Universidad Politécnica
de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Máster Universitario en Inteligencia Artificial

Trabajo Fin de Máster

Clasificación de Aeropuertos Españoles mediante Análisis Topológico de Datos de sus Trayectorias Aéreas y Retrasos

Autor(a): Isabella Bucciarelli Imbrondone
Tutores: Antonio Jiménez Martín, Alfonso Mateos Caballero

Madrid, Julio - 2025

Este Trabajo Fin de Máster se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Máster
Máster Universitario en Inteligencia Artificial

Título: Clasificación de Aeropuertos Españoles mediante Análisis Topológico de Datos de sus Trayectorias Aéreas y Retrasos

Julio - 2025

Autor(a): Isabella Bucciarelli Imbrondone
Tutores: Antonio Jiménez Martín, Alfonso Mateos Caballero
Departamento de Inteligencia Artificial
ETSI Informáticos
Universidad Politécnica de Madrid

Agradecimientos

Quiero dedicar unas palabras de agradecimiento a todas las personas que, de una u otra manera, han sido parte de este camino y me han acompañado durante la realización de este Trabajo de Fin de Máster (TFM).

En primer lugar, quiero agradecer profundamente a mi familia, a mis amigos y a todas las personas cercanas a mí. Gracias por estar a mi lado en los momentos buenos y en los momentos difíciles, por darme ánimos cuando más los necesitaba, por confiar en mí incluso cuando yo tenía dudas, y por darme su cariño, su apoyo y su compañía en cada etapa de este proceso. Sin ustedes, todo esto habría sido mucho más difícil, y me siento muy afortunada de tenerlos en mi vida.

En segundo lugar, quiero dar las gracias a mis tutores, Alfonso Mateos y Antonio Jiménez, por darme la oportunidad de realizar este TFM y por acompañarme en todo el proceso. Gracias por su tiempo, por resolver mis dudas, por orientarme cuando me sentía perdida y por enseñarme tanto. Todo lo que he aprendido en estos meses ha sido gracias a su apoyo, y siempre estaré agradecida por ello.

También quiero agradecer especialmente a Manuel Cuerno y Fernando Gómez, autores del artículo Cuerno *et al.* (2023), ya que han sido una gran ayuda para llevar a cabo este TFM. Gracias por compartir sus conocimientos, por su colaboración y por guiarme con base en su experiencia. Su trabajo ha sido una referencia clave para poder desarrollar este TFM.

Por último, agradezco a Dios por darme la fuerza, la paciencia y la salud para poder terminar esta etapa. Gracias a Él he podido seguir adelante, superar los obstáculos y lograr este objetivo que tanto significa para mí.

Este Trabajo Fin de Máster es parte del proyecto de I+D+i PID2021-122209OB-C31 y de la Ayuda RED2022-134540-T financiados por MICIU/AEI/10.13039/501100011033.

Resumen

El aumento del tráfico aéreo en los últimos años ha hecho que su gestión (Air Traffic Management, ATM) sea cada vez más compleja. Esto ha dado lugar a una gran cantidad de datos generados por los vuelos, que incluyen información sobre trayectorias, tiempos y otras variables relacionadas. Estos datos son difíciles de analizar porque están altamente interconectados, son multidimensionales, cambian constantemente, presentan una gran variabilidad y, en muchos casos, contienen errores o están incompletos. Todo esto limita la capacidad de los enfoques tradicionales para analizar esta información y detectar patrones, relaciones o anomalías de forma eficaz.

En este contexto, estudios recientes han propuesto el uso del análisis topológico de datos (*Topological Data Analysis*, TDA) como una alternativa para abordar la complejidad de la información generada por el tráfico aéreo. Con este objetivo, una de estas investigaciones desarrolló una metodología basada en técnicas topológicas aplicadas al estudio de trayectorias de aeronaves y los retrasos asociados, centrándose en vuelos que despegan y aterrizan en aeropuertos españoles. Según sus autores, el análisis topológico podría ser una técnica prometedora para superar las limitaciones del análisis de datos en el ámbito de la aviación.

Siguiendo este enfoque, este Trabajo de Fin de Máster toma como base dicha propuesta para ampliar y evaluar el uso del TDA en el análisis de datos de vuelos. Para ello, se desarrolla un modelo que combina análisis topológico de datos con técnicas de aprendizaje automático, con el objetivo de clasificar los aeropuertos españoles según el comportamiento de sus vuelos. El análisis tiene en cuenta tanto los retrasos como las diferencias entre las trayectorias planificadas (las rutas previstas del vuelo) y las trayectorias reales (rutas efectivamente voladas), partiendo de la metodología usada en el estudio previo y adaptándola con ajustes propios para mejorar su aplicabilidad.

Finalmente, la validación del modelo se realiza comparando la clasificación obtenida con los resultados de otros estudios.

Abstract

The increase in air traffic in recent years has made its management (Air Traffic Management, ATM) increasingly complex. This has led to the generation of a large amount of data from flights, including information about trajectories, timings, and other related variables. These data are difficult to analyze because they are highly interconnected, multidimensional, constantly changing, and exhibit high variability. In many cases, they also contain errors or are incomplete. All of this limits the ability of traditional approaches to effectively analyze this information and detect patterns, relationships, or anomalies.

In this context, recent studies have proposed the use of Topological Data Analysis (TDA) as an alternative to address the complexity of information generated by air traffic. With this objective in mind, one of these studies developed a methodology based on topological techniques applied to the analysis of aircraft trajectories and associated delays, focusing on flights that depart from or arrive at Spanish airports. According to its authors, topological analysis could be a promising technique to overcome the limitations of conventional data analysis in the aviation domain.

Following this approach, this Master's Thesis builds upon that proposal to expand and evaluate the use of TDA in flight data analysis. To this end, a model is developed that combines topological data analysis with machine learning techniques, aiming to classify Spanish airports based on the behavior of their associated flights. The analysis considers both delays and differences between planned trajectories (the scheduled routes for the flight) and actual trajectories (the routes actually flown), starting from the methodology used in the previous study and adapting it with custom modifications to enhance its applicability.

Finally, the model is validated by comparing the resulting classification with the outcomes of other studies.

Tabla de contenidos

Capítulo 1

Introducción

En los últimos años, el crecimiento del tráfico aéreo ha hecho que la gestión del tráfico aéreo (*Air Traffic Management*, ATM) sea cada vez más compleja, no solo en cuanto a lo operativo, sino también en lo que respecta al análisis de los datos que se generan. Los distintos sistemas involucrados en el ATM, como los radares, los sistemas de navegación, los registros de vuelo o los sensores de las aeronaves, producen datos variados y difíciles de tratar. Estos datos son multidimensionales, están altamente interconectados, cambian constantemente y, en muchos casos, presentan errores o están incompletos. Estas características hacen que su análisis resulte especialmente complejo. Además, el aumento continuo del número de vuelos ha incrementado la cantidad de datos disponibles, lo que hace cada vez más necesaria la incorporación de herramientas capaces de trabajar con esta información sin perder precisión ni eficiencia. Según datos del Consejo Internacional de Aeropuertos (*Airports Council International*, ACI), en 2024 se registraron cerca de 100 millones de despegues y aterrizajes en todo el mundo, y se estima que esta cifra podría llegar a 149 millones en 2043 y superar los 176 millones en 2053 (Airports Council International (ACI), 2024).

En este contexto, una de las aproximaciones para estudiar los datos generados por los vuelos es a través del análisis de las trayectorias que siguen las aeronaves. Sin embargo, este tipo de análisis también presenta varios desafíos. Las trayectorias de vuelo son secuencias temporales de puntos que incluyen variables como latitud, longitud, altitud, velocidad y tiempo, lo que las convierte en datos complejos, tanto por su naturaleza multidimensional como por su carácter dinámico. Esta complejidad dificulta su visualización, comparación y análisis, especialmente cuando se busca identificar patrones generales o comportamientos comunes (Bian *et al.* 2018). Además, estos datos suelen estar afectados por ruido o errores provocados por condiciones meteorológicas, decisiones humanas o fallos en los sistemas de adquisición (Burmester *et al.* 2018).

Frente a estos desafíos, el estudio realizado por Cuerno *et al.* (2023) propone el uso de análisis topológico de datos (*Topological Data Analysis*, TDA) como una alternativa para el análisis de trayectorias aéreas. El TDA se basa en conceptos de topología algebraica para describir la "forma" de los datos, permitiendo identificar patrones globales, bucles, agrupaciones complejas y relaciones no evidentes, incluso en presencia de ruido, valores atípicos o datos incompletos (Carlsson, 2009). En particular, este estudio emplea la homología persistente, una técnica dentro de TDA que anali-

za cómo cambian las características topológicas de los datos a diferentes niveles de detalle o escalas (Edelsbrunner y Morozov, 2013).

La idea que proponen es que la aplicación de TDA, y en particular de la homología persistente al análisis de datos de vuelos ofrece varias ventajas. En primer lugar, permite capturar propiedades globales del movimiento de los vuelos, yendo más allá de las relaciones locales entre puntos, lo que facilita la detección de características o patrones presentes en ciertos vuelos. Otra ventaja es su robustez frente al ruido y a pequeñas perturbaciones en los datos, como las causadas por condiciones meteorológicas o variaciones en los procedimientos de vuelo, sin necesidad de reducir previamente la dimensionalidad. Por último, el cálculo de la homología persistente puede ejecutarse de forma paralela, lo que ayuda a disminuir el coste computacional.

Hoy en día, TDA, y en particular la homología persistente, se ha aplicado en múltiples campos de la ingeniería. Por ejemplo, en el análisis de redes neuronales en neurociencia (Giusti *et al.* 2016), en la planificación de rutas para robots autónomos (Bhattacharya *et al.* 2013, 2015), y en el estudio del comportamiento de exploración y cobertura de sistemas multi-robot (Bhattacharya *et al.* 2014). Asimismo, ha sido utilizado en otros ámbitos, como el financiero, donde Domínguez Monterroza *et al.* (2023) emplearon TDA para analizar los cambios y la estabilidad del mercado integrado latinoamericano (MILA). Sin embargo, no se han visto aplicaciones de TDA en el ámbito de la aviación.

De esta manera, el presente trabajo surge como una inspiración a partir del estudio realizado por Cuerno *et al.* (2023), con el propósito de profundizar en la aplicación del análisis topológico de datos en el ámbito de la aviación, específicamente en el análisis de datos de vuelo. Para ello, se utiliza el mismo conjunto de datos empleado en dicho estudio, compuesto de vuelos que despegaron o aterrizaron en aeropuertos españoles durante un determinado período de tiempo, y se adapta su metodología original mediante algunas modificaciones y se combina con técnicas de aprendizaje automático (*machine learning*) para clasificar los aeropuertos españoles en función del comportamiento de los vuelos asociados a ellos.

Para contextualizar, hay que saber que cada vuelo suele contar con dos trayectorias: una planificada, establecida en el plan de vuelo, y otra real, que refleja el recorrido efectivamente ejecutado por la aeronave. Analizar las diferencias entre ambas puede revelar patrones operacionales recurrentes, especialmente durante las fases de aproximación o salida. La Figura 1.1 muestra un ejemplo representativo de la trayectoria planificada y real de un vuelo, con la trayectoria planificada en azul y la real en rojo.

Las discrepancias entre ambas trayectorias pueden deberse a diversos factores, como la congestión del espacio aéreo, las condiciones meteorológicas o restricciones operativas en los aeropuertos. Por ello, el análisis de estas desviaciones, junto con los retrasos de llegada, puede proporcionar información valiosa sobre la eficiencia y el funcionamiento de los aeropuertos implicados.

Generalmente, los aeropuertos se clasifican utilizando métricas cuantitativas, como el volumen de pasajeros o el número de operaciones anuales. En este Trabajo de Fin de Máster (TFM) se realiza una clasificación alternativa de los aeropuertos españoles basada en el análisis topológico de los datos de vuelo de todas las aeronaves que despegan o aterrizan en ellos. La clasificación se realiza considerando tanto la distan-

cia (desviación) entre las trayectorias planificadas y las trayectorias reales, como los retrasos registrados en las llegadas. Para ello, como ya se mencionó, se toma como base la metodología utilizada por Cuerno *et al.* (2023) para el análisis topológico de los datos de vuelo, a la cual se le aplican ciertas modificaciones y se combina con técnicas de aprendizaje automático para clasificar los aeropuertos españoles.

Posteriormente, las clasificaciones obtenidas se comparan con los resultados del estudio realizado por Cuerno *et al.* (2023) y otro Trabajo de Fin de Máster que también ha hecho una clasificación de aeropuertos basada en las mismas variables (distancia entre las trayectorias de los vuelos y el retraso que estos presentan) y a partir del mismo conjunto de datos, con el objetivo de validar los resultados y evaluar la efectividad del análisis.

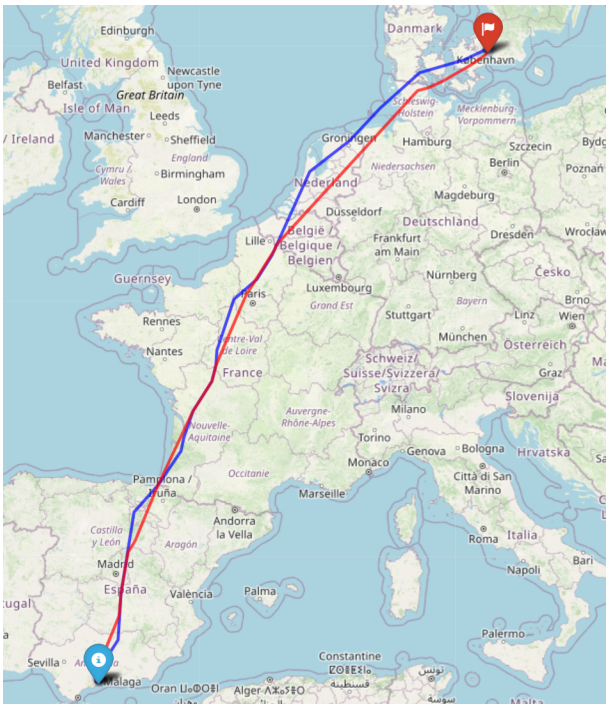

Figura 1.1: Ejemplo de trayectoria planificada (azul) y trayectoria real (roja) de un vuelo.

1.1. Objetivos

A continuación, se presentan el objetivo general y los objetivos específicos de este Trabajo de Fin de Máster (TFM):

1.1.1. Objetivo general

Clasificar los aeropuertos españoles mediante un análisis topológico de datos de vuelos que despegan o aterrizan en España, basándose en las distancias entre las trayectorias planificadas y reales y los retrasos de llegada asociados.

1.1.2. Objetivos específicos

- Recolectar y preprocesar datos de vuelos que despegan o aterrizan en aeropuertos españoles.
- Ajustar y adaptar la metodología propuesta por Cuerno *et al.* (2023).
- Calcular los retrasos y las métricas de distancias entre las trayectorias planificadas y reales.
- Aplicar técnicas de análisis topológico de datos para identificar patrones característicos asociados a cada aeropuerto.
- Medir la similitud entre aeropuertos en función de los patrones topológicos identificados.
- Utilizar algoritmos de aprendizaje automático para clasificar los aeropuertos según el comportamiento de sus vuelos.
- Comparar la clasificación obtenida con los resultados de Cuerno *et al.* (2023), evaluando las diferencias y aportaciones del nuevo enfoque.

