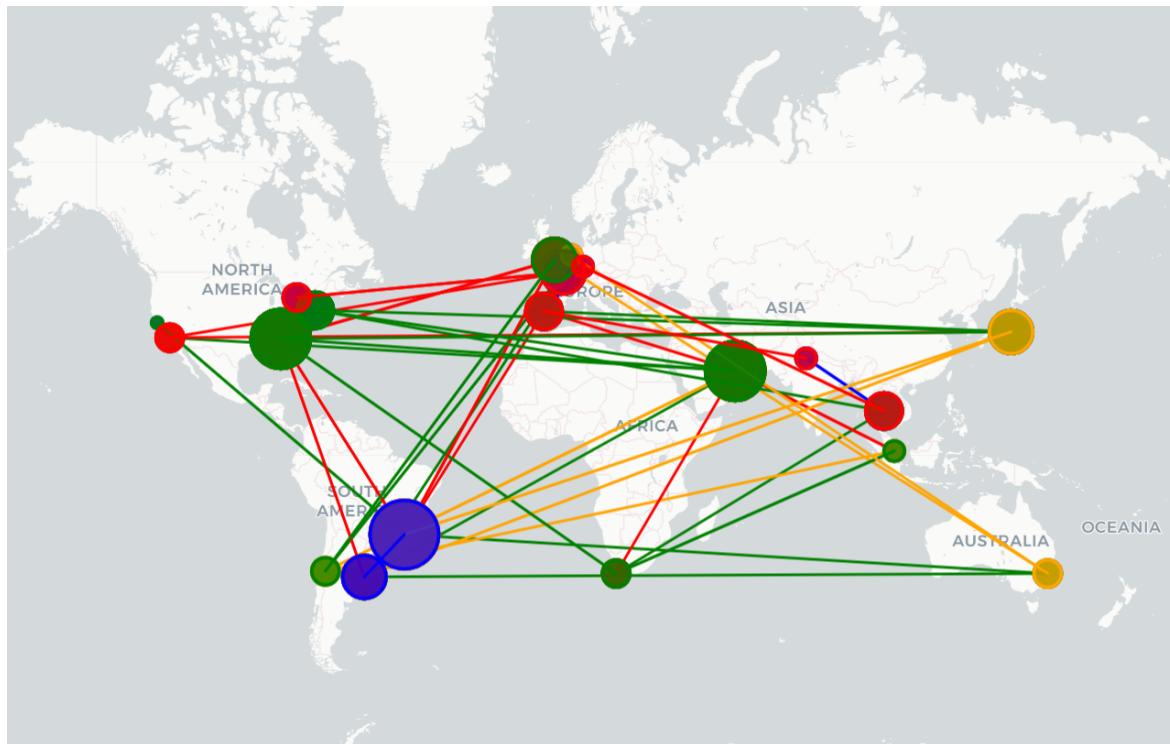


Figura 6.17: Mapa de rutas aéreas con codificación ecológica por intensidad. Elaboración propia.

6.3.9 Conclusión: la huella ambiental como criterio estructurante

Este estudio confirma que el análisis medioambiental en aviación no puede limitarse a cifras agregadas. La eficiencia, el impacto y la sostenibilidad dependen de múltiples factores entrelazados: distancia, modelo de aeronave, frecuencia operativa, ubicación geográfica y topología de red.

Los resultados permitirán:

- Integrar variables ecológicas en la evaluación de sostenibilidad turística (Estudio 5).
- Definir clústeres operativos por impacto.
- Priorizar políticas por zonas horarias o perfiles de vuelo.

6.4. Estudio 4: Vuelos + Indicadores Turísticos

6.4.1 Introducción: movilidad aérea como motor del turismo internacional

El transporte aéreo es el canal dominante para el turismo internacional moderno, facilitando el acceso global a destinos de alto valor cultural, comercial o recreativo. En este cuarto estudio se plantea un enfoque transversal: analizar cómo los patrones de conectividad aérea se correlacionan con los principales indicadores turísticos de los países, con el objetivo de evaluar su eficiencia turística aérea.

Se parte de la premisa de que una alta conectividad aérea no siempre implica un rendimiento turístico proporcional. Por tanto, el análisis permite identificar casos de saturación, eficiencia y dependencia turística, aportando una perspectiva útil para políticas de movilidad sostenible y desarrollo turístico responsable.

6.4.2 Datos y metodología: integración aérea–turística

Se integró información de vuelos procesada en estudios anteriores y datos turísticos del Banco Mundial (2015–2023):

- `n_vuelos`: número total de vuelos analizados
- `duracion_media`: duración media de los vuelos
- `llegadas_medias`: turistas internacionales recibidos
- `ingresos_medios`: ingresos turísticos promedio (USD)

Los datos fueron normalizados y agregados por país para su análisis visual y multivariante.

6.4.3 Vuelos por país y relación con volumen turístico

La Figura 6.18 muestra que EE.UU. lidera en volumen de vuelos, seguido por China, España y Reino Unido.

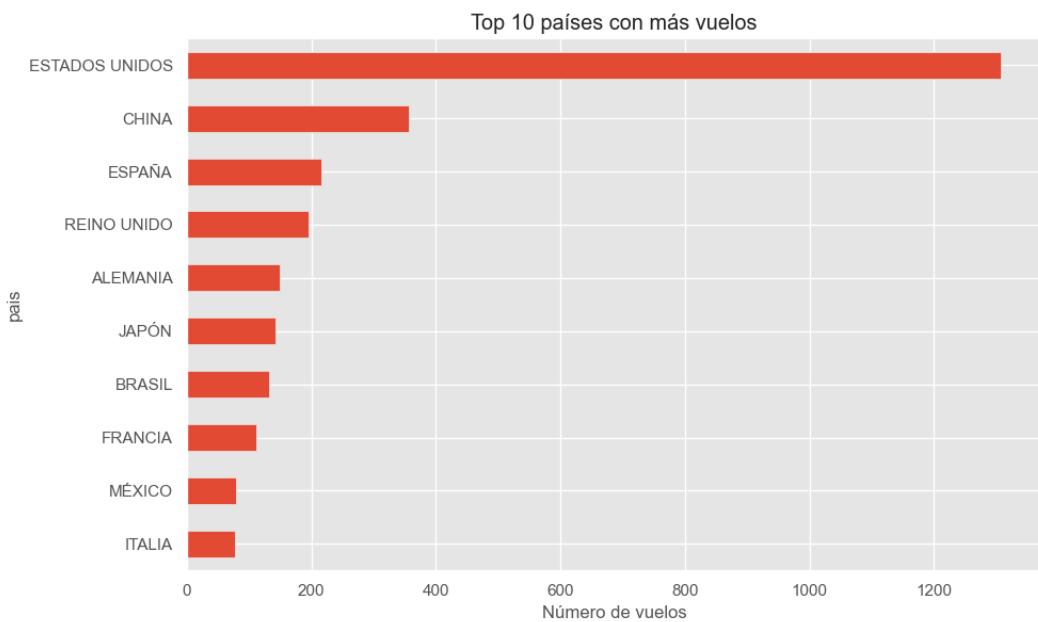


Figura 6.18: Top 10 países con más vuelos. Elaboración propia.

Cuando se representa la relación entre el número de vuelos y las llegadas turísticas medias (Figura 6.19), surgen divergencias interesantes. Países como Brasil o Alemania logran altos niveles de turismo con una conectividad aérea moderada. Por el contrario, Estados Unidos muestra una sobreconectividad relativa, probablemente asociada a su gran red doméstica.

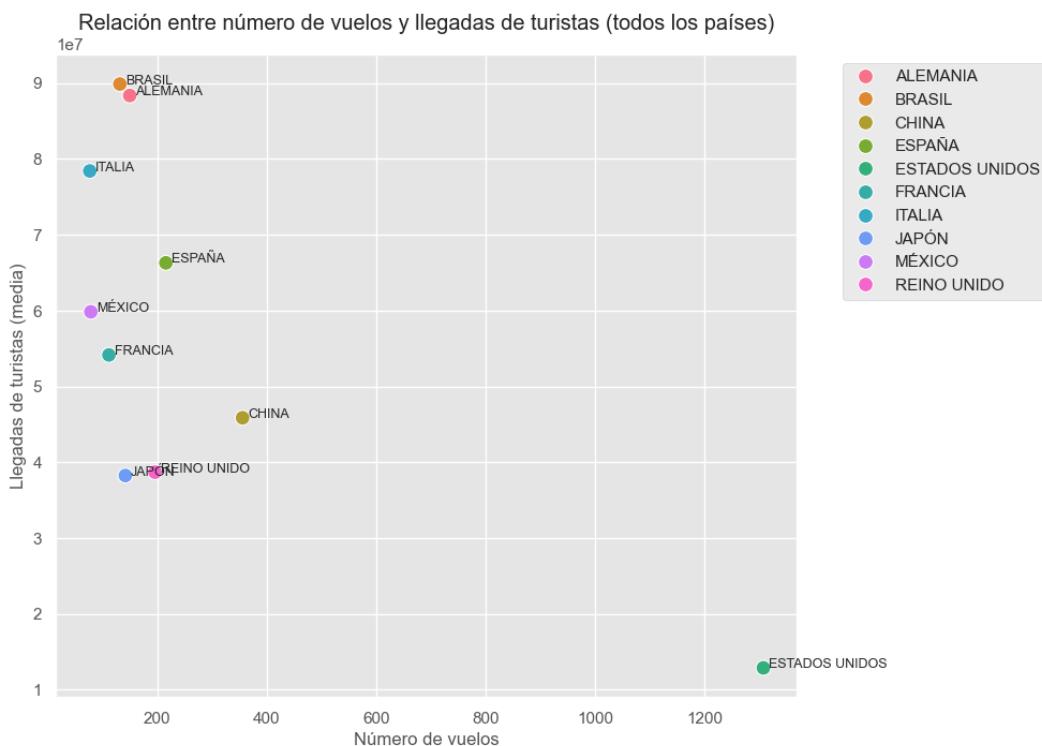


Figura 6.19: Relación entre número de vuelos y llegadas turísticas medias. Elaboración propia.

En la Figura 6.20, restringida a países con más de 50 millones de llegadas, se visualiza que

España destaca por un equilibrio sólido entre conectividad y llegada de visitantes, mientras que Italia y México logran ratios de eficiencia turística destacables con menos vuelos.