1.2. Estructura de la memoria

La presente memoria se estructura del siguiente modo: el Capítulo 2 analiza las tecnologías utilizadas hasta ahora para el análisis de trayectorias de aeronaves, así como las limitaciones que presentan. En el Capítulo 3 se introduce el análisis topológico de datos, explicando los conceptos necesarios para comprender esta técnica, con especial énfasis en la homología persistente. El Capítulo 4 describe en detalle la metodología empleada, las diferencias con respecto al estudio anterior y las herramientas utilizadas para su implementación. A continuación, el Capítulo 5 presenta cómo está formado el conjunto de datos, los procedimientos de limpieza aplicados a los mismos, los resultados de la metodología propuesta, junto con un análisis comparativo frente a los resultados obtenidos en el estudio de Cuerno *et al.* (2023) y otro Trabajo de Fin de Máster. Por último, el Capítulo 6 expone las conclusiones, las principales limitaciones detectadas y posibles líneas de trabajo futuro.

Capítulo 2

Estado del arte

Tradicionalmente, para facilitar el análisis y la visualización de trayectorias de vuelo, se han utilizado técnicas de reducción de dimensionalidad como el análisis de componentes principales (*Principal Component Analysis*, PCA) e incrustación estocástica de vecinos con distribución *t* (*t-distributed stochastic neighbor embedding*, *t*-SNE). Estas herramientas permiten proyectar los datos en espacios de menor dimensión. Por ejemplo, PCA se ha utilizado en diversos trabajos como preprocesamiento para reducir la dimensionalidad de los datos de trayectoria, como en Wang *et al.* (2017) y Yoon y Lee (2025). Además, Zeng *et al.* (2021) emplearon *t*-SNE para representar visualmente patrones de tráfico aéreo, destacando agrupaciones de trayectorias en el espacio aéreo terminal. No obstante, PCA asume que los datos son lineales, por lo que puede no capturar relaciones no lineales importantes en los datos y su interpretación puede resultar compleja. Por otro lado, *t*-SNE, aunque efectiva para la visualización de agrupaciones complejas, es computacionalmente costosa, no siempre reproducible y también requiere interpretación experta.

Además, últimamente se ha extendido el uso de técnicas de aprendizaje automático (*machine learning*) y aprendizaje profundo (*deep learning*) para el análisis de trayectorias de aeronaves. Por ejemplo, Liu y Hansen (2018) desarrollaron un modelo que combina redes neuronales convolucionales (*Convolutional Neural Networks*, CNN) y recurrentes (*Recurrent Neural Networks*, RNN) para predecir trayectorias. De manera similar, Huang *et al.* (2024) propusieron un modelo basado en imágenes de trayectorias y redes neuronales profundas para predecir los tiempos de aterrizaje. Wang *et al.* (2017) desarrollaron un modelo híbrido para la predicción 4D a partir de datos ADS-B (información transmitida automáticamente por aeronaves sobre su posición, velocidad, altitud, etc.). Zeng *et al.* (2021) utilizaron *autoencoders* profundos y modelos de mezcla gaussiana (*Gaussian Mixture Models*, GMM) para representar trayectorias en espacios reducidos y agruparlas por comportamiento. Bolić *et al.* (2022) emplearon *DBSCAN* para categorizar trayectorias en distintos escenarios de espacio aéreo terminal, teniendo en cuenta su variabilidad espacial y temporal. Gariel *et al.* (2011) usaron *K-Means* para identificar desviaciones en las trayectorias.

Sin embargo, este tipo de técnicas presenta importantes limitaciones. En primer lugar, muchos algoritmos de clustering, como *K-Means* o *DBSCAN*, dependen de parámetros como el número de clústeres o el radio de vecindad. Además, algoritmos como *K-Means* asumen distribuciones esféricas u homogéneas de los datos, lo que puede ser problemático cuando las trayectorias tienen estructuras complejas. Por otra par-

te, los modelos de aprendizaje profundo, aunque muy potentes, requieren grandes volúmenes de datos etiquetados y una capacidad computacional elevada. También tienden a comportarse como "cajas negras", dificultando la interpretación de los resultados, lo cual es una limitación importante en contextos como el control del tráfico aéreo. Finalmente, tanto las redes neuronales como los modelos de clustering pueden ser sensibles al ruido y a la calidad de los datos de entrada, como errores en el ADS-B.

Por otro lado, también se han empleado técnicas basadas en grafos para analizar comportamientos entre aeropuertos a partir de sus trayectorias. Por ejemplo, Li y Jing (2021) construyeron un modelo de red basado en trayectorias para analizar la propagación de retrasos entre aeropuertos, modelando las conexiones de tráfico aéreo como un grafo dirigido y ponderado. Sin embargo, hay que tener en cuenta que este tipo de modelos basados en grafos solo permiten evaluar relaciones entre pares de aeropuertos. Asimismo, otros estudios han analizado la estructura y el rendimiento de redes de transporte aéreo. Por ejemplo, Zhou *et al.* (2019) desarrollaron una métrica de eficiencia ponderada para evaluar la robustez y eficiencia de redes aéreas nacionales, considerando pesos como la frecuencia de vuelos y el número de rutas. No obstante, estos enfoques presentan limitaciones, como la sensibilidad a parámetros, la complejidad en la interpretación de resultados y la incapacidad para capturar relaciones entre más de dos aeropuertos.

Frente a las limitaciones de los enfoques tradicionales para el análisis de trayectorias de aeronaves, el análisis topológico de datos (*Topological Data Analysis*, TDA) puede ser utilizado como una alternativa para analizar y extraer características de las trayectorias de vuelos. Como se mencionó anteriormente en el Capítulo 1, aplicar homología persistente al análisis de datos de vuelos ofrece varias ventajas. En primer lugar, permite capturar características globales sin limitarse a relaciones estrictamente locales, lo que facilita la detección de patrones recurrentes, desviaciones o anomalías operacionales. Además, este enfoque permite analizar relaciones complejas de orden superior que no pueden detectarse fácilmente mediante métodos tradicionales. Otra ventaja importante es su robustez frente al ruido y a pequeñas perturbaciones en los datos, como las provocadas por condiciones meteorológicas o variaciones en los procedimientos de vuelo, sin requerir una reducción previa de la dimensionalidad. Asimismo, el cálculo de la homología persistente puede ejecutarse de manera paralela, lo que permite reducir el costo computacional. Por último, los resultados obtenidos mediante TDA pueden integrarse con modelos de aprendizaje automático, facilitando tareas como clasificación, agrupamiento o detección de anomalías a partir de representaciones topológicas.

En el ámbito de la gestión del tráfico aéreo (*Air Traffic Management*, ATM), aún no existen aplicaciones conocidas del análisis topológico de datos. Sin embargo, algunos trabajos han comenzado a explorar su potencial. Por ejemplo, Li *et al.* (2019) analizaron la estructura topológica de superficies aeroportuarias con el objetivo de evaluar su conectividad operativa, demostrando que TDA puede reflejar si un aeropuerto está operando de forma total o parcial.

Por su parte, como se mencionó en el Capítulo 1, Cuerno *et al.* (2023) aplicaron técnicas de homología persistente al análisis de trayectorias de vuelos en aeropuertos españoles. En su estudio, utilizaron un conjunto de datos que incluía vuelos que despegaban o aterrizaban en aeropuertos españoles durante un largo periodo de tiempo.

Para cada vuelo, calcularon dos variables: la distancia entre la trayectoria planificada (prevista antes del vuelo) y la trayectoria real (efectivamente volada), y el retraso en la hora de llegada. A partir de esto, representaron cada vuelo como un punto en un espacio bidimensional definido por esas dos variables.

Posteriormente, agruparon los vuelos según el aeropuerto al que estaban asociados (ya fuera como origen o destino) y aplicaron un análisis topológico a los conjuntos de puntos de cada aeropuerto. De este modo, obtuvieron una representación topológica específica para cada uno, que capturaba el comportamiento conjunto de sus vuelos. Con estas representaciones, calcularon distancias entre aeropuertos y las compararon con la clasificación oficial de AENA (gestor aeroportuario y de navegación aérea en España). El análisis mostró que TDA era capaz de identificar diferencias entre aeropuertos que, según la clasificación de AENA, pertenecían al mismo grupo. Por ejemplo, detectaron que el aeropuerto de Zaragoza, a pesar de ser considerado pequeño, presentaba un comportamiento diferenciado respecto a otros aeropuertos con características similares según dicha clasificación.

Como también se mencionó en el Capítulo 1, en este TFM se retoma el enfoque propuesto por Cuerno *et al.* (2023), basado en el análisis topológico mediante homología persistente aplicado a vuelos en aeropuertos españoles. Para ello, se utiliza el mismo conjunto de datos y se parte de su metodología, introduciendo diversas modificaciones con el objetivo de ampliar y refinar el análisis original. Entre ellas, se incluyen mejoras en el cálculo de la distancia entre trayectorias mediante un ajuste temporal, así como la incorporación de técnicas de aprendizaje automático para clasificar los aeropuertos según sus características topológicas. Estas modificaciones metodológicas permiten realizar un análisis más robusto, detallado y comparativo del comportamiento operacional de los aeropuertos.

Por otro lado, Serrano (2025) desarrolló otro Trabajo de Fin de Máster que también se basa en la metodología propuesta por Cuerno *et al.* (2023), pero explorando un enfoque diferente de análisis topológico mediante el uso del algoritmo *Mapper* (Singh *et al.* 2007). Este algoritmo, que forma parte del análisis topológico de datos, permite representar un conjunto de datos como un grafo: los nodos agrupan observaciones similares y las conexiones entre ellos indican que comparten elementos. *Mapper* es especialmente útil para visualizar la forma general de datos complejos y descubrir patrones que pueden no detectarse con métodos más tradicionales.

Al igual que en este TFM, utiliza el mismo conjunto de datos empleado en el artículo de Cuerno *et al.* (2023), que incluye vuelos que despegan o aterrizan en aeropuertos españoles. Los vuelos los agrupa por aeropuerto y, para cada uno, calcula estadísticas como la media, mediana, desviación típica y rango intercuartílico, tanto de la diferencia entre la trayectoria planificada y la real como del retraso en la llegada asociado a cada vuelo. Estas variables las estandariza y las reduce mediante análisis de componentes principales, y los resultados los utiliza como entrada al algoritmo Mapper para generar un grafo que representa la estructura de los datos.

Posteriormente, utiliza dos algoritmos de agrupamiento distintos (*DBSCAN* y *HDBSCAN*), que aplica tanto de forma global a cada aeropuerto como de manera segmentada por meses. De esta forma, para cada algoritmo de agrupación obtiene una agrupación o representación de como se relacionan los aeropuertos españoles basada en las distancias entre las trayectorias planificadas y reales, así como en el retraso asociado. Además, cuando el análisis se realiza por meses, obtiene dos representaciones

adicionales por cada mes (una por cada algoritmo de agrupación). Los resultados permiten identificar grupos de aeropuertos con comportamientos similares y detectar aeropuertos con comportamientos atípicos o anómalos, también conocidos como outliers funcionales. Asimismo, se observa que algunos aeropuertos geográficamente próximos tienden a agruparse, y que las agrupaciones obtenidas difieren de la clasificación oficial de AENA, ya que están basadas en el comportamiento operativo de los vuelos (retrasos y desviaciones), en lugar de en criterios como el volumen de pasajeros.

Capítulo 3

Análisis topológico de datos

La topología es una rama de las matemáticas que estudia las propiedades de los espacios que se preservan bajo deformaciones continuas, como estiramientos, compresiones o dobleces, pero no bajo rupturas o uniones. Por ejemplo, una goma elástica con forma de círculo puede deformarse en un óvalo o una figura más compleja, pero seguirá teniendo un "agujero" (un bucle topológico) mientras no se corte ni se pegue. Esta propiedad de invarianza permite a la topología describir la "forma" esencial de un espacio, enfocándose en características como componentes conexas, bucles o cavidades.

El análisis topológico de datos (*Topological Data Analysis*, TDA) (Chazal y Michel, 2021) es una disciplina que combina herramientas de la topología algebraica con técnicas computacionales para estudiar la estructura geométrica de los datos. En TDA, los datos se representan como una nube de puntos en un espacio métrico, y el objetivo es identificar características topológicas subyacentes, como agrupaciones (componentes conexas), ciclos o agujeros (bucles) o regiones vacías (cavidades), que persisten a diferentes escalas de observación.

3.1. Nubes de puntos

Como se mencionó, en TDA los datos se representan como una nube de puntos en un espacio geométrico. Cada punto es una observación, y la distancia entre puntos nos dice qué tan similares o cercanos son. La nube de puntos se puede representar como:

$$X = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^d.$$

Por ejemplo, si los puntos están distribuidos en forma de anillo, puede haber un bucle en el centro. Si hay dos grupos separados, hay dos componentes conexas.

3.2. Complejos simpliciales

Para estudiar la forma de los datos de manera matemática, el análisis topológico de datos transforma la nube de puntos en un objeto estructurado llamado complejo simplicial (*simplicial complex*).

Un complejo simplicial es una construcción geométrica formada por la unión de puntos, segmentos, triángulos, tetraedros y sus análogos en dimensiones superiores, denominados símplices.

3.2.1. Símplices

Un k-símplice (*k-simplex*) es el envolvente convexo de un conjunto de $k+1$ puntos afínmente independientes, es decir, puntos que no están contenidos en un subespacio afín de dimensión menor a k. Estas estructuras generalizan la noción de triángulo y tetraedro a dimensiones superiores, y constituyen los elementos básicos de los complejos simpliciales. Por ejemplo:

- Un 0-símplice (*0-simplex*) es un punto.
- Un 1-símplice (*1-simplex*) es un segmento definido por dos puntos.
- Un 2-símplice (*2-simplex*) es un triángulo definido por tres puntos.
- Un 3-símplice (*3-simplex*) es un tetraedro definido por cuatro puntos.
- En general, un k-símplice (*k-simplex*) está definido por $k+1$ puntos afínmente independientes.

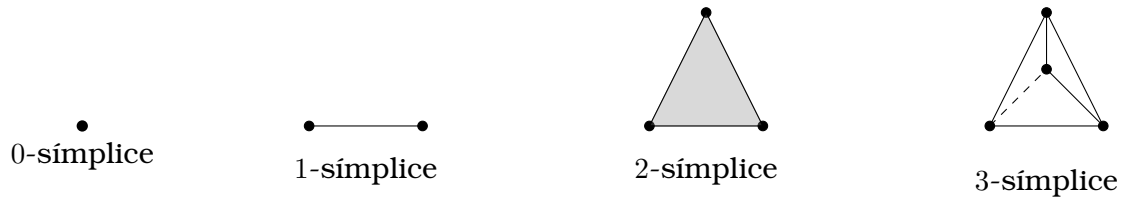

Figura 3.1: Ejemplos de símplices: punto (0D), línea (1D), triángulo (2D), y tetraedro (3D).

Caras de un símplice

Cada k-símplice contiene subestructuras más simples llamadas *caras* (*faces*). Una cara de un k-símplice es cualquier subconjunto de sus vértices. Por ejemplo:

- Un triángulo (2-símplice) tiene 3 aristas (1-símplices) y 3 vértices (0-símplices) como caras.
- Un tetraedro (3-símplice) tiene 4 triángulos (2-símplices), 6 aristas (1-símplices) y 4 vértices (0-símplices) como caras.

Frontera de un símplice

La *frontera* (*boundary*) de un k-símplice está compuesta por la suma formal de todas sus caras de dimensión $k-1$. Por ejemplo, la frontera de un triángulo son sus tres lados, y la de un tetraedro, sus cuatro caras triangulares. En la Figura 3.2 se puede ver gráficamente cuales serían los 1-símplice (aristas) que conforman la frontera de un 2-símplice (triángulo).

Este concepto se formaliza con el operador frontera (*boundary operator*), denotado como ∂_k, que actúa sobre un k-símplice y produce una combinación lineal de sus $(k-1)$-símplices. Por ejemplo:

$$\partial_2([v_0, v_1, v_2]) = [v_1, v_2] - [v_0, v_2] + [v_0, v_1].$$

Esto significa que la frontera de un triángulo orientado está formada por sus tres aristas con una orientación (signo) determinada. La orientación es fundamental para definir correctamente ciclos y fronteras en homología.

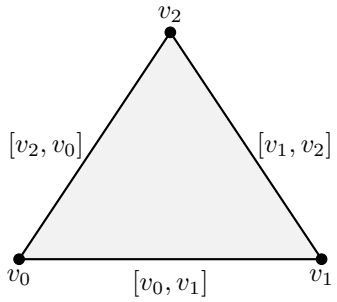

Figura 3.2: Frontera de un 2-símplice.

3.2.2. Definición de complejo simplicial

Un complejo simplicial K es un conjunto de símplices que cumple las siguientes propiedades:

1. Si un símplice $\sigma \in K$, entonces todas sus caras también pertenecen a K.
2. La intersección de dos símplices $\sigma_1, \sigma_2 \in K$ es vacía o una cara común a ambos.

3.2.3. Construcción de complejos

Dado un conjunto de puntos $X \subset \mathbb{R}^d$, se pueden construir distintos tipos de complejos simpliciales mediante un parámetro de proximidad $r > 0$.

3.2.3.1. Complejo de Čech

El complejo de Čech (*Čech Complex*) $\check{C}_r(X)$ se define como el complejo simplicial cuyos k-símplices corresponden a subconjuntos $\{x_{i_0}, \ldots, x_{i_k}\} \subset X$ tales que:

$$\check{C}_r(X) = \left\{ \{x_{i_0}, \ldots, x_{i_k}\} \subset X \mid \bigcap_{j=0}^{k} B_r(x_{i_j}) \neq \emptyset \right\}. \tag{3.1}$$

Donde $B_r(x)$ denota la bola cerrada de radio r centrada en $x \in \mathbb{R}^d$. Es decir, se añade un k-símplice si las bolas de radio r en torno a sus vértices tienen intersección no vacía. Por ejemplo, en la Figura 3.3 se ilustra cómo se forma un 2-símplice en un complejo de Čech: las tres bolas centradas en los vértices correspondientes se intersectan todas en una región común.

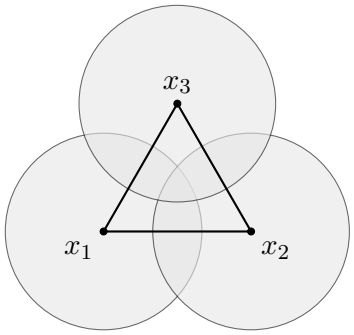

Figura 3.3: Ejemplo de formación de un 2-símplice en el complejo de Čech.

3.2.3.2. Complejo de Vietoris–Rips

El complejo de Vietoris–Rips *(Vietoris–Rips Complex)* $VR_r(X)$ es una relajación computacionalmente más eficiente del complejo de Čech. Se define como el complejo simplicial cuyos k-símplices son aquellos subconjuntos $\{x_{i_0}, \ldots, x_{i_k}\} \subset X$ tales que:

$$\|x_{i_j} - x_{i_l}\| \leq r \quad \text{para todo } 0 \leq j < l \leq k. \tag{3.2}$$

Es decir, se añade un k-símplice si todos sus vértices están a distancia mutua a lo sumo r, lo cual equivale a formar un clique en el grafo de vecindad. Por ejemplo, en la Figura 3.4 se muestra cómo se forma un 2-símplice en un complejo de Vietoris–Rips. En este caso, se añade un 2-símplice si todos los pares de puntos que definen sus vértices se encuentran a una distancia menor o igual que r. Esta condición es equivalente a que las bolas de radio r centradas en dichos puntos se intersecten dos a dos, aunque no necesariamente de forma conjunta. A diferencia del complejo de Čech, no se requiere que la intersección entre todas las bolas sea común, sino únicamente que exista una intersección entre cada par.

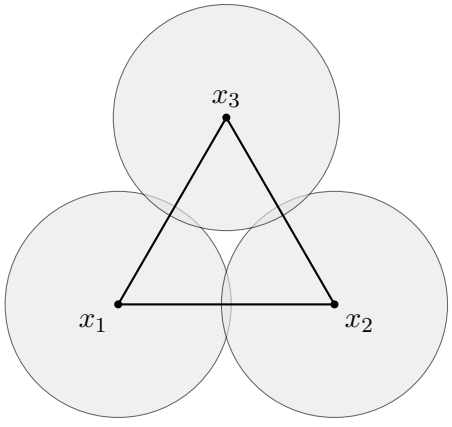

Figura 3.4: Ejemplo de formación de un 2-símplice en el complejo de Vietoris–Rips.

Aunque $VR_r(X)$ no coincide necesariamente con $\check{C}_r(X)$, se cumple que:

$$\check{C}_r(X) \subseteq VR_r(X) \subseteq \check{C}_{2r}(X),$$

lo que implica que el complejo de Vietoris–Rips es una buena aproximación del complejo de Čech y resulta computacionalmente más manejable, especialmente en espacios de alta dimensión donde construir intersecciones exactas de bolas es costoso.

3.3. Homología persistente

La homología persistente (*Persistent Homology*) es el área de TDA que permite analizar cómo evolucionan las características topológicas a medida que varía el parámetro r (Carlsson, 2009; Edelsbrunner y Morozov, 2013). En lugar de estudiar un único complejo simplicial, se construye una filtración, que consiste en una secuencia creciente de complejos simpliciales:

$$K_{r_1} \subseteq K_{r_2} \subseteq \cdots \subseteq K_{r_m},$$

donde $r_1 < r_2 < \cdots < r_m$ son valores crecientes del parámetro, y cada K_{r_i} es un complejo simplicial construido sobre la misma nube de puntos $X \subset \mathbb{R}^d$, ya sea mediante un complejo de Čech $\check{C}_{r_i}(X)$ o un complejo de Vietoris–Rips $VR_{r_i}(X)$.

Una filtración es, por tanto, una familia creciente de complejos simpliciales construidos al ir aumentando el parámetro r. A medida que r crece, se van añadiendo más símplices (vértices, aristas, triángulos, etc.), lo que permite observar cómo surgen, persisten y desaparecen características topológicas como componentes conexas, agujeros o cavidades.

El objetivo de la homología persistente es rastrear estas características a lo largo de la filtración, identificando en qué escala r nacen y en cuál desaparecen. Esto proporciona información sobre la estructura de los datos y permite distinguir entre ruido y patrones significativos. En la Figura 3.5, se puede ver un ejemplo de filtración de un complejo de Vietoris-Rips a partir de tres puntos. A medida que crece el parámetro r, las bolas alrededor de los puntos se intersectan y se añaden aristas y símplices al complejo.

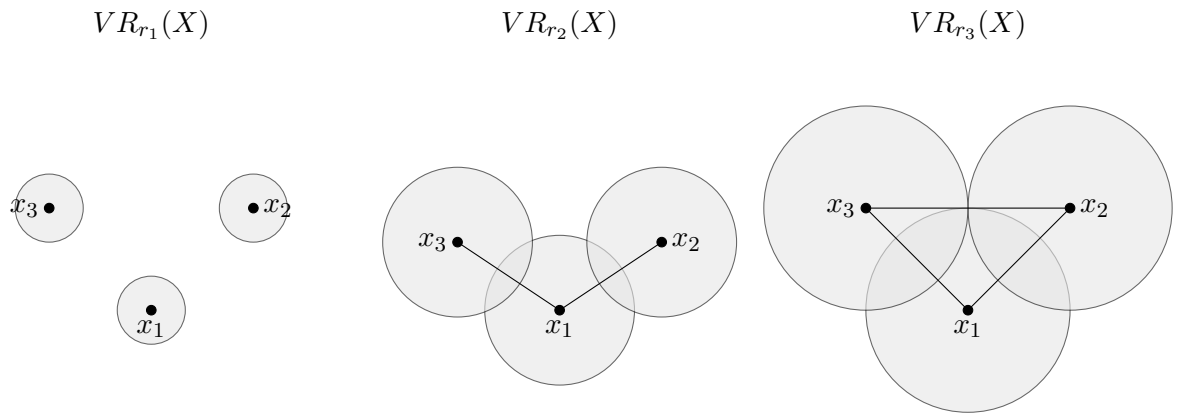

Figura 3.5: Ejemplo de filtración de complejos de Vietoris–Rips.

3.3.1. Grupos de homología

Los grupos de homología (*Homology Groups*) son una herramienta en topología algebraica que permiten describir y cuantificar la presencia de estructuras de diferentes dimensiones en un espacio, como:

- H_0: componentes conexas (agrupaciones separadas)
- H_1: bucles o agujeros (ciclos no rellenables)
- H_2: cavidades (superficies cerradas no delimitadas por una región tridimensional)
- y así sucesivamente

A cada dimensión $k \geqslant 0$ se le asocia un grupo de homología H_k que describe los "agujeros" k-dimensionales en el espacio.

Los grupos de homología se construyen a partir de una cadena, definida por los siguientes elementos:

- Grupo de cadenas C_k: El grupo de cadenas k-dimensionales es el grupo generado por los k-símplices de K. Formalmente, si K tiene k-símplices $\{\sigma_1, \sigma_2, \ldots, \sigma_m\}$, entonces:

$$C_k = \left\{ \sum_{i=1}^{m} a_i \sigma_i \mid a_i \in \mathbb{Z} \right\}, \tag{3.3}$$

  donde los coeficientes a_i son enteros. Una cadena k-dimensional es una combinación lineal de k-símplices, que representa una "superficie" o estructura geométrica en dimensión k.

- Operador frontera ∂_k: El operador frontera es un homomorfismo $\partial_k : C_k \to C_{k-1}$ que asigna a cada k-símplice su frontera, definida como la suma alternada de sus $(k-1)$-caras. El operador frontera actúa linealmente sobre cadenas:

$$\partial_k([v_0, \ldots, v_k]) = \sum_{i=0}^{k} (-1)^i [v_0, \ldots, \hat{v}_i, \ldots, v_k], \tag{3.4}$$

  donde $\hat{v}_i$ indica que se omite el vértice v_i. Esto generaliza el ejemplo anterior para cualquier dimensión. Por ejemplo:

  - Para un 1-símplice (arista) $[v_0, v_1]$, la frontera es:

$$\partial_1([v_0, v_1]) = v_1 - v_0.$$

  - Para un 2-símplice (triángulo) $[v_0, v_1, v_2]$, la frontera es:

$$\partial_2([v_0, v_1, v_2]) = [v_1, v_2] - [v_0, v_2] + [v_0, v_1].$$

  - Para una cadena 2-dimensional, por ejemplo:

$$c = [v_0, v_1, v_2] + [v_1, v_2, v_3],$$

    su frontera se calcula aplicando linealmente ∂_2 a cada símplice:

$$\partial_2(c) = \partial_2([v_0, v_1, v_2]) + \partial_2([v_1, v_2, v_3]) =$$

$$([v_1, v_2] - [v_0, v_2] + [v_0, v_1]) + ([v_2, v_3] - [v_1, v_3] + [v_1, v_2]).$$

    Agrupando términos semejantes:

$$\partial_2(c) = ([v_0, v_1] - [v_0, v_2] - [v_1, v_3] + [v_2, v_3]) + 2[v_1, v_2].$$

Una propiedad del operador frontera es que $\partial_{k-1} \circ \partial_k = 0$, es decir, la frontera de una frontera es siempre nula. Esto refleja que los bordes de una superficie no tienen bordes propios.

- Grupo de ciclos Z_k: El grupo de ciclos k-dimensionales es el núcleo del operador frontera:
$$Z_k = \ker(\partial_k) = \{c \in C_k \mid \partial_k(c) = 0\}. \tag{3.5}$$
Los ciclos son cadenas que no tienen frontera, como un bucle cerrado (en dimensión 1) o una superficie cerrada (en dimensión 2).

- Grupo de fronteras B_k: El grupo de fronteras k-dimensionales es la imagen del operador frontera en la dimensión superior:
$$B_k = \text{im}(\partial_{k+1}) = \{\partial_{k+1}(d) \mid d \in C_{k+1}\}. \tag{3.6}$$
Las fronteras son ciclos que son el borde de una cadena de dimensión superior. Por ejemplo, el contorno de un triángulo es una frontera porque es la frontera de una superficie 2-dimensional.

- Grupo de homología H_k: El grupo de homología en dimensión k se define como el cociente:
$$H_k = Z_k / B_k, \tag{3.7}$$
es decir, los ciclos k-dimensionales módulo los ciclos que son fronteras. Esto captura los "agujeros" que no pueden llenarse con cadenas de dimensión superior. El *número de Betti* $\beta_k = \text{rank}(H_k)$ indica el número de agujeros independientes en dimensión k.

3.4. Representaciones de la homología persistente

Como se dijo anteriormente en la Sección 3.3, la homología persistente permite rastrear cómo estas características (como componentes conexas, bucles o cavidades) aparecen y desaparecen a medida que varía el parámetro de filtración r. Para representar y analizar esta evolución, se utilizan herramientas como los diagramas de persistencia, los códigos de barras y los paisajes de persistencia. Estas representaciones permiten visualizar y cuantificar la robustez de las características topológicas.

3.4.1. Diagrama de persistencia

Un diagrama de persistencia (*Persistence Diagram*) (Edelsbrunner *et al.* 2002) es una representación gráfica de las características topológicas que aparecen y desaparecen en una filtración. Formalmente, para cada dimensión k, el diagrama de persistencia D_k es un conjunto de puntos en el plano $\mathbb{R}^2$, donde cada punto (b, d) representa una característica topológica de dimensión k que nace en el valor de filtración b (*birth*) y muere en el valor d (*death*), con $b \leq d$.

- Nacimiento: Una característica topológica nace cuando un ciclo k-dimensional aparece en el grupo de homología H_k.
- Muerte: Una característica muere cuando el ciclo se convierte en una frontera (es decir, se "rellena" o se conecta con otras estructuras).

El intervalo de persistencia de una característica es el intervalo $[b, d)$, y su persistencia se define como $d - b$, que mide la "duración“ o robustez de la característica. Las características con mayor persistencia (es decir, intervalos largos) son consideradas más significativas, ya que persisten a través de múltiples escalas.

3.4.1.1. Distancias entre diagramas de persistencia

Para poder comparar formalmente distintos diagramas de persistencia, se definen métricas que permiten medir la disimilitud entre ellos. Las dos distancias más comunes en este contexto son:

- Distancia de *Bottleneck* (o del cuello de botella).
- Distancia de *Wasserstein*.

Estas distancias se basan en encontrar una correspondencia (emparejamiento) entre los puntos de dos diagramas de persistencia $D^{(1)}$ y $D^{(2)}$, permitiendo también emparejar puntos con la diagonal en caso de no encontrar una correspondencia directa.

3.4.1.1.1. Distancia de Bottleneck. La distancia de *Bottleneck* mide la mayor distancia entre pares de puntos emparejados en la mejor correspondencia posible. Formalmente:

$$dist_B(D^{(1)}, D^{(2)}) = \min_{\gamma: D^{(1)} \leftrightarrow D^{(2)}} \max_{(a,b)\in\gamma} \|a - b\|_\infty, \tag{3.8}$$

donde γ representa un emparejamiento entre los puntos de $D^{(1)}$y $D^{(2)}$, permitiendo emparejar con la diagonal, y $\|\cdot\|_\infty$ es la norma infinito (máximo valor absoluto entre coordenadas).