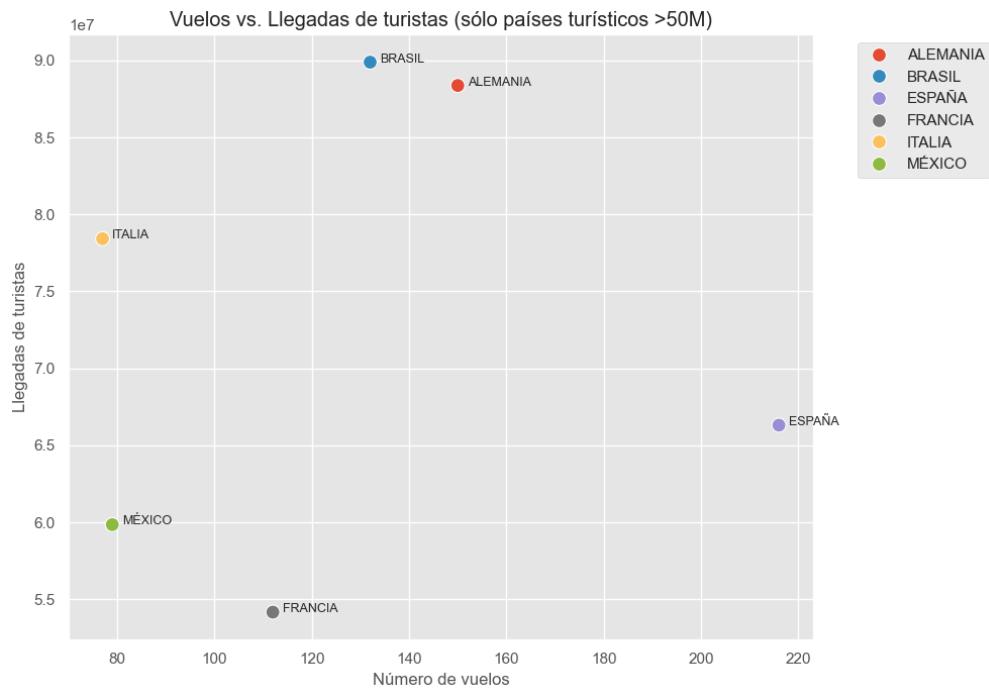


Figura 6.20: Relación vuelos–turismo para países con más de 50M de llegadas. Elaboración propia.

6.4.4 Duración media de vuelos y proyección internacional

La duración media de los vuelos por país puede usarse como proxy de internacionalización. La Figura 6.21 muestra que países como Uzbekistán, Argentina y Bangladés tienen vuelos más largos de media, lo que indica una dependencia de conexiones de largo alcance.

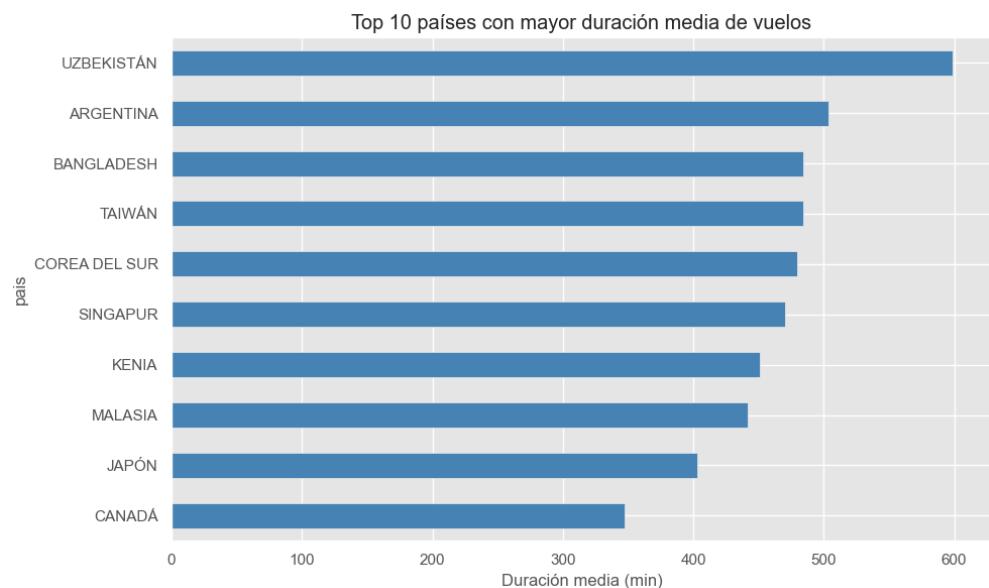


Figura 6.21: Top 10 países con mayor duración media de vuelos. Elaboración propia.

6.4.5 Ingresos y eficiencia en turismo aéreo

A nivel económico, los ingresos turísticos medios ofrecen un indicador directo del retorno. En la Figura 6.22, países como Italia, México y España destacan por su rentabilidad, mientras que otros con muchos vuelos, como EE.UU., presentan cifras inferiores en comparación.

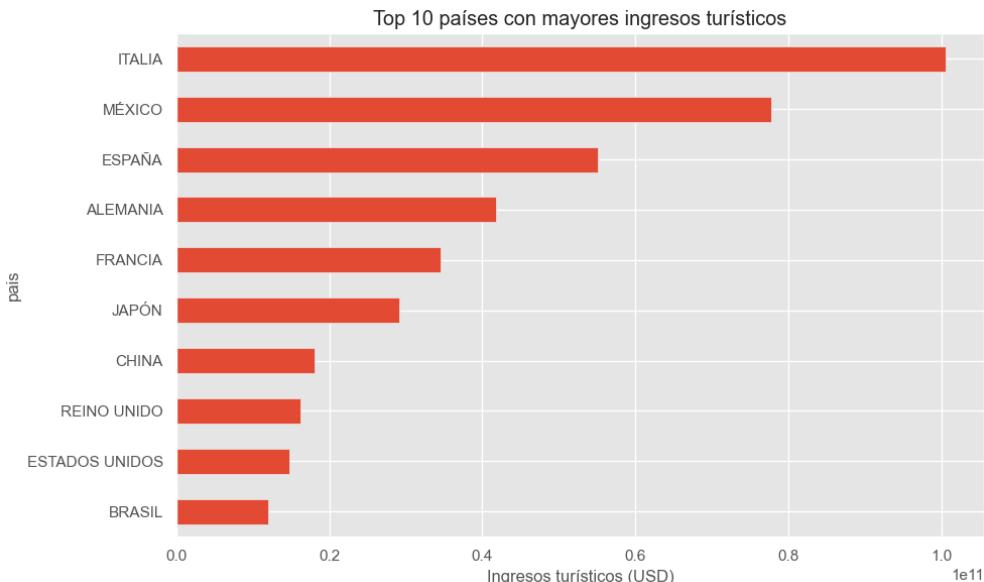


Figura 6.22: Top 10 países por ingresos turísticos medios. Elaboración propia.

Al combinar todas las variables en una tabla por país, se puede observar el equilibrio entre movilidad y retorno económico:

Cuadro 6.3: Indicadores agregados de eficiencia turística aérea.

País	Vuelos	Llegadas Medias	Ingresos Medios (USD)	Duración Media (min)
México	30	59.8M	77.8B	281.6
España	166	66.3M	55.1B	154.1
Alemania	124	88.4M	41.8B	159.7
EE.UU.	1115	12.9M	14.7B	282.9

6.4.6 Análisis de agrupamiento turístico-aeronáutico

Con el objetivo de identificar patrones de comportamiento entre países en cuanto a su eficiencia turística aérea —entendida como la capacidad de generar llegadas e ingresos turísticos en relación con su conectividad aérea— se aplicaron diferentes técnicas de clustering multivariable sobre los vectores construidos para cada país.

a) Clustering exploratorio con K-means y reducción dimensional (PCA / MDS)

En una primera fase, se realizó un agrupamiento de países utilizando el algoritmo K-means, tras reducción dimensional mediante PCA y MDS. Se analizaron diferentes valores de k, siendo k=4 el valor óptimo según la curva del codo.

Este enfoque permitió clasificar a los países en grupos homogéneos que responden a perfiles diferenciados:

Clúster 0: Países de alta eficiencia turística (pocos vuelos, muchas llegadas e ingresos).

Clúster 1: Países con conectividad intermedia y retorno turístico proporcional.

Clúster 2: Economías muy internacionalizadas con vuelos largos y retorno moderado.

Clúster 3: Países con fuerte conectividad pero bajo rendimiento turístico relativo.

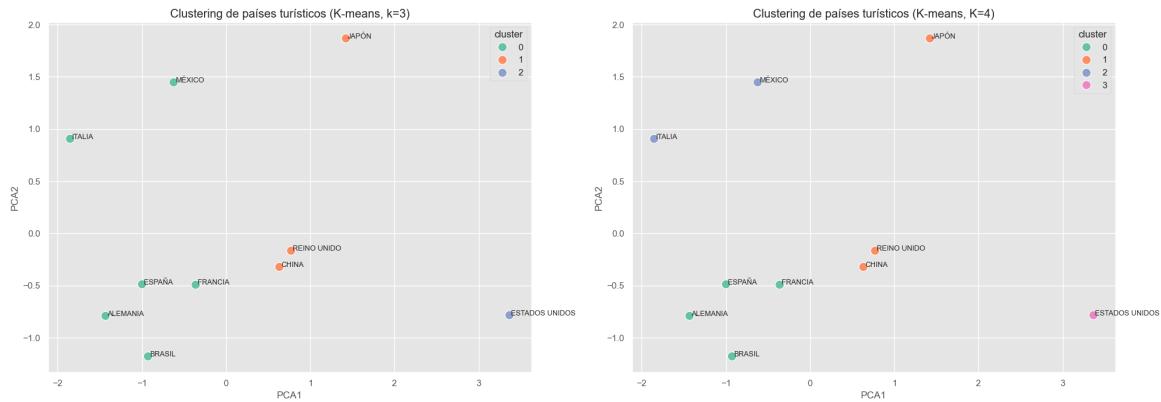


Figura 6.23: Clustering turístico-aeronáutico por PCA ($k=3$ y $k=4$).