3.4.1.1.2. Distancia de Wasserstein. La distancia de *Wasserstein* de orden p generaliza la *Bottleneck*, considerando la suma de las distancias elevadas a la potencia p entre los puntos emparejados:

$$dist_W^p(D^{(1)}, D^{(2)}) = \left(\min_{\gamma: D^{(1)} \leftrightarrow D^{(2)}} \sum_{(a,b)\in\gamma} \|a - b\|^p \right)^{\frac{1}{p}}. \tag{3.9}$$

Esta métrica también permite emparejamientos con la diagonal y penaliza emparejamientos lejanos en función de la potencia p.

3.4.2. Código de barras

El código de barras (Ghrist, 2008) es una representación alternativa del diagrama de persistencia que muestra los intervalos de persistencia como barras horizontales. Cada barra corresponde a un intervalo $[b, d)$ en el diagrama de persistencia, donde:

- El extremo izquierdo de la barra está en b (nacimiento).
- El extremo derecho está en d (muerte).

3.4.3. Paisaje de persistencia

Una forma funcional de representar un diagrama de persistencia es mediante los paisajes de persistencia (*Persistence Landscapes*), introducidos por Bubenik (2015). Esta representación transforma el conjunto discreto de puntos del diagrama en una secuencia de funciones, lo que permite aplicar técnicas de análisis funcional y estadística en contextos como la comparación, agregación o clasificación de datos topológicos.

Sea un diagrama de persistencia $D = \{(b_i, d_i)\}_{i=1}^{N}$. A cada punto se le asocia una función triangular $f_{(b_i,d_i)}(t)$, definida por:

$$f_{(b_i,d_i)}(t) = \max\left(0, \min(t - b_i, d_i - t)\right),$$

la cual tiene soporte en (b_i, d_i), alcanza su máximo en el punto medio, y su altura es $(d_i - b_i)/2$.

El paisaje de persistencia es la colección de funciones $\Lambda = \{\lambda_k\}_{k=1}^{\infty}$, donde cada $\lambda_k(t)$ representa el k-ésimo valor más grande entre las funciones $\{f_{(b_i,d_i)}(t)\}$ evaluadas en $t \in \mathbb{R}$:

$$\lambda_k(t) = k\text{-ésimo valor más grande de } \left\{f_{(b_1,d_1)}(t), \ldots, f_{(b_N,d_N)}(t)\right\}.$$

3.4.4. Paisaje de persistencia promedio

Dado un conjunto de diagramas de persistencia $D^{(1)}, D^{(2)}, \ldots, D^{(m)}$, con sus correspondientes paisajes:

$$\Lambda^{(i)} = \{\lambda_k^{(i)}(t)\}_{k=1}^{\infty}, \quad i = 1, \ldots, m, \tag{3.10}$$

el paisaje de persistencia promedio (*Average Persistent Landscape*) se define como el promedio punto a punto de las funciones correspondientes:

$$\bar{\lambda}_k(t) = \frac{1}{m} \sum_{i=1}^{m} \lambda_k^{(i)}(t), \quad \forall k \in \mathbb{N},\ t \in \mathbb{R}. \tag{3.11}$$

3.4.5. Distancia entre paisajes

La distancia entre dos paisajes de persistencia $\Lambda = \{\lambda_k\}$ y $\Gamma = \{\gamma_k\}$ se puede calcular usando normas funcionales estándar, que miden la diferencia entre funciones.

Las dos formas más comunes son:

- Norma L^p: Para $1 \leq p < \infty$, la distancia se define como:

$$\|\Lambda - \Gamma\|_p = \left(\sum_{k=1}^{\infty} \int_{\mathbb{R}} |\lambda_k(t) - \gamma_k(t)|^p \, dt \right)^{1/p}. \tag{3.12}$$

- Norma suprema L^∞:

$$\|\Lambda - \Gamma\|_\infty = \sup_{k,t} |\lambda_k(t) - \gamma_k(t)|. \tag{3.13}$$

Estabilidad: Una propiedad importante de estas distancias es que son estables con respecto a la distancia *Bottleneck* entre diagramas de persistencia. Esto significa que:

$$\|\Lambda - \Gamma\|_\infty \leq dist_B(D^{(1)}, D^{(2)}), \tag{3.14}$$

donde $D^{(1)}$ y $D^{(2)}$ son los diagramas de persistencia que generan Λ y Γ, respectivamente. Esta propiedad asegura que pequeñas diferencias en los datos no producen grandes diferencias en los paisajes.

En este TFM se utiliza la norma L^2 para calcular la distancia entre paisajes de persistencia:

$$\|\Lambda - \Gamma\|_2 = \left(\sum_{k=1}^{\infty} \int_{\mathbb{R}} (\lambda_k(t) - \gamma_k(t))^2 \, dt \right)^{1/2}. \tag{3.15}$$

Capítulo 4

Metodología

Como se mencionó en el Capítulo 1, este TFM busca realizar un análisis topológico de las trayectorias de vuelo, basándose en la distancia (o desviación) entre las trayectorias de cada vuelo y el retraso asociado. Para ello, el procesamiento de las trayectorias se ha inspirado en el enfoque propuesto por Cuerno *et al.* (2023), incorporando además algunas modificaciones. Por otra parte, a lo largo del desarrollo del trabajo se han empleado diferentes versiones metodológicas, refinadas progresivamente en función de los resultados obtenidos.

4.1. Primera versión

4.1.1. Procesamiento de las trayectorias

En este apartado se detalla el procesamiento de los datos de las trayectorias de vuelo, que son la base para el análisis topológico posterior.

4.1.1.1. Representación de las trayectorias

Cada trayectoria de vuelo está representada como una secuencia ordenada de puntos definidos por el tiempo, latitud, longitud y altitud:

$$P = (t, L, l, a),$$

donde t es el instante de tiempo, L la latitud, l la longitud y a la altitud.

Por cada vuelo se consideran dos trayectorias:

- La trayectoria planificada $T_p = \{(t_j, L_j, l_j, a_j)\}_{j \in J}$, correspondiente a la ruta prevista para ese vuelo.
- La trayectoria real $T_r = \{(t_i, L_i, l_i, a_i)\}_{i \in I}$, que corresponde al recorrido efectivamente realizado por la aeronave.

Aquí, J e I representan los conjuntos de índices de los puntos que componen cada trayectoria, respectivamente.

4.1.1.2. Propiedades de las trayectorias

Estas dos trayectorias suelen cumplir ciertas propiedades:

- Ambas comienzan y finalizan en las mismas coordenadas geográficas, correspondientes a los aeropuertos de origen y destino:

$$(L_{j_0}, l_{j_0}) = (L_{i_0}, l_{i_0}), \quad (L_{j_f}, l_{j_f}) = (L_{i_f}, l_{i_f}).$$

- La altitud en los puntos inicial y final es cero, ya que las aeronaves despegan y aterrizan en tierra:

$$a_{j_0} = a_{j_f} = a_{i_0} = a_{i_f} = 0.$$

- Normalmente, la trayectoria real contiene una mayor cantidad de puntos temporales que la planificada, por lo que sus instantes de tiempo no coinciden necesariamente:

$$\#\{t_i\} > \#\{t_j\}.$$

4.1.1.3. Unificación de instantes de tiempo e interpolación

Para evaluar la distancia (desviación) entre la trayectoria planificada y la real, es necesario comparar ambas trayectorias en instantes temporales coincidentes. La idea es calcular la distancia entre la posición en la que se esperaba que estuviera el vuelo (según la planificación) y la posición real en la que se encontraba en el mismo instante de tiempo. Al repetir este proceso para cada momento en el que están definidas las dos trayectorias, se obtiene una medida de las distancias a lo largo del tiempo. La suma de estas distancias permite cuantificar la desviación total de una trayectoria respecto a la otra.

Por consiguiente, para comparar ambas trayectorias punto a punto, es necesario que estén definidas en los mismos instantes temporales. Para ello, se define un conjunto unificado de instantes de tiempo para cada vuelo, que corresponde a la unión de los instantes presentes en ambas trayectorias:

$$\{t_k\}_{k \in K} = \{t_j\}_{j \in J} \cup \{t_i\}_{i \in I},$$

donde,

$$t_0 = \text{mín}(t_{j_0}, t_{i_0}), \quad t_{k_f} = \text{máx}(t_{j_f}, t_{i_f}).$$

Una vez hecho esto, para hacer que las dos trayectorias estén definidas en todos los puntos de t_k se hace una interpolación lineal en el tiempo. Es decir, para valores no definidos en las trayectorias, se realiza una interpolación lineal de la siguiente manera:

1. Si para un tiempo t_{k_l} no existe un punto en la trayectoria planificada T_p, igual para la trayectoria real, es decir, $t_{k_l} \notin \{t_j\}_{j\in J}$, pero si existe el punto anterior y el siguiente, es decir, $t_{k_{l-1}}, t_{k_{l+1}} \in \{t_j\}_{j\in J}$, se crea un nuevo punto: $p = (t_{k_l}, L_{k_l}, l_{k_l}, a_{k_l})$, mediante una interpolación lineal de los puntos anterior y siguiente.
2. De otra manera, si se tiene el punto $t_{k_{l-1}} \in \{t_j\}_{j\in J}$, pero el siguiente no está en la trayectoria planificada, igual para la trayectoria real, es decir, $t_{k_{l+1}} \notin \{t_j\}_{j\in J}$, entonces se busca el primer tiempo t_{k_p} que esté definido en la trayectoria planificada tal que $k_p > k_l$, y se interpola k_{l-1} con k_p para obtener todos los puntos faltantes entre k_{l-1} y k_p
3. Si los tiempos de la trayectoria planificada, igual para la trayectoria real, no cubren completamente el intervalo de la trayectoria real, es decir, si: $t_{j0} > t_0$ y $t_{jf} < t_{kf}$, se establecen los valores de los extremos de la trayectoria planificada de la siguiente forma: $(t_k, L_k, l_k, a_k) = (t_k, L_{j0}, l_{j0}, a_{j0}), \quad$ para $k < j_0$ y $(t_k, L_k, l_k, a_k) = (t_k, L_{jf}, l_{jf}, a_{jf}), \quad$ para $k > j_f$.

4.1.1.4. Cálculo de la distancia entre trayectorias

Una vez definidas ambas trayectorias en los mismos instantes de tiempo, se calcula la distancia o desviación entre ellas. Para esto, se utiliza la distancia de Haversine para las coordenadas geográficas (latitud y longitud) y la diferencia absoluta para la altitud. Así, para cada instante de tiempo t_k, la distancia entre los puntos correspondientes de las trayectorias planificada y real se define como:

$$\text{dist}_{t_k}\left((L_i, l_i, a_i), (L_j, l_j, a_j)\right) = \sqrt{\text{dist}_H\left((L_i, l_i), (L_j, l_j)\right)^2 + |a_i - a_j|^2}, \tag{4.1}$$

donde dist_H es la distancia de Haversine entre las coordenadas geográficas. Tanto la distancia Haversine como la diferencia de altitud se calcula en kilómetros (km), de modo que la distancia total para cada punto se expresa en km.

La distancia total o desviación entre las trayectorias se calcula como la suma de las distancias en cada punto a lo largo de toda la trayectoria:

$$\text{dist}_{\text{Desv}}(T_p, T_r) = \sum_{l=0}^{k_f} \text{dist}_{t_l}((L_{j_l}, l_{j_l}, a_{j_l}), (L_{i_l}, l_{i_l}, a_{i_l})) \tag{4.2}$$

También, se calculan otras métricas de distancia como la distancia máxima, que es el valor máximo entre todas las distancias calculadas en cada punto de la trayectoria (máximo valor de $dist_{tk}$), así como también la distancia media, que es la media de la suma de las distancias en cada instante de tiempo y la desviación típica de las distancias en cada instante de tiempo de la trayectoria. Estas se pueden definir como:

- Distancia máxima: el valor máximo de las distancias puntuales a lo largo de la trayectoria:

$$\text{dist}_{\text{max}}(T_p, T_r) = \max_{0\le l\le k_f} \text{dist}_{t_l}((L_{j_l}, l_{j_l}, a_{j_l}), (L_{i_l}, l_{i_l}, a_{i_l})). \tag{4.3}$$

- Distancia media: la media aritmética de las distancias puntuales:

$$\text{dist}_{\text{media}}(T_p, T_r) = \frac{\text{dist}_{\text{Desv}}(T_p, T_r)}{k_f + 1} \tag{4.4}$$

- Desviación típica: la desviación estándar de las distancias en cada punto:

$$\sigma_{\text{dist}}(T_p, T_r) = \sqrt{\frac{1}{k_f + 1} \sum_{l=0}^{k_f} \left(\text{dist}_{t_l}((L_{j_l}, l_{j_l}, a_{j_l}), (L_{i_l}, l_{i_l}, a_{i_l})) - \text{dist}_{\text{media}}(T_p, T_r)\right)^2} \tag{4.5}$$

Aquí, k_f denota el índice correspondiente al último instante de tiempo del conjunto unificado $\{t_k\}_{k \in K}$. Dado que los índices temporales k se enumeran desde 0 hasta k_f, el número total de instantes considerados es $k_f + 1$. Por esta razón, en el cálculo de la media y la desviación estándar se divide entre $k_f + 1$ y no solo entre k_f.

4.1.2. Creación de la nube de puntos

Obtenidas las métricas que cuantifican la desviación entre la trayectoria planificada y la trayectoria real para cada vuelo, es necesario representar estos valores de forma estructurada para aplicar técnicas de análisis topológico de datos. Para ello, se construye una nube de puntos para cada combinación de día y aeropuerto.

Cada punto de estas nubes representa un vuelo registrado en un aeropuerto determinado en un día concreto. Así, si se dispone de datos correspondientes a N aeropuertos durante D días, se obtienen en total $N \times D$ nubes de puntos, cada una conteniendo un número de puntos igual a la cantidad de vuelos realizados en dicha combinación.

Representación del vector de características

Para cada vuelo, se construye un vector de características en un espacio de cinco dimensiones, compuesto por las siguientes variables:

- $\text{dist}_{\text{Desv}}$: desviación total de la trayectoria.
- dist_{max}: distancia máxima entre la trayectoria planificada y la real.
- $\text{dist}_{\text{media}}$: distancia media.
- σ_{dist}: desviación típica de las distancias.
- $r = t_{i_f} - t_{j_f}$: retraso del vuelo, medido en minutos como la diferencia entre el tiempo real de llegada y el planificado.

Se consideran dos formas distintas de construir estos vectores, con el objetivo de ver cómo afecta la representación de los datos a la posterior clasificación de los aeropuertos:

- Caso 1 – Todas las distancias positivas: Se utiliza un vector de cinco dimensiones con las variables mencionadas como punto de la trayectoria, manteniendo siempre valores positivos para las distancias, independientemente del retraso.

$$\text{Para cada vuelo: } x = (\text{dist}_{\text{Desv}}, \text{dist}_{\text{max}}, \text{dist}_{\text{media}}, \sigma_{\text{dist}}, r) \in \mathbb{R}^5$$

- Caso 2 – Posibles distancias negativas: Se usa la misma estructura que en el caso anterior, pero si el retraso es negativo ($r < 0$), las distancias también se transforman a valores negativos.

$$\text{Para cada vuelo: } x = \begin{cases} (\text{dist}_{\text{Desv}}, \text{dist}_{\text{max}}, \text{dist}_{\text{media}}, \sigma_{\text{dist}}, r)\,, & \text{si } r \geq 0 \\ (-\text{dist}_{\text{Desv}}, -\text{dist}_{\text{max}}, -\text{dist}_{\text{media}}, -\sigma_{\text{dist}}, r)\,, & \text{si } r < 0 \end{cases}$$

4.1.3. Análisis topológico de las nubes de puntos mediante homología persistente

Una vez obtenidas las nubes de puntos, se aplica a cada una de ellas un análisis topológico de datos utilizando homología persistente en dimensión cero (H_0). Para ello, se utiliza la filtración de Vietoris-Rips, que permite construir complejos simpliciales a partir de los datos, generando como resultado un diagrama de persistencia para cada nube.