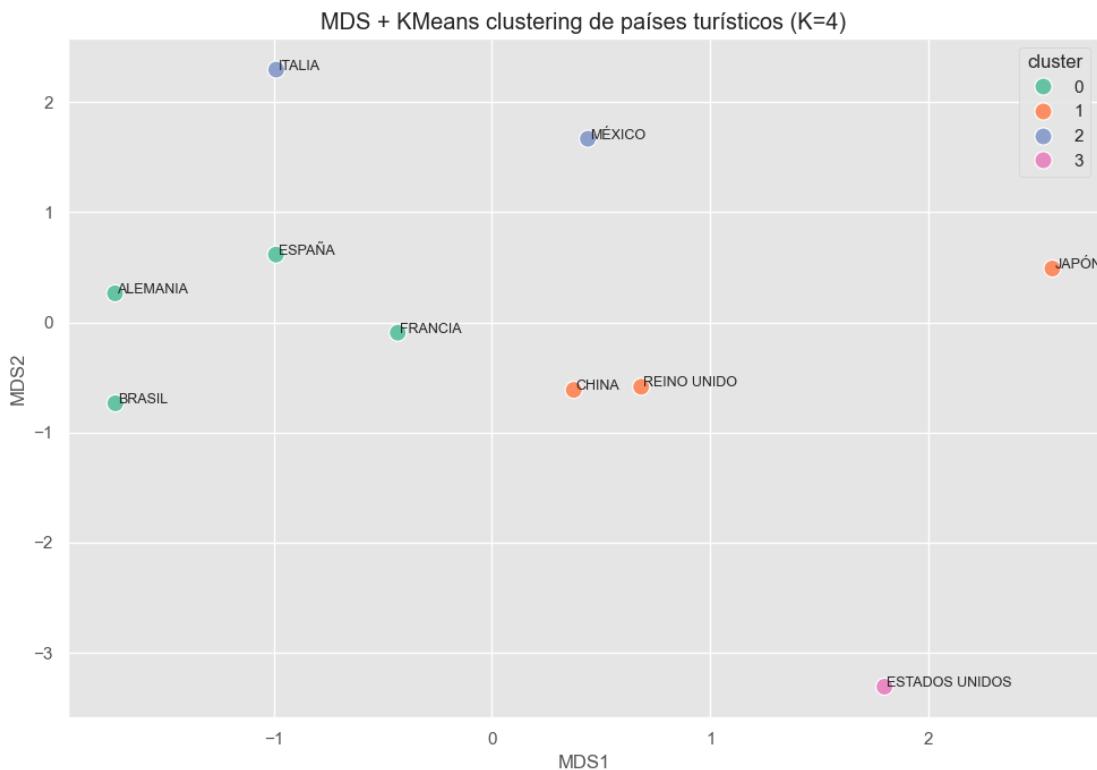


Figura 6.24: Proyección de clustering en MDS 2D ($k=4$).

Este análisis preliminar sirvió como base para validar la separación entre perfiles país y preparar la aplicación de algoritmos más avanzados.

b) Clustering jerárquico (Ward) Como segunda aproximación, se aplicó un modelo de clustering jerárquico utilizando el método de enlace Ward, basado en distancias euclidianas. El dendrograma obtenido (Figura 6.25) revela la existencia de dos grandes ramas.

- Una compuesta por México e Italia, con características comunes de eficiencia en rentabilidad turística pese a su conectividad moderada.
- Una rama más amplia con subdivisiones internas, que agrupa países como Alemania, Francia, Japón, China o EE.UU., con perfiles más heterogéneos en retorno, duración y número de vuelos.

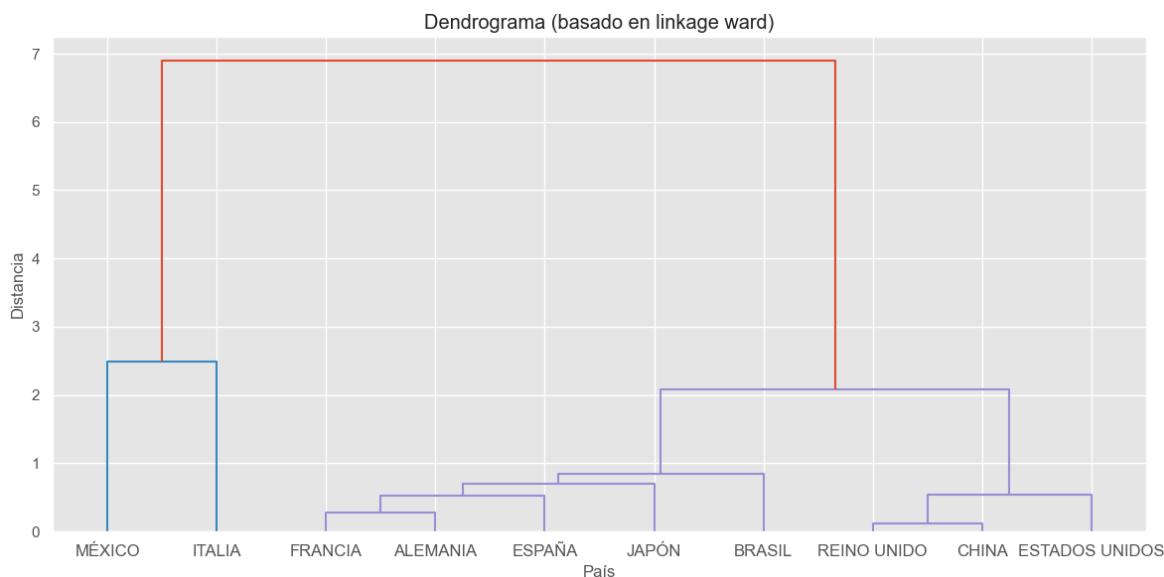


Figura 6.25: Dendrograma jerárquico de países turísticos (linkage Ward).

Este análisis jerárquico aporta una visión estructural de las similitudes inter-país, sin necesidad de prefijar el número de clústeres.

c) DBSCAN

Por último, se aplicó el algoritmo DBSCAN, orientado a la detección de clústeres con forma arbitraria y la identificación de outliers. El resultado, proyectado mediante MDS 2D (ver Figura 6.28), destaca a México e Italia como observaciones aisladas, lo que refuerza su carácter singular en términos de eficiencia turística aérea.

El resto de países, agrupados en un clúster denso, comparten un equilibrio más convencional entre conectividad, duración de vuelos y retorno económico.

6.5. ESTUDIO 5: ANÁLISIS COMBINADO VUELOS, TURISMO Y EMISIONES DE CO₂61

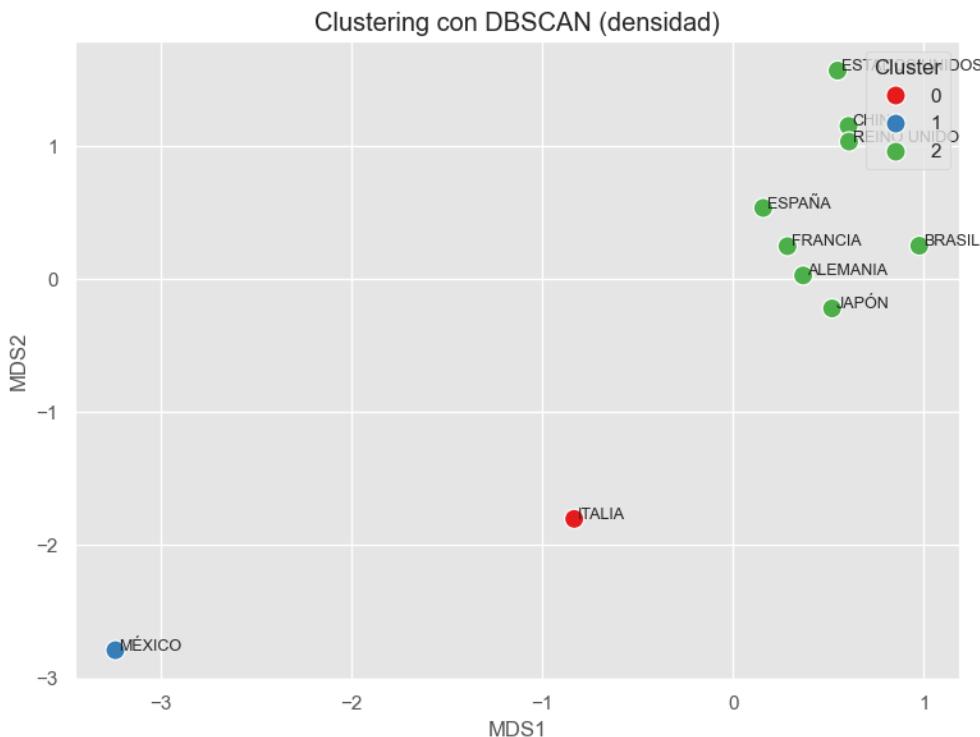


Figura 6.26: Clustering de países turísticos con DBSCAN proyectado en MDS.

6.4.7 Conclusión: eficiencia y dependencia turística desde el cielo

Este análisis ofrece una mirada sistémica sobre cómo los países convierten su conectividad aérea en valor turístico. Conclusiones clave:

- Existen casos altamente eficientes (México, Italia, Brasil).
- EE.UU. presenta sobredimensionamiento operativo respecto a retorno.
- Se pueden construir indicadores de sostenibilidad turística aérea.

Estos resultados fundamentan el Estudio 5, donde se integran las tres dimensiones —operativa, ecológica y turística— para evaluar la sostenibilidad turística aérea global.

6.5. Estudio 5: Análisis combinado Vuelos, Turismo y Emisiones de CO₂

6.5.1 Introducción: un enfoque integrado para la sostenibilidad turística aérea

Este estudio propone una visión holística al integrar operaciones de vuelo, eficiencia medioambiental y comportamiento turístico. Se pretende evaluar la sostenibilidad turística aérea de los países combinando:

- Conectividad aérea (número de vuelos, duración).

- Impacto ambiental (emisiones totales y medias de CO₂ por vuelo).
- Indicadores turísticos (llegadas internacionales, ingresos por turismo, dependencia del PIB).

Este enfoque permite caracterizar a los países no solo por su nivel de actividad, sino por cómo equilibran movilidad, impacto y retorno económico, ayudando a identificar rutas sostenibles y modelos replicables.

6.5.2 Datos cruzados: vuelos, emisiones y turismo por país

Se elaboró un dataset unificado por país con las siguientes variables:

Cuadro 6.4: Variables cruzadas por país.

Variable	Significado
n_vuelos	Total de vuelos registrados
co2_total_kg	Emisiones totales de CO ₂
media_emisiones	Emisiones medias por vuelo
llegadas_turistas	Media anual de llegadas internacionales
ingresos_turisticos_usd	Ingresos turísticos medios anuales (USD)
co2_por_millon_usd	CO ₂ por millón de USD generado

Los datos revelaron una gran variabilidad entre países. Por ejemplo, España presenta más vuelos y emisiones totales que EE.UU., pero también muchos más turistas, lo que reduce su ratio de CO por retorno económico:

Cuadro 6.5: Comparativa de eficiencia turística-aérea por país.