Dado que se tiene una nube de puntos por cada día y por cada aeropuerto, el total de diagramas de persistencia generados será $N \times D$. Cada uno de estos diagramas resume la evolución de los componentes conexos en función del parámetro de filtración.

A continuación, se calcula el paisaje de persistencia (*Persistence Landscape*, PL) asociado a cada diagrama. Este paisaje transforma el diagrama de persistencia en una función matemática, lo que facilita su comparación y análisis cuantitativo.

Para medir la similitud entre paisajes de persistencia, se utiliza la norma L_2. En particular, se calcula la distancia entre cada par de paisajes de persistencia de todos los aeropuertos en el mismo día. Como resultado, para cada día $k = 1, \ldots, D$, se construye una matriz de distancias M_k de tamaño $N \times N$, donde cada entrada $M_k(i, j)$ representa la distancia L_2 entre los paisajes de persistencia de los aeropuertos i y j en ese día. Por lo que, en total se obtienen D matrices de distancia: $M_1, M_2, \ldots, M_D$.

Posteriormente, se calcula una matriz de distancias promedio entre aeropuertos a lo largo de los días. Esta matriz promedio se define como:

$$\bar{M}_{i,j} = \frac{1}{M_{i,j}^{\text{cont}}} \sum_{k=1}^{D} M_k(i, j), \tag{4.6}$$

donde $M_{i,j}^{\text{cont}}$ es el número de días en los que ambos aeropuertos i y j tienen datos válidos (es decir, vuelos registrados). Si alguno de los dos aeropuertos no tiene vuelos en un determinado día, ese día se excluye del promedio.

Como resultado, se obtiene una matriz $\bar{M} \in \mathbb{R}^{N \times N}$ que representa la distancia promedio entre aeropuertos basada en la topología de las trayectorias de vuelo a lo largo del tiempo.

4.1.4. Clasificación

Con la matriz promedio de distancias $\bar{M}$ calculada, el siguiente paso es agrupar los aeropuertos en clústeres de comportamiento similar. Para evaluar el número de clústeres por el que esta representado la matriz de distancia, se utiliza la métrica conocida

como Silhouette Score (Rousseeuw, 1987), que evalúa que tan bien se ha agrupado cada aeropuerto dentro de su clúster. Para esto, el algoritmo compara la distancia promedio de un aeropuerto con los de su propio grupo frente a la distancia promedio con los de otros grupos. Un valor alto de Silhouette Score indica que el aeropuerto está bien clasificado dentro de su clúster.

Una vez evaluado el número de clústeres, se aplica el algoritmo de Agrupamiento Aglomerativo (*Agglomerative Clustering*). Este método jerárquico parte considerando cada aeropuerto como un clúster independiente y, en cada iteración, une los pares de clústeres más cercanos hasta alcanzar el número deseado de grupos.

Se ha elegido este algoritmo de clustering por las siguientes razones:

- Trabaja directamente con matrices de distancia, sin requerir que los datos estén embebidos en un espacio vectorial.
- Permite especificar explícitamente el número de clústeres deseado.

Dado que el algoritmo opera directamente sobre $\bar{M}$, no se requiere ninguna transformación adicional de los datos.

Finalmente, se analiza la clasificación obtenida para los aeropuertos españoles en cada uno de los casos considerados. Se comparan las diferentes clasificaciones, se examinan sus coincidencias y divergencias, y se discute la validez de los resultados.

4.2. Segunda versión

En la primera versión de la metodología, la comparación entre la trayectoria planificada y la trayectoria real se realiza directamente, considerando los instantes temporales originales de cada una. Sin embargo, esto implica que las diferencias en el tiempo de los vuelos (por ejemplo, retrasos en la salida o llegada) influyan directamente en el cálculo de las métricas de distancia, aumentando la distancia entre trayectorias incluso cuando estas son muy parecidas. Es decir, si un vuelo presenta un retraso muy grande, la desviación entre sus trayectorias sería, a su vez, muy grande, a pesar de que las trayectorias en sí sean muy parecidas. Por lo tanto, se puede decir que el retraso del vuelo también se refleja en la desviación que se está calculando.

Como ya se está tomando en cuenta el retraso como una variable independiente, para evitar que la diferencia de tiempos influya en el cálculo de la distancia, se propone una segunda versión de la metodología en la cual se ajusta la escala temporal de la trayectoria real antes de igualar lo instantes de tiempo (véase la Sección 4.1.1.3), de forma que su instante inicial y final coincidan con los de la trayectoria planificada. Con esta transformación, ambas trayectorias se alinean temporalmente, y cualquier desviación calculada se debe únicamente a diferencias espaciales entre ambas trayectorias, y no a desajustes temporales.

Para realizar esto, para cada vuelo se calcula una recta que pasa por dos puntos, los puntos iniciales y finales de la trayectoria, es decir, se calcula una recta afín definida por los extremos temporales de la trayectoria real y planificada, con el objetivo de ajustar todos los tiempos intermedios de la trayectoria real a la escala temporal de la planificada y haciendo coincidir sus extremos temporales. Específicamente, se considera una transformación lineal del tiempo de la forma:

$$t_i^{\text{ajustado}} = m \times t_i + n,$$

donde t_i representa el tiempo original del punto i-ésimo de la trayectoria real, y t_i^{ajustado} es su correspondiente tiempo ajustado.

Para determinar los parámetros m y n, se imponen las siguientes condiciones de ajuste:

$$\begin{cases} t_{j0} = m \times t_{i0} + n \\ t_{jf} = m \times t_{if} + n \end{cases},$$

donde, t_{j0} y t_{jf} son los tiempos inicial y final de la trayectoria planificada, y t_{i0} y t_{if} son los tiempos inicial y final de la trayectoria real.

Resolviendo este sistema de ecuaciones se obtiene:

$$m = \frac{t_{jf} - t_{j0}}{t_{if} - t_{i0}}, \quad n = t_{j0} - m \times t_{i0}.$$

Una vez calculados los parámetros m y n, se aplica la transformación a todos los puntos de la trayectoria real. Sea cada punto de la trayectoria real representado como una tupla $[t_i, L_i, l_i, a_i]$, el punto ajustado se obtiene como:

$$[t_i^{\text{ajustado}}, L_i, l_i, a_i],$$

donde:

$$t_i^{\text{ajustado}} = m \times t_i + n.$$

Este proceso hace que todos los puntos ajustados estén alineados con la escala temporal de la trayectoria planificada. Luego, una vez ajustada la escala temporal, se procede como en la primera versión: se interpolan ambas trayectorias en instantes de tiempo comunes (véase la Sección 4.1.1.3) y se calcula la desviación punto a punto.

4.3. Tercera versión

En esta tercera versión, el objetivo sigue siendo ajustar los instantes temporales de la trayectoria real para que coincidan con los de la trayectoria planificada, evitando que las diferencias de tiempo (por ejemplo, retrasos) influyan directamente en el cálculo de la distancia entre ambas trayectorias.

A diferencia de la versión anterior (véase la Sección 4.2), en la cual se realiza un ajuste lineal entre los tiempos iniciales y finales, esta versión propone una aproximación más precisa basada en el progreso relativo a lo largo de la ruta. Esto se debe a que una simple transformación lineal no tiene en cuenta otros posibles aspectos como puede ser las diferencias en velocidad entre ambas trayectorias. Por ejemplo, si la trayectoria real recorre la ruta más rápido o más lento que la planificada, los puntos

intermedios no estarán bien alineados, y se generará una desviación mas grande debida al desfase de los tiempos, no a la forma real de la trayectoria.

4.3.1. Cálculo del porcentaje de progreso

Para realizar esto, primero hay que calcular el porcentaje de progreso de cada punto de ambas trayectorias. Este porcentaje indica qué parte del recorrido total se ha completado en cada instante. Por ejemplo, si un punto tiene un progreso de 0.25, significa que está ubicado aproximadamente al 25 % del trayecto total.

Dadas las dos trayectorias:

- La trayectoria planificada $T_p = \{(t_j, L_j, l_j, a_j)\}_{j \in J}$
- La trayectoria real $T_r = \{(t_i, L_i, l_i, a_i)\}_{i \in I}$

Primero, se calcula la distancia entre cada par de puntos consecutivos, asignando una distancia de cero al punto inicial. Este cálculo se realiza utilizando la misma fórmula empleada para la desviación (véase la Sección 4.1.1.4), pero expresando el resultado en metros (m) en lugar de kilómetros (km). De esta manera, el cálculo de las distancias para la trayectoria planificada es:

$$d_j = \sqrt{dist_{\mathrm{H}}((L_j, l_j), (L_{j-1}, l_{j-1}))^2 + |a_j - a_{j-1}|^2}, \quad \forall j \in J,\ j > 0, \tag{4.7}$$

y para la trayectoria real:

$$d_i = \sqrt{dist_{\mathrm{H}}((L_i, l_i), (L_{i-1}, l_{i-1}))^2 + |a_i - a_{i-1}|^2}, \quad \forall i \in I,\ i > 0. \tag{4.8}$$

Luego, se calcula la distancia acumulada hasta cada punto:

$$D_j = \sum_{k=1}^{j} d_k \quad \forall j \in J,\ j > 0 \quad \text{y} \quad D_i = \sum_{k=1}^{i} d_k \quad \forall i \in I,\ i > 0. \tag{4.9}$$

Finalmente, el porcentaje de progreso para cada punto se define como:

$$p_j = \frac{D_j}{D_{|J|}} \quad \forall j \in J,\ j > 0 \quad \text{y} \quad p_i = \frac{D_i}{D_{|I|}} \quad \forall i \in I,\ i > 0, \tag{4.10}$$

donde $D_{|J|}$ y $D_{|I|}$ representan la distancia total de las trayectorias planificada y real, respectivamente.

4.3.2. Ajuste temporal basado en el porcentaje de progreso

Como resultado del cálculo anterior, se obtienen dos listas:

- $\{p_j\}_{j \in J}$: porcentaje acumulado de progreso para cada punto de la trayectoria planificada.
- $\{p_i\}_{i \in I}$: porcentaje acumulado de progreso para cada punto de la trayectoria real.

El objetivo es reasignar un nuevo instante de tiempo a cada punto de la trayectoria real, en función de su porcentaje de progreso p_i, de modo que se corresponda con los tiempos de la trayectoria planificada. Para esto, se procede del siguiente modo:

Para cada punto $(t_i, L_i, l_i, a_i) \in T_r$, se toma su valor de progreso p_i, y se busca un intervalo en la lista de progresos de la trayectoria planificada tal que:

$$p_j \leq p_i \leq p_{j+1}, \quad \text{con } j \in J. \tag{4.11}$$

Una vez encontrado este intervalo, se interpola linealmente entre los tiempos planificados correspondientes t_j y t_{j+1}, de la siguiente forma:

$$t_i^{(ajustado)} = t_j + \left(\frac{p_i - p_j}{p_{j+1} - p_j} \right) \times (t_{j+1} - t_j). \tag{4.12}$$

Este proceso se repite para todos los puntos $i \in I$, generando una nueva trayectoria real ajustada temporalmente, donde las marcas de tiempo han sido modificadas, pero las posiciones espaciales (L_i, l_i, a_i) se mantienen.

4.4. Especificación de aportaciones realizadas a la metodología

Como se mencionó previamente, este TFM se basa en la metodología propuesta por Cuerno *et al.* (2023). No obstante, se han introducido una serie de modificaciones al enfoque original. A continuación, se detallan las aportaciones metodológicas realizadas en el desarrollo de este TFM:

4.4.1. Representación de la nube de puntos

En el estudio realizado por Cuerno *et al.* (2023), cada vuelo es representado en la nube de puntos mediante dos variables: la distancia total entre la trayectoria planificada y la trayectoria real, y el retraso de llegada. Además, las distancias se transforman a valores negativos cuando el retraso asociado al vuelo es negativo.

En este TFM (véase la Sección 4.1.2), en vez de utilizar solo la distancia total entre la trayectoria planificada y real y el retraso, cada vuelo se representa como un punto en un espacio de cinco dimensiones, considerando las siguientes variables: distancia total entre trayectoria planificada y real, distancia media entre trayectorias, distancia máxima, desviación típica de las distancias y retraso de llegada.

Asimismo, se implementaron dos variantes en la representación de los datos: una en la que todas las distancias permanecen positivas, independientemente del retraso, y otra en la que las distancias se convierten en negativas cuando el retraso es negativo. Esta doble representación permite analizar cómo afecta el tratamiento del retraso a la estructura topológica resultante.

4.4.2. Cálculo de distancia entre aeropuertos

En la metodología original, una vez aplicado el análisis topológico y obtenido los diagramas y paisajes de persistencia por día y aeropuerto, la distancia entre aeropuertos

se calcula a partir del paisaje de persistencia promedio. Este se obtiene promediando los paisajes diarios de un mismo aeropuerto. Posteriormente, se calcula la distancia entre dos aeropuertos como la distancia entre sus respectivos paisajes promedio, utilizando la norma suprema.

En este TFM (véase la Sección 4.1.1.4), se propone una alternativa metodológica. En lugar de promediar previamente los paisajes de persistencia, se calcula la distancia entre aeropuertos como el promedio de las distancias entre todos los pares de paisajes de persistencia generados en los días en que ambos aeropuertos están definidos.

4.4.3. Clasificación de aeropuertos

El estudio realizado por Cuerno *et al.* (2023), no llegó a hacer una clasificación de aeropuertos, si no que unicamente calculó las distancias entre aeropuertos, tal y como se indicó anteriormente (véase la Sección 4.4.2). A partir de los valores obtenidos, comparan las distancias entre aeropuertos pertenecientes al mismo grupo según la clasificación de AENA, evaluando los casos en los que ciertos aeropuertos no se comportaban como el resto de su grupo (es decir, aquellos que presentaban una distancia considerablemente mayor respecto a los demás aeropuertos del mismo grupo).

Sin embargo, en este TFM se amplió el análisis, combinando el análisis topológico con algoritmos de aprendizaje automático para realizar una clasificación de aeropuertos a partir de las distancias calculadas entre ellos. Esto permite identificar de manera más detallada las discrepancias y los aeropuertos que presentan mayor similitud entre sí, según las características topológicas derivadas del comportamiento de sus vuelos.

4.4.4. Cálculo de la distancia entre trayectorias

En la primera versión de este TFM, el cálculo de la distancia entre la trayectoria planificada y la real de un vuelo se realizó siguiendo el mismo procedimiento que en Cuerno *et al.* (2023), sin ajustes temporales específicos.

Sin embargo, en versiones posteriores de la metodología (véase las Secciones 4.2 y 4.3), se introdujo un refinamiento consistente en el ajuste de las marcas temporales de las trayectorias. Este ajuste permite alinear de forma más precisa los puntos correspondientes entre la trayectoria planificada y la real, minimizando el efecto de variables como el retraso, las diferencias de tiempo o la velocidad en el cálculo de la distancia entre trayectorias. Así, se obtiene una medida más representativa de la desviación espacial pura entre ambas trayectorias.

4.5. Herramientas de implementación

La implementación de la metodología y el análisis de datos se realizó utilizando el lenguaje de programación Python, apoyado en diversas bibliotecas. A continuación, se detallan las principales herramientas utilizadas:

- *scikit-learn* (*sklearn*): Para la aplicación de técnicas de clustering, como *Agglomerative Clustering*, así como para el uso de la métrica *Silhouette Score*.
- *ripser*: Para el cálculo de la filtración de Vietoris-Rips y el cálculo de diagramas de persistencia a partir de complejos de Vietoris–Rips.