País	Vuelos	Emisiones (kg)	Llegadas (M)	Ingresos (USD)	CO ₂ / M USD
España	60	319,513	66.3	5.5e+10	0.0965
Alemania	31	138,387	88.3	4.2e+10	0.1067
China	53	512,225	45.9	1.8e+10	0.4347
EE.UU.	43	489,593	12.8	1.5e+10	0.7293
Brasil	42	263,281	89.9	1.2e+10	0.5193

6.5.3 Correlaciones entre variables: tensiones y sinergias

El análisis de correlación (Figura 6.27) reveló que las emisiones medias están negativamente correlacionadas con los indicadores turísticos, lo que sugiere que los países más turísticos suelen tener emisiones más controladas por unidad de ingreso o visitante.

6.5. ESTUDIO 5: ANÁLISIS COMBINADO VUELOS, TURISMO Y EMISIONES DE CO₂63

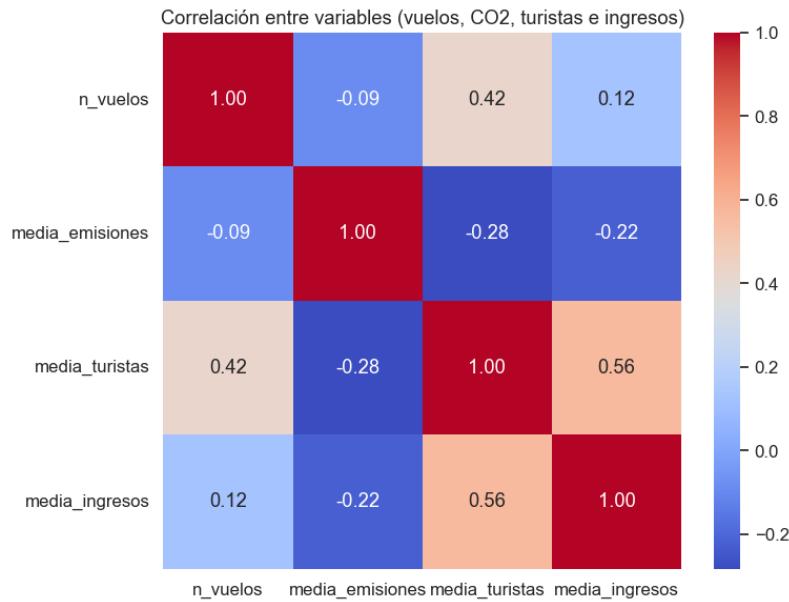


Figura 6.27: Matriz de correlación entre vuelos, emisiones, turistas e ingresos.

Además, se observó que el número de vuelos guarda una relación moderada con el turismo y baja con las emisiones, lo que respalda la necesidad de considerar eficiencia y no volumen.

Complementariamente, el gráfico de dispersión múltiple (Figura 6.30) muestra relaciones no lineales y evidencia la heterogeneidad en los perfiles nacionales.

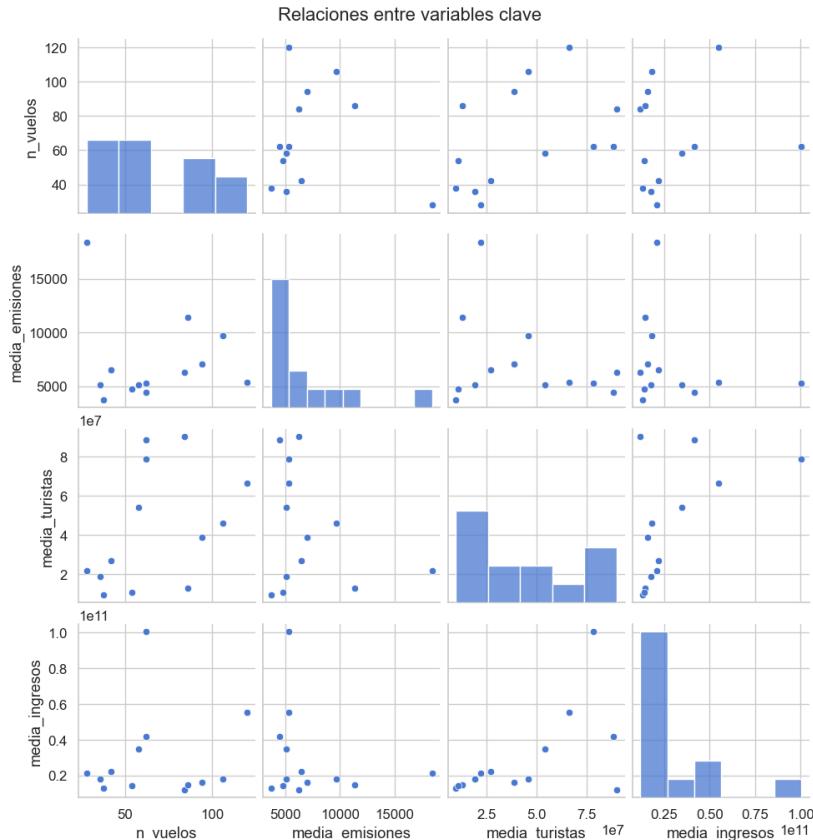


Figura 6.28: Relaciones entre variables clave. Elaboración propia.

6.5.4 Clasificación de países según perfil turístico-sostenible

Se construyó un clustering sobre las variables normalizadas `n_vuelos`, `media_emisiones`, `media_ingresos` y `co2_por_millon_usd` para categorizar los países según su perfil turístico-ambiental.

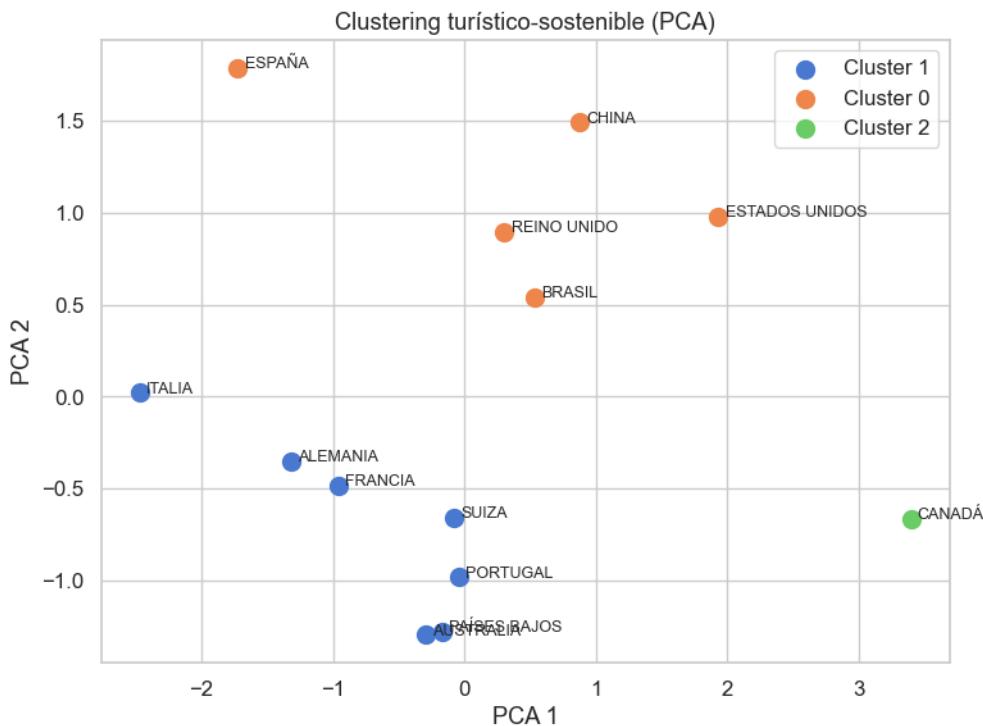


Figura 6.29: Clustering turístico-sostenible por PCA. Elaboración propia.

Clústeres obtenidos:

- **Cluster 0:** países con vuelos intensivos, ingresos medios y alta huella de carbono (EE.UU., China, Reino Unido).
- **Cluster 1:** países equilibrados con buena conversión entre emisiones y retorno (Alemania, Francia, Portugal, Suiza).
- **Cluster 2:** casos de eficiencia extrema con bajo impacto y alto ingreso (Canadá, Italia).

6.5.5 Comparativa radar: perfiles país de sostenibilidad aérea

Se sintetizaron los perfiles clave mediante un gráfico de radar con los países más representativos. Este enfoque permite observar:

EE.UU. como el país con mayor número de vuelos y emisiones por ingreso generado.

Italia como el más eficiente: menor CO₂ por ingreso turístico, alta rentabilidad.

España con una relación balanceada entre volumen y sostenibilidad.

6.5. ESTUDIO 5: ANÁLISIS COMBINADO VUELOS, TURISMO Y EMISIONES DE CO₂65

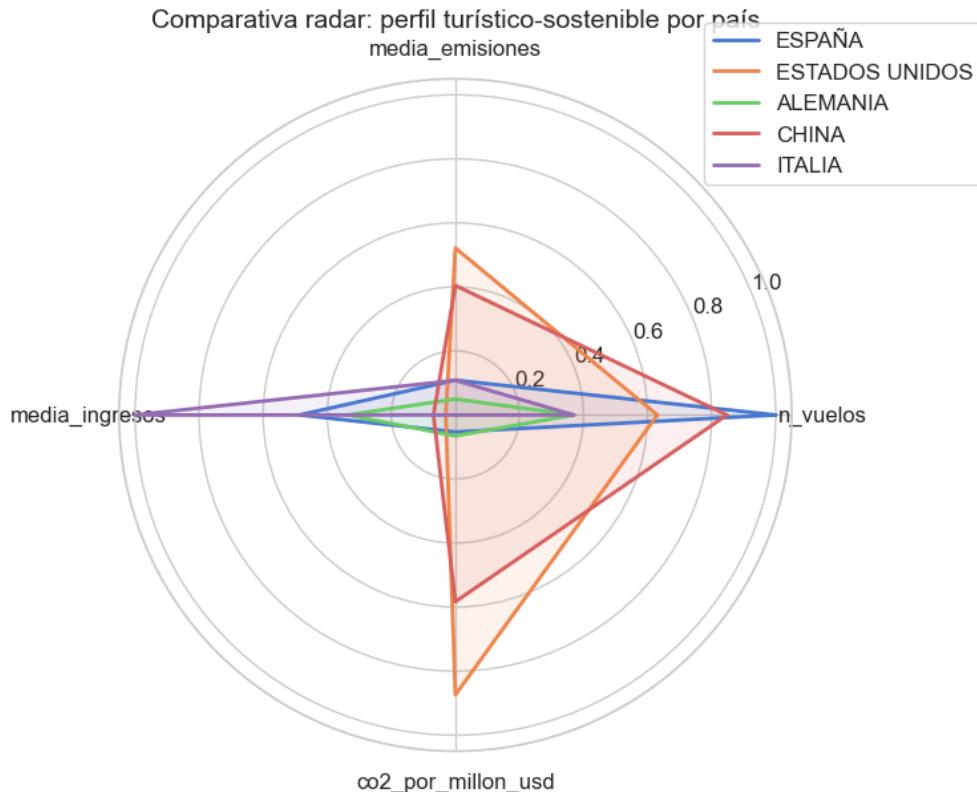


Figura 6.30: Radar comparativo de perfiles turístico-sostenibles por país.