- *persim*: Para la obtención y visualización de paisajes de persistencia y visualización de diagramas de persistencia. Se emplearon las funciones *plot_diagrams*, *PersLandscapeExact* y *plot_landscape_simple*.
- *seaborn*: Para la creación de diagramas de cajas.
- *matplotlib*: Librería base para todas las visualizaciones generadas en el proyecto.
- *pandas* y *NumPy*: Utilizadas extensamente para el manejo, transformación y análisis de datos.
- *folium*: Utilizada para la visualización de trayectorias sobre mapas.

Capítulo 5

Resultados

5.1. Análisis y limpieza de datos

5.1.1. Conjunto de datos

Para realizar el estudio, se cuenta con los mismos datos de trayectorias utilizados en el artículo Cuerno *et al.* (2023), el cual se puede consultar en Cuerno (2025). Este conjunto de datos contiene los ficheros correspondientes a las trayectorias planificadas y reales de vuelos que tuvieron como origen o destino un aeropuerto español durante un periodo de 217 días en el año 2018, específicamente desde el 25 de marzo hasta el 27 de octubre de ese año. El conjunto de datos está dividido en dos ficheros: uno que contiene todas las trayectorias planificadas, organizadas por día, y otro que incluye todas las trayectorias reales, también divididas por día.

5.1.2. Representación de aeropuertos

Los aeropuertos están representados mediante su código ICAO (*International Civil Aviation Organization*), que consta de cuatro letras y es establecido por la OACI, la organización de aviación civil internacional. Específicamente, en la Tabla 5.3 se pueden ver todos los aeropuertos españoles sobre los cuales se ha aplicado el análisis, junto con sus códigos ICAO correspondientes y las ciudades a las que pertenecen.

5.1.3. Limpieza de datos

Al analizar el conjunto de datos y comenzar con el estudio de las trayectorias, se detectaron casos de trayectorias erróneas y duplicadas. Por esta razón, se decidió realizar una limpieza de datos antes de aplicar la metodología. Esta limpieza consistió en eliminar los vuelos cuyas trayectorias fueron identificadas como incorrectas, así como aquellos considerados redundantes para el análisis. A continuación, se describen los tipos de casos encontrados en el conjunto de datos:

- Vuelos que fueron cancelados.
- Vuelos que despegaron pero no llegaron a su destino (desviados).
- Vuelos con marcas de tiempo duplicadas, es decir, con dos puntos registrados en el mismo instante pero con coordenadas geográficas diferentes.

- Vuelos cuyo aeropuerto de origen y destino es el mismo.
- Vuelos cuyo aeropuerto de origen o destino está identificado con el código ZZZZ, el cual se utiliza cuando el lugar de salida o llegada no corresponde a un aeropuerto registrado, ya que puede tratarse de un vuelo militar.
- Vuelos duplicados, es decir, que aparecen más de una vez en el conjunto de datos.

Los vuelos cancelados, los que no llegaron a su destino, aquellos con marcas de tiempo duplicadas, los que tienen el mismo aeropuerto de origen y destino y los que utilizan el código ZZZZ fueron eliminados del conjunto de datos. En cuanto a los vuelos duplicados, estos pueden deberse a distintos casos:

1. Vuelos planificados para despegar un día y llegar al día siguiente.
2. Vuelos que debían despegar un día y llegar al siguiente, pero que despegaron antes de lo previsto y terminaron saliendo y llegando el mismo día (día anterior).
3. Vuelos que debían despegar un día y llegar al siguiente, pero que despegaron con retraso y finalmente salieron y llegaron el mismo día (día siguiente).
4. Vuelos planificados para despegar y llegar el mismo día, pero que salieron con retraso y llegaron al día siguiente.
5. Vuelos planificados para despegar y llegar el mismo día, pero que despegaron el día anterior y llegaron al día siguiente.

Como se puede observar, estos vuelos aparecen duplicados porque estaban planificados para despegar y llegar en días distintos, o porque en la práctica despegaron y aterrizaron en días diferentes. Esto provoca que el mismo vuelo aparezca en los registros de ambos días dentro del conjunto de datos.

En el archivo de vuelos planificados, la trayectoria planificada de estos vuelos es siempre la misma, ya que se trata del mismo vuelo. Sin embargo, en el archivo de vuelos reales, la trayectoria puede ser idéntica o diferente. Debido a esto, cuando existen trayectorias reales distintas para un vuelo duplicado, se ha decidido conservar aquella con mayor número de puntos registrados.

Los vuelos que despegan y aterrizan en días diferentes, se han clasificado según su día real de llegada. Así, en los casos 1, 3, 4 y 5, los vuelos se consideran como pertenecientes al día siguiente, ya que ese es su día real de llegada. En el caso 2, el vuelo se asigna al día anterior, pues aunque estaba planificado para llegar un día después, en realidad aterrizó antes de lo previsto. Esta asignación se realiza porque, como se explicó en la Sección 4.1.2, el análisis topológico de trayectorias se realiza por separado para cada día y cada aeropuerto, por lo que los vuelos deben estar asociados a un día específico.

5.1.3.1. Porcentajes de vuelos eliminados en cada caso

En las Tablas 5.1 y 5.2 se presentan las métricas relacionadas con el número de vuelos en cada uno de los casos mencionados anteriormente. La Tabla 5.1 muestra el número total de vuelos en el conjunto de datos, la cantidad de vuelos eliminados, el total de vuelos restantes tras la limpieza y el número de vuelos en cada categoría.

Por su parte, la Tabla 5.2 presenta esos mismos valores expresados como porcentaje con respecto al tamaño original del conjunto de datos.

En la Tabla 5.1 se observa que el conjunto de datos inicial contiene 1,134,813 vuelos. Tras eliminar los vuelos cancelados, aquellos que no llegaron a destino, los que tienen instantes de tiempo duplicados, los que tienen el mismo aeropuerto de origen y destino, los que tienen como código de aeropuerto *ZZZZ* y los vuelos duplicados, se eliminaron un total de 36,241 vuelos, lo que representa un 3.1 % de los datos, como se indica en la Tabla 5.2. Finalmente, el conjunto de datos quedó con 1,098,572 vuelos, equivalentes al 96.80 % del total inicial.

Además, en la Tabla 5.1, los casos 6 y 7 corresponden a vuelos cuyo día de llegada no se encuentra dentro del rango de fechas del análisis. Es decir, el caso 6 representa vuelos que llegaron un día antes del inicio del análisis (25 de marzo de 2018), por lo que fueron eliminados. El caso 7 corresponde a vuelos que llegaron después del último día del análisis (27 de octubre de 2018), y también fueron eliminados, ya que, como se mencionó, los vuelos se registran según su fecha de llegada.

Tabla 5.1: Número de vuelos en el conjunto de datos, número de vuelos eliminados y número de vuelos por categoría. Primera limpieza.

Categoría	**Total**
Total Vuelos Iniciales	1,134,813
Total Vuelos Eliminados	36,241
Total Vuelos Resultantes	1,098,572
0 - Vuelos Cancelados	25
1 - Vuelos que no llegaron a su destino (desviados)	1,518
2 - Vuelos con instantes de tiempo duplicados	203
3 - Vuelos que tienen el mismo origen y destino	4,219
4 - Vuelos que tienen como aeropuerto destino o origen "ZZZZ"	202
5 - Vuelos eliminados porque están duplicados	29,881
6 - Vuelos que llegaron el día antes que el día de inicio	5
7 - Vuelos que llegaron el día después del día de fin	188

Tabla 5.2: Porcentaje de vuelos resultantes en el conjunto de datos, porcentaje de vuelos eliminados y porcentaje de vuelos por categoría. Primera limpieza.

Categoría	**Porcentaje**
0 - Vuelos Cancelados	0.0022 %
1 - Vuelos que no llegaron a su destino (desviados)	0.1337 %
2 - Vuelos con instantes de tiempo duplicados	0.0178 %
3 - Vuelos que tienen el mismo origen y destino	0.3717 %
4 - Vuelos que tienen como aeropuerto destino o origen "ZZZZ"	0.0178 %
5 - Vuelos eliminados porque están duplicados	2.6331 %
6 - Vuelos que llegaron el día antes que el día de inicio	0.0004 %
7 - Vuelos que llegaron el día después del día de fin	0.0165 %
Total Vuelos Eliminados	3.1935 %
Total Vuelos Resultantes	96.8064 %

Tabla 5.3: Lista de aeropuertos españoles empleados en el análisis, con su respectivo código ICAO y la ciudad donde se encuentran.

Código	Nombre del Aeropuerto	Ciudad
LEAB	Albacete	Albacete
LEAL	Alicante-Elche Miguel Hernández	Alicante
LEAM	Almería	Almería
LEAS	Asturias	Oviedo
LEBA	Córdoba	Córdoba
LEBB	Bilbao	Bilbao
LEBG	Burgos-Villafría	Burgos
LEBZ	Badajoz	Badajoz
LEBL	Josep Tarradellas Barcelona-El Prat	Barcelona
LECO	A Coruña	A Coruña
LECU	Madrid-Cuatro Vientos	Madrid
LEGE	Girona-Costa Brava	Girona
LEGR	Federico García Lorca Granada-Jaén	Granada
LEHC	Huesca-Pirineos	Huesca
LEIB	Ibiza	Ibiza
LEJR	Jerez	Jerez de la Frontera
LELL	Sabadell	Sabadell
LELN	León	León
LEMD	Adolfo Suárez Madrid-Barajas	Madrid
LEMG	Málaga-Costa del Sol	Málaga
LEMH	Menorca	Mahón
LEPA	Palma de Mallorca	Palma de Mallorca
LEPP	Pamplona	Pamplona
LERS	Reus	Reus
LERJ	Logroño-Agoncillo	Logroño
LESA	Salamanca	Salamanca
LESB	Son Bonet	Palma de Mallorca
LESO	San Sebastián	San Sebastián
LEST	Santiago-Rosalía de Castro	Santiago de Compostela
LEVC	Valencia	Valencia
LEVD	Valladolid	Valladolid
LEVT	Vitoria	Vitoria-Gasteiz
LEVX	Vigo	Vigo
LEZL	Sevilla	Sevilla
LEXJ	Santander-Seve Ballesteros	Santander
LEZG	Zaragoza	Zaragoza
GCFV	Fuerteventura	Puerto del Rosario
GCGM	La Gomera	San Sebastián de La Gomera
GCHI	El Hierro	Valverde
GCLA	La Palma	Santa Cruz de La Palma
GCLP	Gran Canaria	Las Palmas de Gran Canaria
GCRR	César Manrique-Lanzarote	Arrecife
GCTS	Tenerife Sur	Granadilla de Abona
GCXO	Tenerife Norte	San Cristóbal de La Laguna
GEML	Melilla	Melilla

5.1.4. Diagramas de valores obtenidos en la primera versión de la metodología

Tras la limpieza de datos y el cálculo de la distancia entre trayectorias de vuelo, así como de los retrasos asociados, se elaboraron diagramas de cajas para visualizar la distribución de valores de cada variable, de acuerdo con la primera versión de la metodología. En las Figuras 5.1 y 5.2 se muestran los diagramas correspondientes

a las variables de distancia y retraso, calculadas para cada vuelo y agrupadas por aeropuerto.

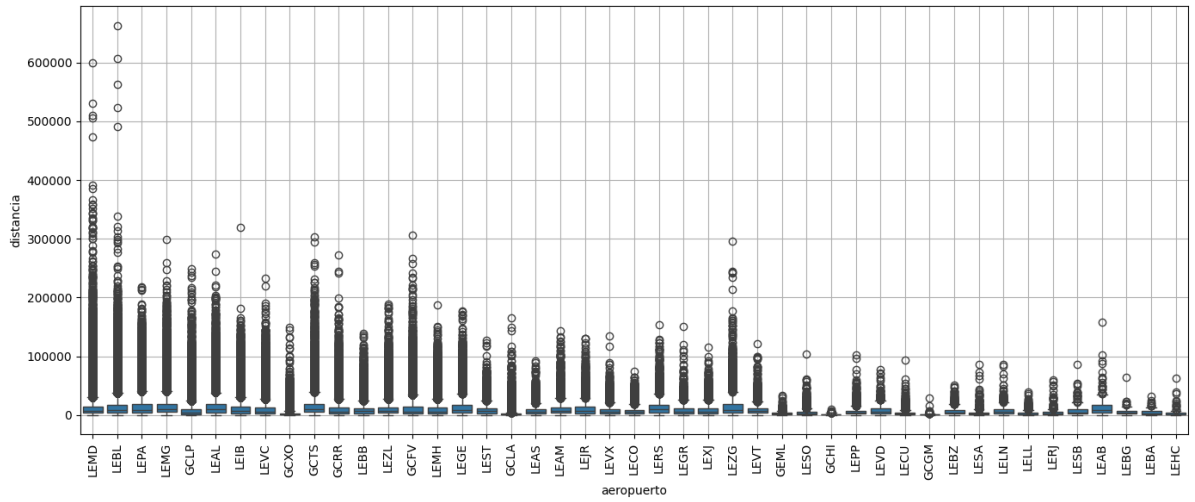

Figura 5.1: Diagrama de cajas de los valores de distancia por aeropuerto. Primera versión de la metodología.

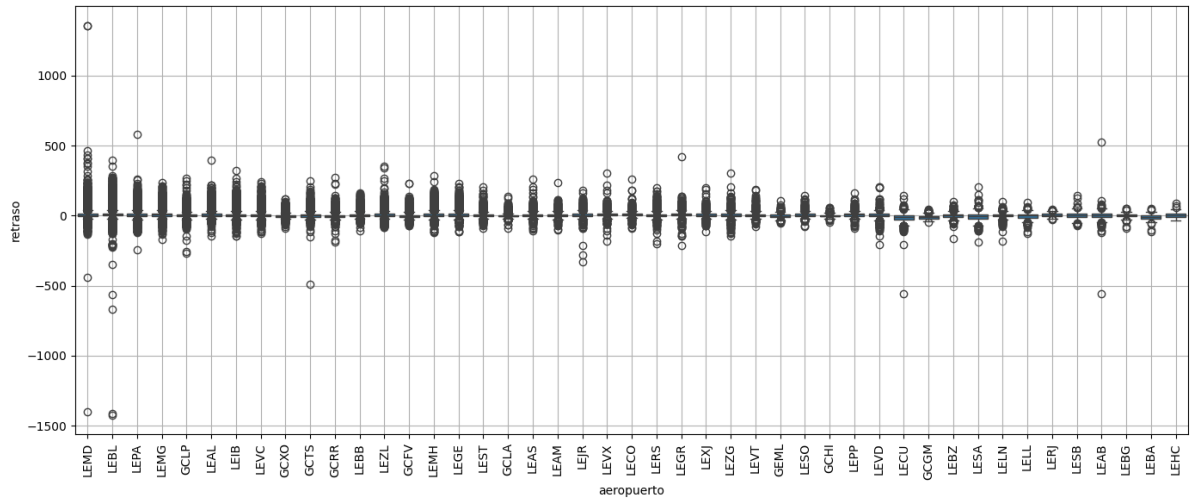

Figura 5.2: Diagrama de cajas de los valores de retraso por aeropuerto. Primera versión de la metodología.

Como se puede observar en el diagrama de caja de los valores de distancia (Figura 5.1), aeropuertos como Madrid y Barcelona (LEMD, LEBL) presentan vuelos con distancias muy elevadas (superiores a 600,000 km). Estos casos fueron revisados manualmente, y se detectó que suelen corresponder a vuelos con grandes retrasos. Aunque las trayectorias espaciales fueron similares, la diferencia significativa en el tiempo generó una distancia alta entre la trayectoria planificada y la real.