Este tipo de visualización es útil para benchmarking y análisis estratégico de rutas, incentivando modelos replicables basados en rendimiento y responsabilidad.

6.5.6 Conclusiones: un modelo de sostenibilidad turística aérea

Este análisis integrado permite formular las siguientes conclusiones:

- La sostenibilidad aérea no depende del volumen, sino del balance entre vuelos, emisiones y retorno económico.
- Países como Italia, Alemania o Portugal muestran que es posible combinar movilidad eficiente con bajo impacto.
- La métrica CO₂ por millón USD es una herramienta sólida para evaluar la eficiencia ambiental del turismo aéreo.

Capítulo 7

Algoritmos de clustering y detección de patrones

7.1. Justificación del enfoque

El análisis exploratorio desarrollado en los capítulos anteriores ha revelado la existencia de estructuras latentes en los datos relacionados con el tráfico aéreo comercial, las emisiones de CO₂ y los indicadores turísticos. No obstante, las observaciones realizadas hasta ahora han sido predominantemente descriptivas, fundamentadas en estadísticas agregadas, cruces de variables y visualizaciones bivariadas.

Para ir más allá del análisis superficial y descubrir patrones complejos o agrupaciones emergentes en los datos multidimensionales, resulta imprescindible aplicar técnicas de *clustering* no supervisado. Estos algoritmos permiten identificar subconjuntos homogéneos dentro de conjuntos de datos sin necesidad de etiquetas previas, lo cual es ideal para contextos donde se desconoce la estructura subyacente de los datos, como ocurre con las rutas aéreas globales o la segmentación de países por perfil turístico-operativo.

Fundamentos teóricos: clustering en ciencia de datos

En términos generales, los algoritmos de clustering buscan minimizar una función de disimilitud intra-clúster y maximizar la separación entre grupos. Dado un conjunto de datos $X = \{x_1, x_2, \dots, x_n\}$ con $x_i \in R^d$, el objetivo es encontrar una partición $C = \{C_1, C_2, \dots, C_k\}$ tal que:

$$\sum_{j=1}^k \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

se minimice, donde μ_j es el centroide del clúster C_j .

Este tipo de enfoque permite abstraer la complejidad dimensional de los datos y generar representaciones más interpretables para el diseño de políticas de optimización y sostenibilidad.

Objetivos del capítulo

El objetivo de este capítulo no es únicamente aplicar modelos de agrupamiento por motivos técnicos, sino hacerlo con una orientación estratégica para el proyecto. Específicamente, se pretende:

- Detectar agrupaciones de vuelos similares en cuanto a duración, intensidad de emisiones y origen/destino.
- Agrupar tipos de aeronaves por perfil de uso y eficiencia media.
- Clasificar países según su comportamiento turístico-sostenible, como ya se introdujo en el Estudio 5.
- Evaluar la robustez de los patrones encontrados a través de distintas técnicas: centroides (K-means), densidad (DBSCAN), jerarquía (Aglomerativo) y componentes principales (PCA, MDS).

Por qué clustering y no clasificación supervisada

Este enfoque no supervisado es especialmente adecuado por varias razones:

- **Ausencia de etiquetas previas:** no existen clases conocidas a priori para vuelos, aviones o países.
- **Exploración estructural:** se busca descubrir estructuras o relaciones internas, no predecir una variable específica.
- **Transferibilidad:** los clústeres pueden utilizarse posteriormente como etiquetas para sistemas de recomendación o modelos supervisados de predicción de rendimiento.

Selección de algoritmos

En los siguientes apartados se detallarán los algoritmos seleccionados, con sus respectivas motivaciones:

- **K-Means:** eficiente para agrupamientos basados en proximidad euclídea. Útil para datos normalizados y de forma convexa.
- **DBSCAN:** identifica agrupaciones arbitrarias y permite detectar outliers.
- **Clustering jerárquico (Ward):** visualización mediante dendrogramas, útil para observar niveles de similitud.
- **Birch:** explorados en fases intermedias del análisis de rutas, por su tolerancia a formas irregulares y datos ruidosos.

Cada técnica será aplicada sobre subconjuntos distintos del proyecto (vuelos, tipos de aeronave, países), adaptando la dimensionalidad mediante PCA o MDS para facilitar la visualización y comprensión.

Relación con los objetivos globales del TFG

Este capítulo actúa como nodo central de articulación en los análisis. Los clústeres obtenidos permitirán:

- Proponer rutas más eficientes dentro de un mismo perfil operativo.
- Rediseñar trayectorias según agrupaciones de emisiones y demanda.
- Integrar tipologías de país para priorizar rutas de menor impacto o mayor retorno.

En resumen, el *clustering* permite transformar una colección compleja de datos multifuente en un sistema estructurado de toma de decisiones, alineado con los objetivos de sostenibilidad, eficiencia y análisis comparativo del presente TFG.

7.2. Selección y configuración de algoritmos

La selección de algoritmos de agrupamiento no supervisado debe responder tanto a la naturaleza del conjunto de datos como a los objetivos analíticos específicos. En este proyecto se han trabajado tres grandes tipos de objetos:

- Vuelos individuales (con atributos como duración, región, emisiones, internacionalidad).
- Modelos de aeronave (usos, duración media, eficiencia, clustering).
- Países (con indicadores de vuelos, turismo y emisiones agregados).

Cada uno de estos conjuntos presenta estructuras y desafíos distintos. Por tanto, se ha optado por una combinación de algoritmos complementarios, capaces de ofrecer robustez analítica y riqueza interpretativa.

1. K-Means: agrupamiento por proximidad euclídea

El algoritmo K-Means se ha utilizado como método base en prácticamente todos los subconjuntos del proyecto. Sus ventajas son bien conocidas:

- Escalabilidad.
- Interpretación sencilla de los centroides.
- Buen rendimiento con datos normalizados.

La principal decisión metodológica en este caso fue la determinación del número óptimo de clústeres k . Para ello, se utilizaron dos enfoques:

- Curva del codo (elbow method): minimiza la inercia intra-clúster total.
- Coeficiente de Silueta: evalúa la coherencia interna de cada clúster.

Ejemplo: para el agrupamiento de tipos de avión, se optó por $k = 4$ tras evaluar la silueta y validar visualmente los grupos en proyección PCA.

2. DBSCAN: agrupamiento por densidad y detección de outliers

El algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise) ha sido utilizado para detectar:

- Vuelos atípicos en términos de emisiones extremas o duración anómala.
- Países outliers con perfiles únicos en sostenibilidad turística.

La principal configuración de este algoritmo radica en:

- ε : radio de vecindad.
- `min_samples`: número mínimo de puntos por vecindad.

Ambos parámetros fueron ajustados empíricamente utilizando mapas de dispersión y la visualización resultante de proyecciones MDS. Se lograron agrupaciones de alta calidad, con outliers significativos (como EE.UU., México o Italia en determinados análisis).

3. Clustering jerárquico (Agglomerative + linkage Ward)

El enfoque jerárquico se implementó principalmente sobre:

- Países (para comparar similitudes estructurales globales).
- Agrupaciones de rutas (para construir árboles de decisión por distancia y CO₂).

El método Ward linkage fue seleccionado por su capacidad para minimizar la varianza intra-grupo a cada paso de fusión. Se aplicó sobre matrices de distancia normalizadas, y el resultado fue representado mediante dendrogramas, que facilitaron una interpretación progresiva y clara de los niveles de agrupamiento.

4. Algoritmos complementarios: Birch, Spectral y Mean Shift

Estos métodos se utilizaron en fases exploratorias con los siguientes propósitos:

- **Birch**: probar escalabilidad sobre datasets de vuelos simulados y clustering preliminar de aeropuertos.
- **Spectral clustering**: estudiar agrupaciones no convexas en entornos multiconectados (p. ej., vuelos con hub europeo común).
- **Mean Shift**: sin suponer el número de clústeres, útil para identificar núcleos naturales en mapas bidimensionales.

Aunque no fueron seleccionados como métodos finales, sus resultados respaldaron algunas configuraciones de K-Means y DBSCAN.

Preprocesamiento previo: normalización y reducción de dimensionalidad

Antes de aplicar cualquier algoritmo, todos los conjuntos de datos fueron sometidos a:

- **Normalización por min-max scaling** para evitar que variables como duración o emisiones dominaran las distancias.
- **Reducción de dimensionalidad** para facilitar visualizaciones:
 - PCA (Análisis de Componentes Principales) en datos con estructura lineal.
 - MDS (Escalado Multidimensional) en matrices de distancia.

Estas técnicas ayudaron no solo a mejorar la eficiencia computacional, sino también a interpretar los resultados de clustering a través de gráficos de dispersión, mapas de calor o figuras radar.

Síntesis del enfoque aplicado

Cuadro 7.1: Síntesis de algoritmos por conjunto de análisis.

Conjunto	Algoritmo principal	Complementarios	Proyección
Vuelos comerciales	K-Means (k=6)	DBSCAN, Spectral	MDS, PCA
Modelos de aeronave	K-Means (k=4)	Birch, Agglomerative	PCA
Países (turismo/CO ₂)	K-Means (k=3/4)	DBSCAN, Jerárquico	PCA, Radar, Dendro

Esta diversidad metodológica ha permitido abordar la complejidad del problema desde distintos ángulos y extraer conclusiones robustas y útiles.

7.3. Resultados y análisis de agrupaciones

Una vez aplicados los distintos algoritmos de clustering y técnicas de proyección, se procedió a interpretar los resultados obtenidos desde un enfoque estructural y comparativo. Esta sección sintetiza los principales hallazgos según el objeto analizado (vuelos, aeronaves, países), prestando especial atención a la coherencia interna de los clústeres y a su utilidad interpretativa.

7.3.1 Agrupaciones de vuelos comerciales

A partir de una muestra representativa de vuelos comerciales reales (más de 5.000 observaciones), se aplicaron técnicas de clustering para detectar tipologías operativas. Las variables consideradas incluyeron:

- Duración del vuelo (exttduracion_min)
- Hora de salida (extthora_salida_num)
- Día de la semana (exttdia_semana)

- Delta horario (extttdelta_utc)
- Si el vuelo es internacional o no

El análisis con K-Means ($k = 6$) arrojó resultados con buena separación (Coef. Silueta ≈ 0.35). Se identificaron clústeres claramente diferenciados:

- Vuelos cortos diurnos domésticos (alta frecuencia, duración < 2h)
- Vuelos nocturnos intercontinentales (larga duración, internacional)
- Casos atípicos con duración negativa (errores o valores extremos)

Estos últimos fueron cruciales para la depuración de datos y validación posterior.

Visualizaciones relacionadas: MDS de clustering de vuelos, boxplots de duración por clúster, dendrograma jerárquico.

7.3.2 Clustering de tipos de aeronaves

En el análisis de más de 2.800 vuelos clasificados por tipo de aeronave, se utilizaron atributos como:

- Duración media de vuelo por modelo
- Número total de vuelos
- Participación relativa por tipo

Aplicando K-Means ($k = 4$) y proyectando en 2D mediante PCA, se obtuvieron grupos coherentes:

- Clúster de aeronaves regionales con duración < 90 min
- Aviones de medio radio como A320 o B737 (segmento central)
- Aeronaves de largo recorrido (B777, A350, B741)
- Jets ejecutivos o vuelos atípicos con baja frecuencia

La proyección en espacio PCA confirmó la distribución esperada por volumen y rango.

Visualizaciones asociadas: PCA clustering de aeronaves, gráfico de boxplots por clúster y gráfico de recuento total de vuelos por grupo.

7.3.3 Clasificación de países según perfil turístico-operativo

Este fue uno de los análisis más ricos e interpretativos del proyecto. A partir de un dataset cruzado con datos de vuelos, emisiones y turismo, se aplicaron distintos modelos de agrupamiento a nivel país:

- K-Means ($k = 4$): reveló agrupaciones según eficiencia turística, volumen de vuelos y ratio de emisiones.

- Clustering jerárquico (Ward): visualizado en dendrogramas, permitió observar la similitud estructural entre pares de países.
- DBSCAN: detectó outliers relevantes como México o Italia.
- Radar chart: mostró el perfil multidimensional de países clave.

Los clústeres resultantes reflejaron patrones consistentes:

Cuadro 7.2: Clústeres por características turístico-operativas.

Clúster	Características principales
C0	Alta conectividad y alto CO ₂ por ingreso (EE.UU., China, Reino Unido)
C1	Equilibrio entre vuelos, ingresos y huella (Alemania, Francia, Portugal, Suiza)
C2	Alta eficiencia turística-sostenible (Italia, Países Bajos, Australia, Canadá)
Outliers	Países que rompen el patrón y requieren análisis individual (México, Japón)

Visualizaciones clave: radar comparativo, PCA clustering de sostenibilidad, matriz de correlaciones.

7.3.4 Visualización sintética: flujos internacionales de conexión

Como complemento final, se elaboró un diagrama Sankey circular que representa los flujos internacionales de vuelos entre países del estudio. Este gráfico:

- Está organizado siguiendo el orden longitudinal horario del globo (de este a oeste).
- Muestra la intensidad de conexión entre pares de países.
- Reafirma visualmente la existencia de hubs de conexión aérea global (EE.UU., Reino Unido, Alemania, Japón).

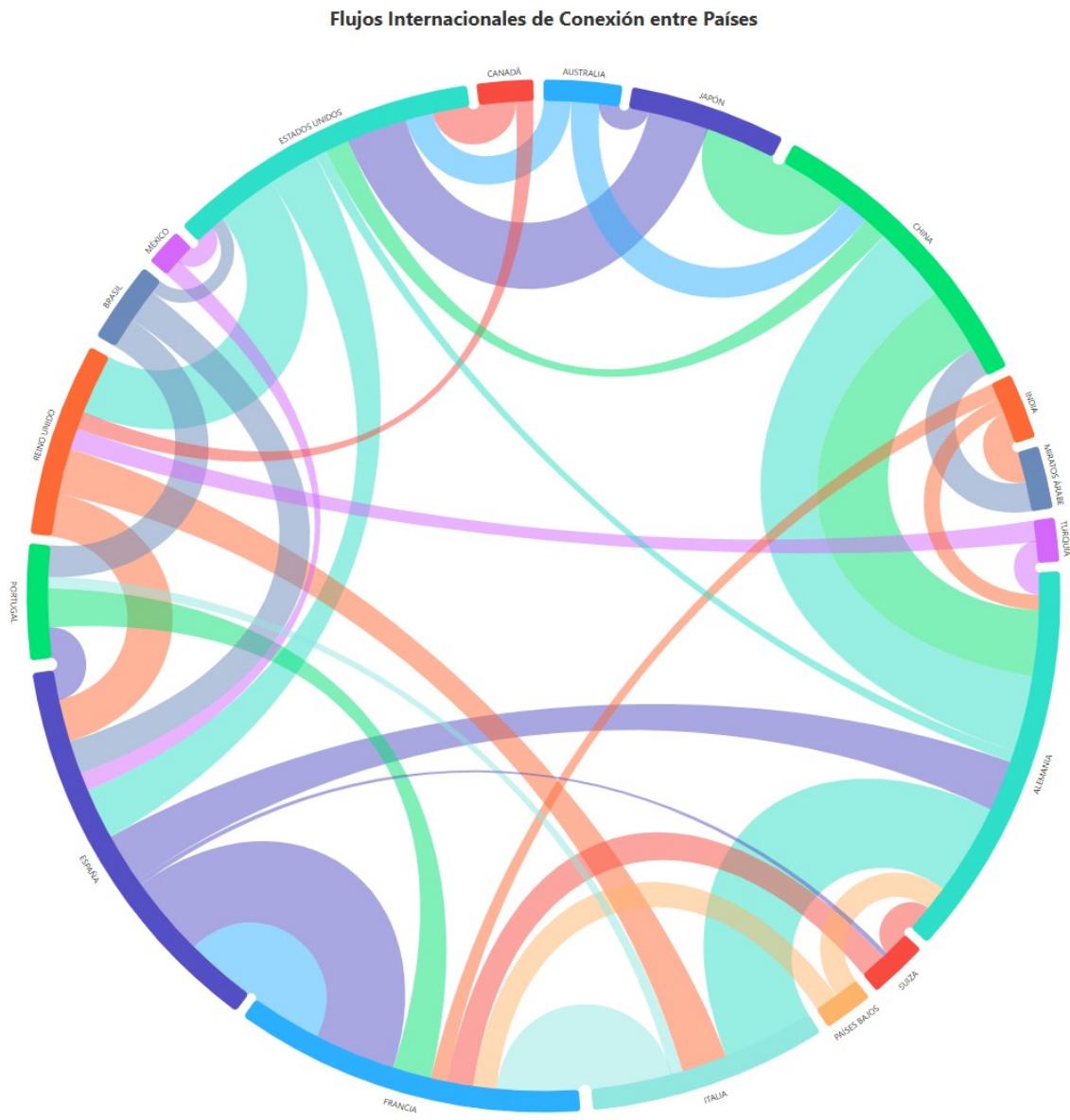


Figura 7.1: Diagrama circular de flujos internacionales entre países (ordenado por huso horario).

Esta representación aporta una dimensión espaciotemporal al análisis, destacando no solo la frecuencia de conexión, sino también la posicionalidad geoestratégica de ciertos países dentro de las redes globales.

Síntesis de hallazgos clave

- Los algoritmos aplicados han permitido validar de forma robusta las tipologías de vuelos, aviones y perfiles nacionales.
 - Se han detectado casos extremos y clústeres con sentido operacional y económico.
 - Las visualizaciones multivariantes y los modelos no supervisados ofrecen una base sólida para la optimización posterior de rutas y la planificación de estrategias de sostenibilidad.

Este capítulo demuestra que el análisis no supervisado, bien aplicado, es capaz de extraer conocimiento estructural de datasets complejos y reales, abriendo paso al siguiente objetivo: la optimización de rutas aéreas desde una perspectiva técnica y ambiental.

Capítulo 8

Discusión crítica y evaluación

8.1. Validación de resultados

La validación de un análisis exploratorio y predictivo como el desarrollado en este Trabajo de Fin de Grado no puede abordarse desde una lógica única o puramente cuantitativa. En ausencia de un modelo supervisado con etiquetas de referencia, la validación se basa en tres pilares complementarios:

1. Coherencia interna de los clústeres obtenidos

Los algoritmos de clustering aplicados sobre diferentes dimensiones (vuelos, aeronaves, países) han mostrado consistencia y lógica interpretativa en los agrupamientos obtenidos. Por ejemplo:

- En el análisis de vuelos, los clústeres separaban claramente trayectos regionales de rutas intercontinentales, en función de duración y diferencia horaria.
- En los tipos de avión, el modelo distinguió correctamente aeronaves de corto, medio y largo radio, reflejando su uso operativo real.
- En el caso de países, los grupos generados mediante K-Means, PCA y MDS agrupaban por eficiencia turística-aeronáutica, diferenciando países como Italia o Canadá (alta rentabilidad y bajo impacto) de otros como EE.UU. o China (alta intensidad operativa y ambiental).

Este nivel de coherencia respalda la validez estructural de los agrupamientos obtenidos, incluso sin etiquetas externas.

2. Convergencia entre distintas técnicas

En múltiples casos, los resultados generados por diferentes algoritmos convergieron o se reforzaron mutuamente:

- Los clústeres detectados por K-Means en países fueron consistentes con las estructuras observadas en el dendrograma jerárquico.

- Los outliers detectados por DBSCAN (como México o Italia) coincidieron con puntos extremos en los gráficos radar y en la distribución del ratio CO₂ por ingreso.
- Las proyecciones 2D mediante PCA y MDS revelaron separaciones similares, confirmando la solidez de la segmentación.

Esta triangulación de métodos aporta robustez a los hallazgos, y reduce el riesgo de que los resultados sean artefactos de una técnica concreta.

3. Validación por interpretación cualitativa

Además de la validación técnica, los resultados fueron contrastados desde una perspectiva contextual basada en el conocimiento del sector aeronáutico y turístico:

- La relación negativa entre emisiones medias y llegadas de turistas valida la hipótesis de que la eficiencia ambiental no está necesariamente reñida con la rentabilidad turística.
- La identificación de países con alto impacto por millón de USD generado (como China o EE.UU.) coincide con las alertas internacionales sobre sostenibilidad y huella ecológica del turismo de larga distancia.
- El grafo circular de flujos internacionales aportó una capa de interpretación espacial coherente con los hubs observados: Europa Occidental, Asia Oriental y Norteamérica como ejes dominantes.