Por otro lado, en el diagrama de caja de los valores de retraso (Figura 5.2) se observan valores negativos muy extremos, lo que indica que algunos vuelos despegaron con mucha antelación. En ciertos casos, esta antelación llegó a alcanzar casi 1500 minutos, es decir, alrededor de 25 horas antes de la hora planificada. Dado que se trata de una situación altamente inusual, se decidió revisar manualmente los casos con adelantos iguales o superiores a 400 minutos (aproximadamente 6 horas).

El análisis reveló que se trataba de vuelos operativamente normales, cuya planificación indicaba una hora determinada, pero que finalmente despegaron con muchas horas de antelación, incluso en el día anterior.

Considerando que estos casos son puntuales y que podrían corresponder a vuelos especiales como vuelos presidenciales o de personajes importantes, cuya programación no es pública o se modifica por razones operativas, se optó por conservarlos en el conjunto de datos.

5.1.5. Diagramas de valores obtenidos en la segunda versión de la metodología

Dado que anteriormente se observó que los valores de distancia entre trayectorias estaban influenciados por el retraso de los vuelos, se han recalculado dichas distancias ajustando previamente las marcas temporales de la trayectoria real. Este ajuste permite alinear temporalmente la trayectoria real con la planificada, de acuerdo con la segunda versión de la metodología propuesta (véase la Sección 4.2).

Como resultado, en las Figuras 5.3 y 5.4 se presentan los diagramas de cajas correspondientes a los valores de distancia y retraso por aeropuerto, obtenidos tras aplicar esta segunda versión de la metodología.

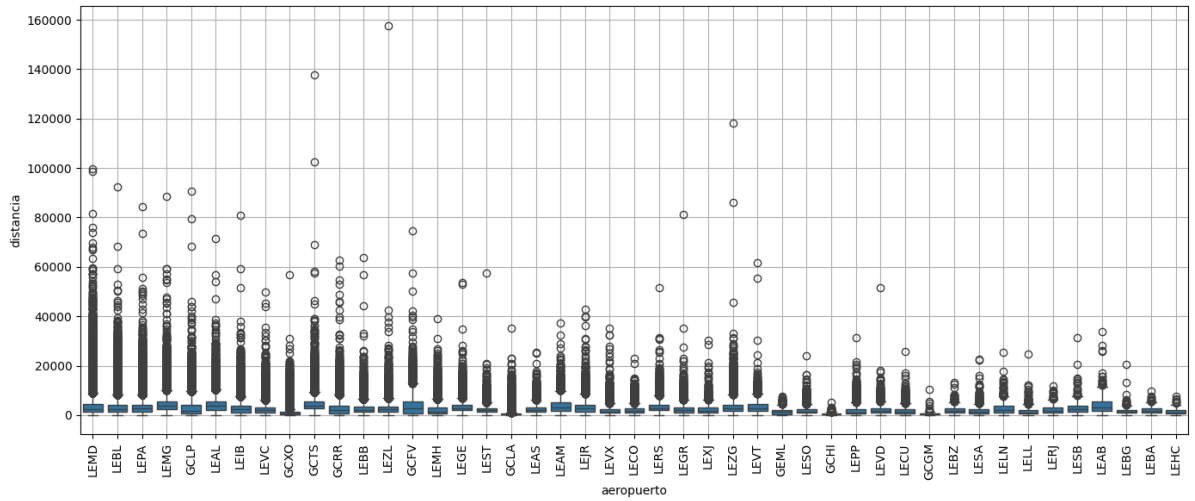

Figura 5.3: Diagrama de cajas de los valores de distancia por aeropuerto. Segunda versión de la metodología.

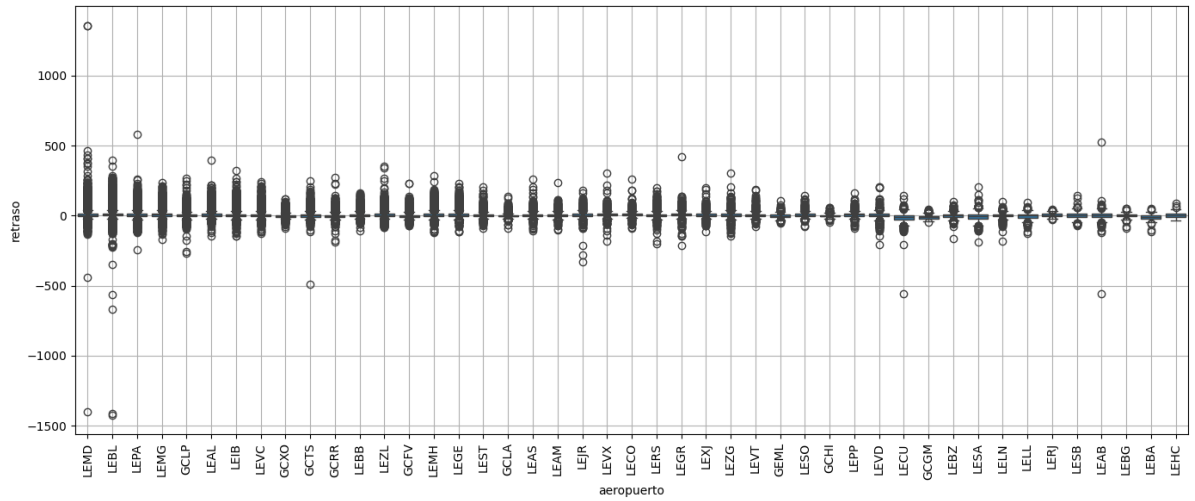

Figura 5.4: Diagrama de cajas de los valores de retraso por aeropuerto. Segunda versión de la metodología.

Como se observa en el diagrama de cajas de las distancias ajustadas (Figura 5.3), al corregir las marcas temporales de la trayectoria real, los valores de distancia entre trayectorias disminuyen considerablemente, con un máximo aproximado de 160,000 km, en contraste con los cerca de 600,000 km observados previamente (Figura 5.1).

No obstante, dado que la mayoría de los valores se sitúan por debajo de los 100,000 km, se procedió a revisar manualmente las trayectorias con distancias superiores para comprender las causas. En esta revisión, se identificó que el vuelo con mayor distancia (aproximadamente 160,000 km), correspondiente al aeropuerto de Sevilla (LEZL), presentaba un comportamiento anómalo: según los datos, un vuelo proveniente de Estocolmo, al aproximarse a Sevilla, en lugar de aterrizar, regresó a Francia antes de volver finalmente a Sevilla. Por este motivo, se consideró que dicho vuelo era erróneo y fue eliminado del conjunto de datos.

El resto de los vuelos con distancias superiores a 100,000 km se consideraron válidos, dado que pueden explicarse por desvíos importantes motivados por condiciones meteorológicas adversas, congestión en el aeropuerto de destino, entre otras causas.

Por otra parte, dado que no es habitual que los vuelos lleguen con una antelación significativa (más de una hora antes), se realizó un análisis detallado de los retrasos. Se encontró que algunos vuelos, aunque despegaron con pocos minutos de adelanto, puntuales o incluso con retraso, llegaron con más de una hora de antelación. Esto puede ser inusual, ya que, aunque se sabe que los vuelos suelen contar con márgenes en sus horarios para evitar retrasos, que un vuelo salga puntual o tarde y llegue más de una hora antes resulta atípico.

Se revisaron algunos casos particulares. Por ejemplo, un vuelo entre el aeropuerto de Madrid-Cuatro Vientos (LECU) y Albacete (LEAB), planificado para despegar a las 9:00 horas y llegar a las 19:00 horas (tiempo inusualmente largo para un trayecto tan corto), terminó despegando a las 9:00 horas y llegando a las 10:00 horas. Este caso se consideró como erróneo. Por ello, se decidió eliminar los vuelos que cumplieran con las siguientes condiciones: haber salido, como máximo, 20 minutos antes de la

hora planificada, o puntuales o con retraso, pero que hubieran llegado con más de una hora de antelación al destino.

A continuación, en las Tablas 5.4 y 5.5, se puede observar el número y el porcentaje de vuelos que quedan en el conjunto de datos tras eliminar estos vuelos erróneos y los que ya se habían mencionado anteriormente en la Sección 5.1.3.

Tabla 5.4: Número de vuelos en el conjunto de datos, número de vuelos eliminados y número de vuelos por categoría. Segunda limpieza.

Categoría	**Total**
Total Vuelos Iniciales	1,134,813
Total Vuelos Eliminados	36,463
Total Vuelos Resultantes	1,098,350
0 - Vuelos Cancelados	25
1 - Vuelos que no llegaron a su destino (desviados)	1,518
2 - Vuelos con instantes de tiempo duplicados	203
3 - Vuelos que tienen el mismo origen y destino	4,219
4 - Vuelos que tienen como aeropuerto destino o origen "ZZZZ"	202
5 - Vuelos eliminados porque están duplicados	29,881
6 - Vuelos que llegaron el día antes que el día de inicio	5
7 - Vuelos que llegaron el día después del día de fin	188
8 - Vuelos que fueron detectados como erróneos	222

Tabla 5.5: Porcentaje de vuelos resultantes en el conjunto de datos, porcentaje de vuelos eliminados y porcentaje de vuelos por categoría. Segunda limpieza.

Categoría	**Porcentaje**
0 - Vuelos Cancelados	0.0022 %
1 - Vuelos que no llegaron a su destino (desviados)	0.1337 %
2 - Vuelos con instantes de tiempo duplicados	0.0178 %
3 - Vuelos que tienen el mismo origen y destino	0.3717 %
4 - Vuelos que tienen como aeropuerto destino o origen "ZZZZ"	0.0178 %
5 - Vuelos eliminados porque están duplicados	2.6331 %
6 - Vuelos que llegaron el día antes que el día de inicio	0.0004 %
7 - Vuelos que llegaron el día después del día de fin	0.0165 %
8 - Vuelos que fueron detectados como erróneos	0.0195 %
Total Vuelos Eliminados	3.2131 %
Total Vuelos Resultantes	96.7868 %

Como se puede ver en las Tablas 5.4 y 5.5, el número total de vuelos detectados como erróneos ha sido de 222. De estos, 221 corresponden a vuelos que han llegado al menos una hora antes de su hora planificada y que han salido puntuales, y uno es el vuelo de Sevilla que se ha considerado incorrecto. De esta forma, el conjunto de datos ha quedado con un total de 1,098,350 vuelos, lo que representa un 96.78 % del total inicial.

5.1.6. Diagramas de valores obtenidos en la tercera versión de la metodología

Para mejorar aún más el cálculo de la distancia entre las trayectorias y evitar que factores como la velocidad o la longitud del tramo recorrido por la aeronave influyan en el cálculo, se han vuelto a alinear los tiempos de la trayectoria real con los de la planificada, esta vez con mayor precisión. Para ello, se ha seguido la metodología

planteada en la Sección 4.3. Una vez hecho esto, y tras eliminar los vuelos mencionados anteriormente, en las Figuras 5.5 y 5.6 se pueden ver los diagramas de cajas de las variables de distancia y retraso por aeropuerto según la tercera versión de la metodología.

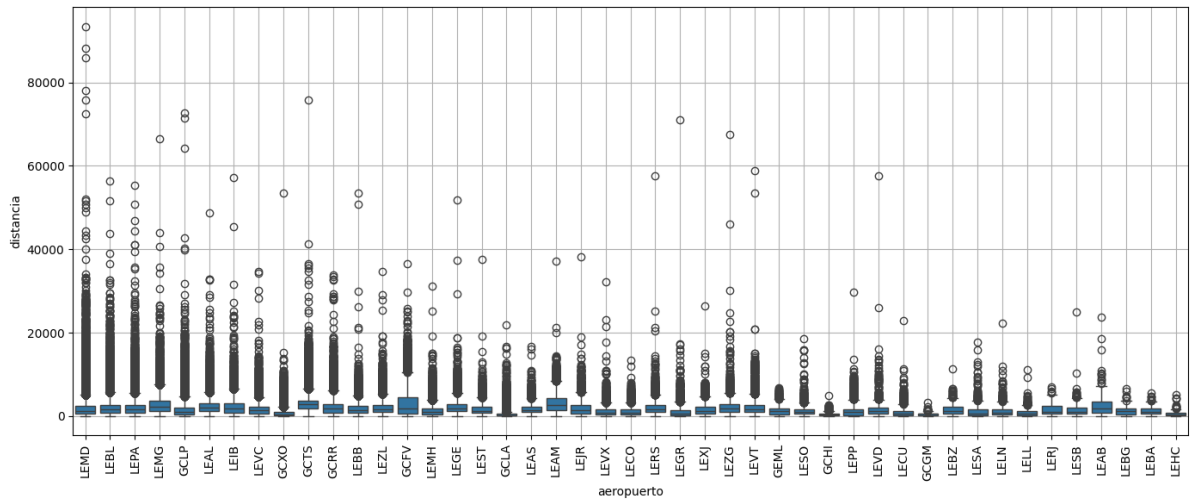

Figura 5.5: Diagrama de cajas de los valores de distancia por aeropuerto. Tercera versión de la metodología.

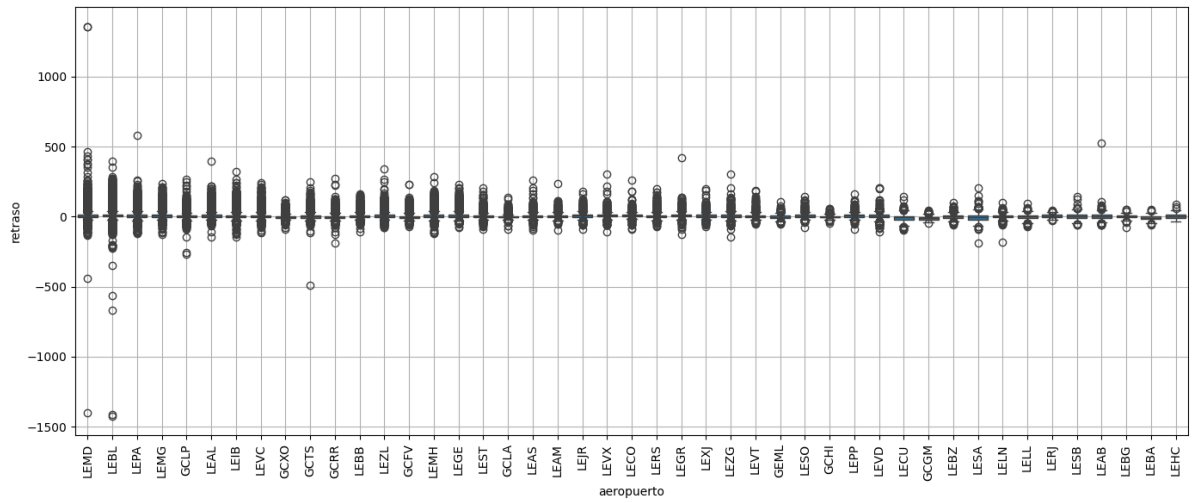

Figura 5.6: Diagrama de cajas de los valores de retraso por aeropuerto. Tercera versión de la metodología.

Como se puede ver en las Figuras 5.5 y 5.6, los valores de desviación ahora han sido mucho más bajos, sin superar los 90,000 km. Además, algunos de los valores de adelantos muy grandes, como el que correspondía al aeropuerto de Albacete (LEAB) en el diagrama anterior (Figura 5.4), ya no aparecen, ya que esos vuelos han sido eliminados. Esto indica que la limpieza de los datos ha funcionado correctamente,

eliminando valores extraños que podían afectar al análisis. Aparte que el ajuste de los valores de distancia permite trabajar con valores más representativos.