En conjunto, los resultados del análisis muestran coherencia técnica, sentido contextual y solidez estructural, lo que valida su utilidad tanto desde una perspectiva académica como como base para la toma de decisiones informada en entornos reales.

8.2. Limitaciones y posibles sesgos

Pese al rigor aplicado en todas las fases del proyecto, es importante reconocer que existen diversas limitaciones inherentes al diseño del estudio, a la disponibilidad de datos y a las técnicas empleadas. Estas limitaciones no invalidan los resultados, pero deben ser tenidas en cuenta a la hora de interpretar las conclusiones o de extrapolar los hallazgos a entornos operativos reales.

1. Limitaciones en la calidad y cobertura de los datos

Buena parte de los datos utilizados en este TFG provienen de fuentes abiertas y no estructuradas, extraídas mediante técnicas de web scraping desde plataformas como FlightAware, OpenFlights y OpenSky Network. Aunque se aplicaron filtros y procesos de limpieza cuidadosos, se identifican varios condicionantes:

- Datos incompletos o inconsistentes: no todos los vuelos contaban con información completa (horarios, duración, zonas horarias), lo que obligó a eliminar registros.
- Desfase o asincronía: no todos los datasets estaban actualizados al mismo año o periodo, especialmente en las fuentes turísticas del Banco Mundial.

- Sesgo geográfico: hubo mayor densidad de datos disponibles en regiones como Europa y EE.UU., mientras que algunas zonas de África o Asia presentaron baja cobertura.

Estas cuestiones introducen asimetrías en la representatividad del análisis y pueden afectar a la solidez comparativa entre países.

2. Supuestos necesarios para la estimación de emisiones

El cálculo de emisiones de CO₂ se realizó a partir de modelos simplificados basados en:

- Tipo de aeronave.
- Duración del vuelo.
- Fórmulas empíricas adaptadas desde bases como ICAO.

Si bien esto permite obtener valores aproximados, se asumen importantes supuestos simplificadores:

- No se tuvo en cuenta la ocupación real del vuelo ni la carga útil.
- No se distinguieron fases del vuelo (despegue, crucero, aterrizaje) ni condiciones atmosféricas.
- Se aplicó un modelo lineal de emisiones por minuto, sin contemplar no linealidades propias del consumo real.

En consecuencia, las cifras deben interpretarse como estimaciones relativas, útiles para análisis comparativos, pero no como valores absolutos certificados.

3. Suposiciones analíticas en el clustering

Aunque se aplicaron múltiples algoritmos de clustering y reducciones de dimensionalidad (PCA, MDS), estos métodos suponen:

- La existencia de estructuras separables, lo cual no siempre es garantizable.
- La necesidad de elegir arbitrariamente valores de k en algunos modelos.
- La sensibilidad a outliers o escalas mal normalizadas (algo mitigado con min-max scaling).

Además, el clustering no supervisado no ofrece una medida directa de “exactitud”, lo que implica que las agrupaciones son válidas en función de su coherencia interna y capacidad explicativa, pero no verificables de forma tradicional.

4. Métricas simplificadas de eficiencia o sostenibilidad

Conceptos clave como “eficiencia turística aérea” o “sostenibilidad por vuelo” se han formalizado mediante indicadores sintéticos (como CO₂ por millón USD generado), pero esto conlleva simplificaciones:

- No se ha modelado la cadena completa de valor del turismo (alojamiento, servicios, etc.).
- No se ha considerado el impacto social, territorial o cultural del flujo turístico.
- El enfoque ambiental se ha centrado en emisiones directas, sin contemplar otras externalidades.

Esto implica que las métricas usadas son útiles como *proxy* o índice comparativo, pero no capturan la complejidad plena del fenómeno turístico ni su sostenibilidad integral.

5. Limitación temporal: análisis estático

Finalmente, este estudio tiene una naturaleza fundamentalmente transversal y estática. No se ha incorporado análisis de series temporales, tendencias ni variaciones estacionales. Esto puede ser especialmente relevante en contextos como:

- Cambios post-pandemia (COVID-19)
- Incremento del turismo doméstico vs. internacional
- Evolución de las normativas de aviación y sostenibilidad

El modelo, por tanto, ofrece una fotografía en alta resolución de una situación agregada, pero no una proyección dinámica de su evolución.

Síntesis crítica

En conjunto, estas limitaciones son propias de cualquier estudio de carácter exploratorio y multidisciplinar. Lejos de invalidar el trabajo, su identificación y comunicación explícita refuerzan la transparencia, honestidad científica y madurez metodológica del proyecto.

Estas limitaciones también señalan posibles líneas futuras de mejora que se detallarán en el Capítulo 9.

8.3. Relevancia práctica y social

Uno de los principales valores de este Trabajo de Fin de Grado reside en su capacidad para trasladar técnicas avanzadas de Ciencia de Datos a problemas reales de carácter social, económico y medioambiental, como es la planificación y evaluación de rutas aéreas comerciales. A lo largo del proyecto se ha demostrado que la disponibilidad de datos abiertos, combinada con herramientas analíticas rigurosas, permite obtener insights estratégicos relevantes tanto para el sector del transporte como para las políticas públicas de turismo sostenible.

1. Aplicabilidad al sector aeronáutico

Desde el punto de vista del transporte aéreo, el proyecto permite:

- Identificar rutas operativamente ineficientes que podrían ser objeto de revisión o reconfiguración (por baja ocupación, alto impacto o redundancia).
- Comparar perfiles de aeronaves según su rendimiento medio en diferentes trayectos, con utilidad para decisiones de flota.
- Detectar hubs de conexión subóptimos o excesivamente contaminantes, facilitando estrategias de desvío o escalamiento inteligente.

Todo ello puede ser incorporado por aerolíneas, gestores aeroportuarios o entidades reguladoras como la Organización de Aviación Civil Internacional (OACI) en sus modelos de simulación o análisis de red.

2. Contribución a la sostenibilidad y transición verde

Uno de los focos centrales del TFG ha sido la incorporación de variables ambientales, específicamente las emisiones de CO₂. A partir del análisis desarrollado, se ha demostrado que:

- No siempre existe correlación positiva entre número de vuelos y rentabilidad turística.
- Algunos países presentan elevadas emisiones con bajo retorno económico, lo que puede cuestionar su modelo de movilidad aérea internacional.
- Se pueden identificar perfiles sostenibles de país, que maximizan ingresos o llegadas minimizando el impacto medioambiental.

Este tipo de información resulta clave para avanzar hacia modelos de turismo sostenible y descarbonización del transporte, alineados con los Objetivos de Desarrollo Sostenible (ODS), especialmente el ODS 13 (Acción por el clima) y el ODS 12 (Producción y consumo responsables).

3. Valor para el análisis estratégico y geopolítico

El enfoque georreferenciado del proyecto, reforzado por visualizaciones como el diagrama circular de flujos internacionales, aporta una dimensión espacial útil para:

- Analizar la centralidad o dependencia de ciertos países en redes aéreas globales.
- Detectar regiones sobrerepresentadas en emisiones o infrarrepresentadas en retorno turístico.
- Ofrecer soporte visual y analítico para negociaciones internacionales, subsidios a rutas, o rediseños de conectividad.

Estas perspectivas pueden ser útiles tanto para organismos multilaterales (como Eurocontrol, ICAO o WTTC), como para entidades nacionales de planificación o diplomacia económica.

4. Relevancia académica y multidisciplinar

Desde un enfoque universitario, el proyecto pone en valor:

- La capacidad de la Ciencia de Datos como herramienta transversal, aplicable a retos complejos que combinan tecnología, economía y sostenibilidad.
- El uso de fuentes abiertas para crear modelos analíticos sin depender de bases de datos privativas o institucionales.
- La combinación de competencias de Ingeniería Informática (automatización, scraping, clustering) y Administración de Empresas (eficiencia, análisis coste-beneficio), reflejando una formación integral y aplicada.

Este enfoque puede inspirar futuras líneas de investigación y proyectos de estudiantes interesados en aplicar tecnología a la toma de decisiones con impacto real.

Conclusión del capítulo

Este proyecto no se limita a un ejercicio de modelado técnico, sino que ha demostrado una clara orientación a resolver problemas complejos del mundo real. Sus resultados son aplicables, sus métodos replicables y su enfoque alineado con los desafíos actuales del transporte, el medio ambiente y el turismo global.

La discusión crítica desarrollada en este capítulo consolida el valor del trabajo realizado y sienta las bases para su ampliación futura, que será abordada en el siguiente y último capítulo del documento.

Capítulo 9

Conclusiones y líneas futuras

9.1. Conclusiones principales

El presente Trabajo de Fin de Grado ha demostrado la viabilidad y el valor de aplicar técnicas de Ciencia de Datos al análisis multidimensional de rutas aéreas comerciales. A través de una combinación de scraping, integración de fuentes abiertas, modelado geoespacial y algoritmos de aprendizaje no supervisado, se han logrado extraer patrones y relaciones que trascienden el simple conteo de vuelos.

Entre las principales conclusiones, destacan:

- Los datos abiertos son suficientes y relevantes para modelar de forma detallada las operaciones de transporte aéreo, emisiones asociadas y retornos turísticos, incluso sin acceso a datos internos de aerolíneas o autoridades.
- Es posible caracterizar eficiencia operativa y sostenibilidad ambiental en base a métricas como la duración promedio por modelo de aeronave, la relación CO₂ por pasajero, o el retorno turístico por vuelo.
- El uso de técnicas de clustering (K-Means, DBSCAN, Birch, jerárquicos) ha permitido segmentar rutas, tipos de aeronave y países en grupos representativos según su perfil operativo y su impacto ambiental/turístico, generando conocimiento estructurado.
- El análisis geopolítico y económico, combinado con visualizaciones como diagramas de Sankey, proyecciones MDS y mapas de conectividad, ha ofrecido una dimensión analítica integradora, clave para la toma de decisiones.
- En conjunto, se ha construido una metodología reproducible, con potencial aplicación real y escalabilidad hacia entornos más complejos o específicos.

9.2. Propuestas de mejora

A pesar de los resultados satisfactorios, se identifican diversas áreas susceptibles de mejora:

- **Ampliación de horizontes temporales:** Incluir datos históricos a lo largo de varios años permitiría observar tendencias y evaluar el impacto de eventos como pandemias, conflictos o cambios legislativos.

- **Incorporación de datos de ocupación real:** La inclusión de información sobre número de pasajeros por vuelo permitiría ajustar las emisiones por pasajero y mejorar las métricas de eficiencia medioambiental.
- **Normalización de zonas horarias y corrección de errores:** Algunos vuelos presentan duraciones negativas debido a errores en la alineación horaria. Se podría implementar una normalización automática mediante bases de datos de IATA o TZ World.
- **Evaluación de escalas y vuelos conectados:** Actualmente, el análisis es punto a punto. Incorporar trayectos completos con escalas permitiría una visión más holística de las rutas y sus impactos.
- **Uso de modelos predictivos más avanzados:** Si bien se han probado regresores clásicos (Random Forest, Gradient Boosting), incorporar redes neuronales o técnicas de aprendizaje profundo podría mejorar la precisión en la estimación de emisiones y tiempos de vuelo.

9.3. Aplicaciones y posibles extensiones

Este trabajo constituye una base robusta para futuras investigaciones o desarrollos tecnológicos, tanto en el ámbito académico como en el sector profesional. Algunas líneas concretas incluyen:

- **Diseño de dashboards interactivos:** Con herramientas como Dash, Streamlit o Power BI, se podrían construir interfaces para que usuarios del sector consulten en tiempo real el rendimiento de rutas, aerolíneas o países.
- **Sistemas de ayuda a la decisión (DSS)** para organismos de aviación civil, donde se combinen modelos de optimización y análisis de impacto para rediseñar rutas o evaluar autorizaciones de vuelo.
- **Modelado de eficiencia turística a nivel regional:** expandiendo el enfoque país-país a regiones o ciudades receptoras, lo cual sería de gran valor para destinos turísticos en auge.
- **Análisis de resiliencia aérea:** mediante simulaciones de cancelaciones masivas, bloqueos de espacio aéreo o restricciones de combustible, se podría evaluar la robustez de las redes aéreas.
- **Extensión al transporte intermodal:** incorporando conexiones ferroviarias o marítimas, lo cual permitiría optimizar rutas desde una visión de movilidad sostenible integral.

Reflexión final

Este TFG no solo ha sido una oportunidad para aplicar conocimientos técnicos y metodológicos, sino también para tomar conciencia del potencial transformador de los datos cuando se orientan hacia retos reales, como la sostenibilidad del transporte, la eficiencia operativa y la gestión turística.

La intersección entre ingeniería, economía y medio ambiente constituye un campo fértil, donde proyectos como este pueden contribuir activamente a la mejora de sistemas globales, y donde el perfil profesional del estudiante se proyecta hacia una visión integral, aplicada y comprometida con el futuro.

Bibliografía

- [1] Agencia Estatal de Seguridad Aérea (AESA). Limitaciones de tiempo de vuelo y actividad de tripulaciones. <https://www.seguridadaerea.gob.es/es/ambitos/operaciones-aereas/limitaciones-de-tiempo-de-vuelo-y-actividad-de-tripulaciones>, 2025. Consultado en 2025.
- [2] Carbon Footprint Ltd. Calculadora de huella de carbono. <https://calculator.carbonfootprint.com/calculator.aspx?lang=es&tab=3>. Consultado en 2025.
- [3] EUROCONTROL. Bada - base of aircraft data. <https://www.eurocontrol.int/model/bada>, 2025. Consultado en 2025.
- [4] European Environment Agency. Aviation and shipping emissions. <https://www.eea.europa.eu>, 2022. Consultado en 2025.
- [5] FlightAware. Seguimiento de vuelos en tiempo real. <https://es.flightright.com/>, 2025. Consultado en 2025.
- [6] Folium Contributors. Folium: Python data. leaflet.js maps. <https://python-visualization.github.io/folium/>. Consultado en 2025.
- [7] G. Gurtner and F. Valenti. Performance assessment of free route airspace in europe. *Journal of Air Transport Management*, 95:102109, 2021.
- [8] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference (SciPy)*, pages 11–15, 2008.
- [9] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.
- [10] John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science Engineering*, 9(3):90–95, 2007.
- [11] IATA. Annual review 2020. <https://www.iata.org/en/publications/annual-review/>, 2020. Consultado en 2025.
- [12] ICAO. 2019 environmental report. <https://www.icao.int/environmental-protection/>, 2019. Consultado en 2025.
- [13] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, et al. Jupyter notebooks – a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87–90, 2016.

- [14] Wes McKinney. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, pages 51–56, 2010.
- [15] Ministerio para la Transición Ecológica y el Reto Demográfico. Viajes por motivos de trabajo y emisiones de co2. <https://www.miteco.gob.es/>, 2025. Consultado en 2025.
- [16] OAG Aviation Worldwide. Plataforma de datos de programación aérea. <https://www.oag.com/es>, 2025. Consultado en 2025.
- [17] OpenAI. Chatgpt (modelo gpt-4) como herramienta de asistencia técnica y editorial. <https://chat.openai.com/>, 2025. Consultado de forma continua durante el desarrollo del TFG.
- [18] OpenSky Network. Opensky network data interface. <https://opensky-network.org>, 2025. Consultado en 2025.
- [19] Organización Mundial del Turismo. Barómetro del turismo mundial. <https://www.unwto.org/es/barometro-del-turismo-mundial>, 2023. Consultado en 2025.
- [20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [21] Jeff Reback et al. pandas-dev/pandas: Pandas. *Zenodo*, 2020.
- [22] Guido Van Rossum and Fred L. Drake Jr. *Python 3 Reference Manual*. CreateSpace, 2009.
- [23] Martin Schäfer, Matthias Strohmeier, Vincent Lenders, Ivan Martinovic, and Matthias Wilhelm. Bringing up opensky: A large-scale ads-b sensor network for research. pages 351–362, 2014.
- [24] SESAR Joint Undertaking. Sesar 3 – shaping europe’s digital sky. <https://www.sesarju.eu>, 2025. Consultado en 2025.
- [25] Pauli Virtanen et al. Scipy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 17:261–272, 2020.
- [26] Michael Waskom. Seaborn: Statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.
- [27] Wikipedia. Web scraping. https://es.wikipedia.org/wiki/Web_scraping. Consultado en 2025.
- [28] D. Zhang, X. Han, X. Sun, and S. Wang. Flight trajectory clustering and similarity detection using a sliding window-based approach. *IEEE Transactions on Intelligent Transportation Systems*, 18(4):990–1002, 2017.
- [29] J. Zhao, M. Zanin, and M. Baptista. Clustering of aircraft trajectories using dbscan and kernel density estimation. *Transportation Research Part C: Emerging Technologies*, 97:118–137, 2018.

Anexo A. Repositorio del código fuente

Todo el código desarrollado durante la realización del presente Trabajo de Fin de Grado, así como los scripts de procesamiento, análisis, visualización y recursos auxiliares, se encuentra disponible de forma pública y organizada en el siguiente repositorio:

[https://github.com/JosanFdez/Proyecto-Fin-de-Grado-Ingenieria-
Informatica.git](https://github.com/JosanFdez/Proyecto-Fin-de-Grado-Ingenieria-Informatica.git)

Este repositorio incluye:

- Estructura completa del proyecto en formato Jupyter Notebook.
- Scripts de *scraping* y consumo de APIs para la adquisición de datos aeronáuticos y turísticos.
- Procesos de limpieza, transformación, normalización y fusiones entre *datasets*.
- Algoritmos aplicados (EDA, *clustering*, predicción, modelos de optimización).
- Archivos CSV y Excel con datos originales y resultados intermedios.
- Gráficos generados automáticamente y visualizaciones geoespaciales.
- Versión final del documento en PDF, junto con materiales anexos.

Este entorno ha sido diseñado con un enfoque modular, reproducible y transparente, facilitando así su reutilización o ampliación futura para tareas académicas, investigadoras o aplicadas al sector aeronáutico-turístico.

Anexo B. Glosario

ADS-B: Acrónimo de *Automatic Dependent Surveillance–Broadcast*. Sistema de vigilancia que permite conocer la posición y velocidad de una aeronave en tiempo real mediante señales emitidas por el propio avión.

API: Acrónimo de *Application Programming Interface*. Interfaz de programación que permite acceder y consumir servicios externos, como datos de vuelos en tiempo real.

Clustering: Técnica de aprendizaje automático no supervisado que agrupa elementos similares según sus características. Empleado para segmentar vuelos, aeronaves, países o rutas.

CO₂ estimado: Valor numérico que representa las emisiones de dióxido de carbono generadas por un vuelo. Se estima en función de la duración y el tipo de aeronave.

DBSCAN: Acrónimo de *Density-Based Spatial Clustering of Applications with Noise*. Algoritmo de agrupamiento por densidad, utilizado para detectar estructuras complejas en datos espaciales.

EDA: Acrónimo de *Exploratory Data Analysis*. Fase inicial del análisis de datos, centrada en explorar, visualizar y comprender patrones antes del modelado.

FlightAware: Plataforma en línea que ofrece datos de vuelos comerciales en tiempo real, utilizada como fuente de datos en el proyecto.

Grafo: Estructura matemática compuesta por nodos y aristas. Se emplea para modelar rutas aéreas como conexiones entre aeropuertos.

ICAO: Acrónimo de *International Civil Aviation Organization*. Organismo especializado de la ONU encargado de regular la aviación civil internacional.

K-Means: Algoritmo de agrupamiento que divide los datos en k grupos según similitud interna. Utilizado para clasificar rutas o perfiles turísticos.

OpenSky: Red colaborativa de sensores ADS-B para la recolección y estudio de datos aeronáuticos abiertos. Fuente clave en el TFG.

PCA: Acrónimo de *Principal Component Analysis*. Técnica de reducción de dimensionalidad usada para representar gráficamente agrupamientos o simplificar variables.

Python: Lenguaje de programación principal del trabajo. Utilizado para análisis de datos, scraping, modelado y visualización.

Radar chart: Gráfico en forma de araña utilizado para comparar múltiples métricas simultáneamente. Muy útil en análisis multicriterio de países o aeronaves.

Scraping: Proceso automatizado de extracción de datos desde páginas web. Empleado para obtener información de vuelos, aerolíneas y aeronaves.

Silueta: Métrica para evaluar la calidad de un *clustering*, indicando qué tan bien está asignado cada punto a su grupo.

Sostenibilidad aérea: Medida del impacto medioambiental del transporte aéreo en relación a su retorno económico y turístico. Concepto clave en el análisis.

Trayectoria aérea: Ruta que sigue un avión desde su origen hasta su destino. Analizada para evaluar eficiencia y emisiones.

UTC: Acrónimo de *Universal Time Coordinated*. Referencia global de tiempo utilizada para estandarizar los horarios de los vuelos.