5.2. Aplicación de las técnicas de análisis topológico

En este apartado se presentan los resultados y valores obtenidos tras aplicar la metodología propuesta en este TFM. Todos estos resultados y valores se han obtenido utilizando la tercera versión de la metodología, ya que esta integra las mejoras incorporadas progresivamente durante el desarrollo del proyecto y constituye el enfoque final adoptado.

Es importante destacar que, como se explicó en la Sección 4.1.2, el análisis se ha aplicado sobre dos casos en la construcción de las nubes de puntos, que difieren en la forma en que se representan los vuelos en las mismas:

- En el primer caso, todas las distancias entre trayectorias planificadas y reales se mantienen como valores positivos, independientemente de si el vuelo llegó adelantado o con retraso.
- En el segundo caso, las distancias se transforman en valores negativos cuando el vuelo presenta un adelanto (retraso negativo), con el fin de explorar si esta codificación influye en los resultados del análisis.

A partir de estos dos casos, se aplica de forma independiente el análisis topológico sobre las nubes de puntos, obteniendo dos clasificaciones distintas de los aeropuertos.

A continuación, se presentan los resultados obtenidos de la siguiente manera. En primer lugar se muestra un ejemplo que permite visualizar cómo se representan las nubes de puntos, los diagramas de persistencia y los paisajes de persistencia. Para ello, se muestran las nubes, diagramas y paisajes obtenidos para ciertos aeropuertos en ciertos días específicos. El objetivo de estas visualizaciones es ofrecer una representación clara de cada una de las etapas del proceso metodológico.

En segundo lugar, se analizan las matrices de distancias entre aeropuertos generadas con la tercera versión de la metodología, extrayendo subconjuntos de matrices para compararlas con los resultados obtenidos en el estudio de Cuerno *et al.* (2023).

Posteriormente, se presentan los valores y las gráficas del Silhouette Score obtenidos al aplicar esta métrica a cada una de las matrices de distancias generadas en los distintos casos considerados por la metodología. Este mismo análisis se realiza también sobre una matriz de distancias que contiene los valores de distancia entre aeropuertos obtenidos por el estudio de Cuerno *et al.* (2023), con el objetivo de establecer una comparación entre ambos enfoques. Aunque dicho estudio no incluyó un proceso de clasificación, los autores proporcionaron las distancias calculadas entre aeropuertos como resultado de su metodología. A partir de estas distancias, en este TFM se construyó la matriz correspondiente y se le aplicó el cálculo del *Silhouette Score*. Además, también se le aplicó el algoritmo *Agglomerative Clustering*, obteniendo así también una clasificación de los aeropuertos españoles según las distancias obtenidas en el estudio. Esto permite comparar de forma directa los resultados obtenidos en ambos estudios.

Por último, se presentan las clasificaciones finales obtenidas en este TFM, correspondientes a los dos casos considerados de la metodología, así como la derivada de la matriz de distancias del estudio de Cuerno *et al.* (2023). También se incluye, a modo comparativo, la agrupación de aeropuertos españoles obtenida en otro Trabajo de Fin de Máster que se basa en las mismas variables y mismo conjunto de datos pero utilizando otra técnica de análisis topológico (Serrano, 2025).

5.2.1. Nubes de puntos, diagramas de persistencia y paisajes de persistencia

Una vez limpiados los datos y calculadas las métricas de distancia para las trayectorias, se construyen las nubes de puntos, donde cada punto representa un vuelo correspondiente a un día y un aeropuerto. Estos puntos incorporan tanto las métricas de distancia calculadas entre la trayectoria planificada y la real, como el retraso de llegada asociado al vuelo. Como se explicó en la Sección 3.1, estas nubes de puntos constituyen la estructura sobre la cual se aplica el análisis topológico mediante homología persistente.

A continuación, se muestran ejemplos de las nubes de puntos para los aeropuertos de Valladolid (LEVD), Santander (LEXJ), Ibiza (LEIB) y Madrid-Barajas (LEMD), correspondientes a los días 3 y 4 de julio de 2018. En cada caso, se presentan las dos formas de construcción de nubes: por un lado, el primer caso, donde todas las distancias son positivas, y por otro, el segundo caso, donde las distancias toman valores negativos cuando hay adelanto. Para facilitar la representación, únicamente se han tenido en cuenta las variables de distancia total y retraso en la visualización de las nubes de puntos.

De esta manera, las Figuras 5.7, 5.9, 5.11 y 5.13 muestran las nubes de puntos generadas para los aeropuertos de Valladolid (LEVD), Santander (LEXJ), Ibiza (LEIB) y Madrid-Barajas (LEMD), respectivamente, correspondientes a los días 3 y 4 de julio de 2018 y al primer caso de la metodología. Y las Figuras 5.8, 5.10, 5.12 y 5.14 presentan las nubes de puntos de esos mismos aeropuertos, en los mismos días, pero bajo el segundo caso de la metodología.

Como se puede observar, en todas las nubes de puntos correspondientes al primer caso de la metodología (Figuras 5.7, 5.9, 5.11 y 5.13), los puntos están entre el primer y el cuarto cuadrante, ya que los valores de distancia son siempre positivos y el retraso puede ser tanto positivo como negativo. Por otra parte, al analizar las nubes de puntos del segundo caso de la metodología (Figuras 5.8, 5.10, 5.12 y 5.14), se puede ver que los puntos se sitúan entre el primer y el tercer cuadrante, ya que en este caso los valores de distancia se transforman en negativos cuando el retraso también lo es.

Además, al comparar las nubes de puntos de los distintos aeropuertos, se puede notar claramente la diferencia en la cantidad de vuelos que manejan. Por ejemplo, al observar las nubes correspondientes al aeropuerto de Valladolid (Figuras 5.7 y 5.8), se aprecia que cada nube contiene muy pocos puntos, lo que indica que este aeropuerto tiene un tráfico aéreo bajo. Posteriormente, al analizar las nubes correspondientes al aeropuerto de Santander (Figuras 5.9 y 5.10), se nota un ligero aumento en la cantidad de puntos en comparación con Valladolid, aunque sigue siendo un número reducido.

Por otro lado, en las nubes de puntos del aeropuerto de Ibiza (Figuras 5.11 y 5.12), se observa un incremento considerable en el número de vuelos por día, lo que refleja una mayor actividad aérea. Finalmente, en el caso del aeropuerto de Madrid-Barajas (Figuras 5.13 y 5.14), la densidad de puntos es notablemente superior, confirmando que se trata de un aeropuerto con un volumen de tráfico muy alto, tanto en número de vuelos como en diversidad de retrasos.

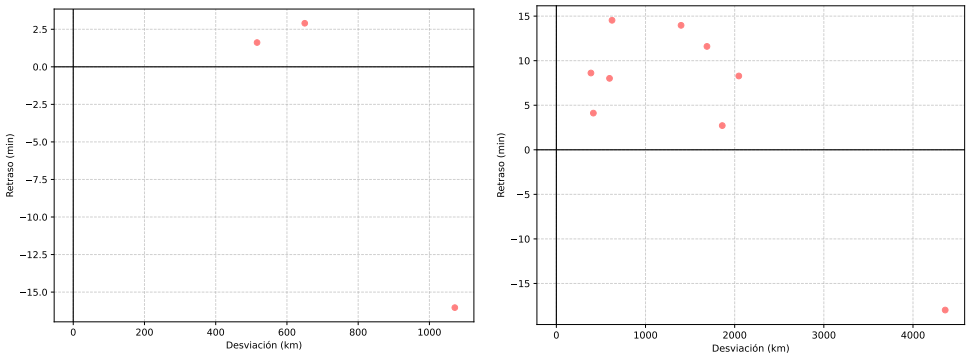

Figura 5.7: Nubes de puntos para el aeropuerto de Valladolid (LEVD) en los días 3 (izquierda) y 4 (derecha) de julio de 2018, generadas en el caso 1 de la metodología.

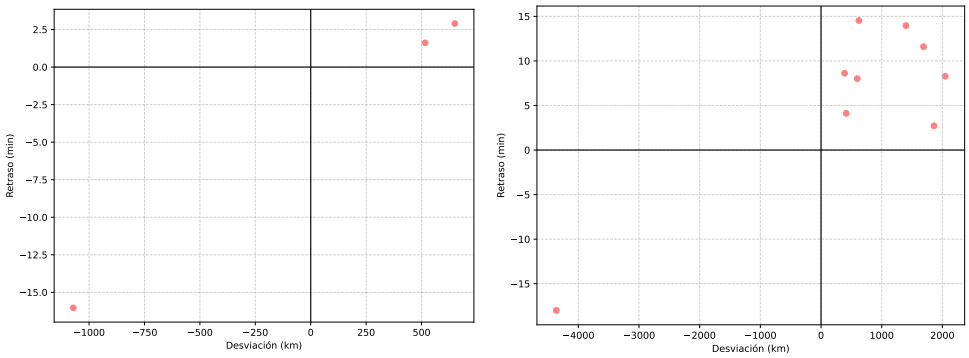

Figura 5.8: Nubes de puntos para el aeropuerto de Valladolid (LEVD) en los días 3 (izquierda) y 4 (derecha) de julio de 2018, generadas en el caso 2 de la metodología.

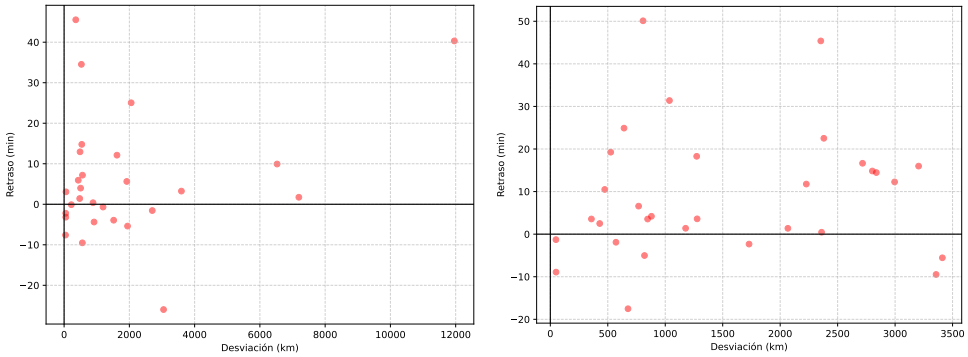

Figura 5.9: Nubes de puntos para el aeropuerto de Santander (LEXJ) en los días 3 (izquierda) y 4 (derecha) de julio de 2018, generadas en el caso 1 de la metodología.

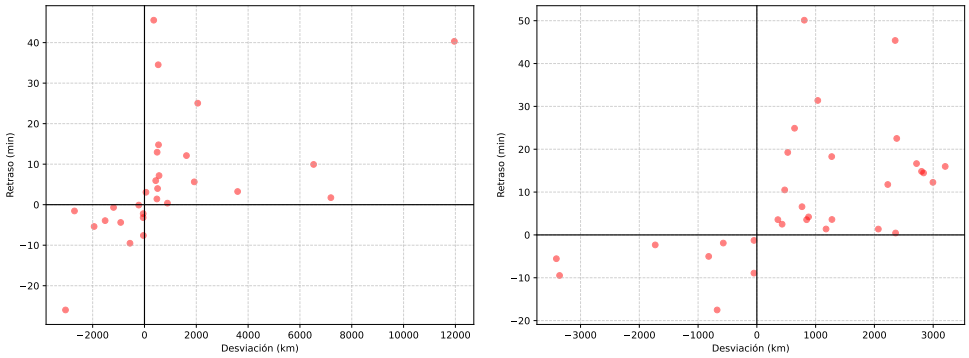

Figura 5.10: Nubes de puntos para el aeropuerto de Santander (LEXJ) en los días 3 (izquierda) y 4 (derecha) de julio de 2018, generadas en el caso 2 de la metodología.

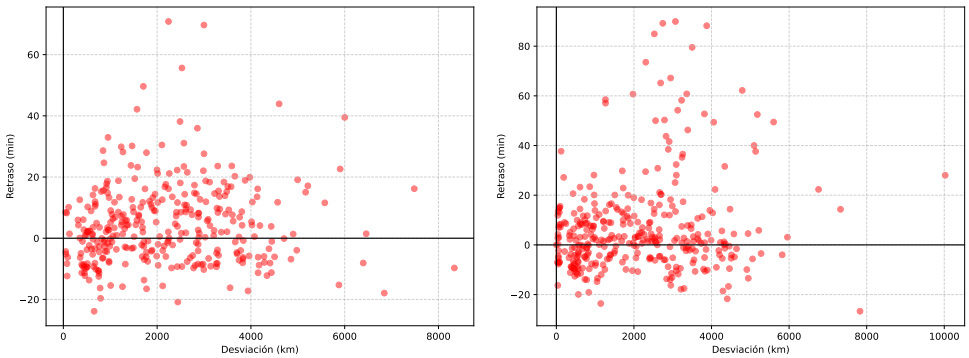

Figura 5.11: Nubes de puntos para el aeropuerto de Ibiza (LEIB) en los días 3 (izquierda) y 4 (derecha) de julio de 2018, generadas en el caso 1 de la metodología.

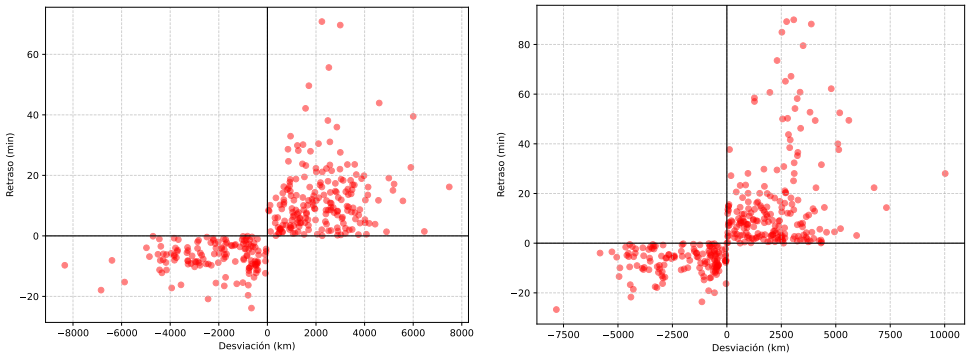

Figura 5.12: Nubes de puntos para el aeropuerto de Ibiza (LEIB) en los días 3 (izquierda) y 4 (derecha) de julio de 2018, generadas en el caso 2 de la metodología.

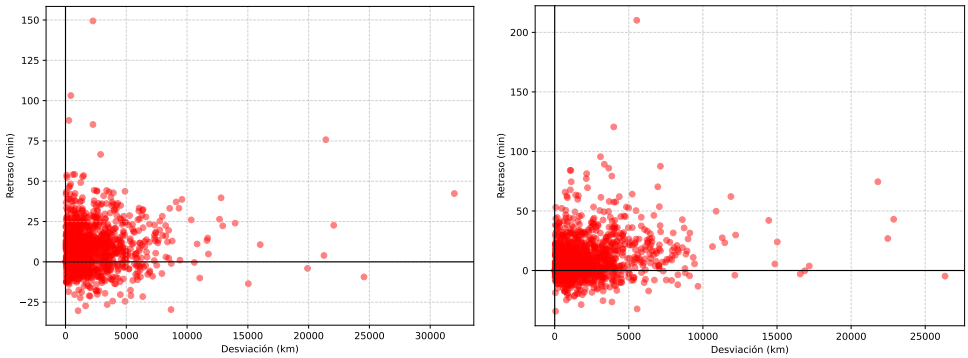

Figura 5.13: Nubes de puntos para el aeropuerto de Madrid-Barajas (LEMD) en los días 3 (izquierda) y 4 (derecha) de julio de 2018, generadas en el caso 1 de la metodología.

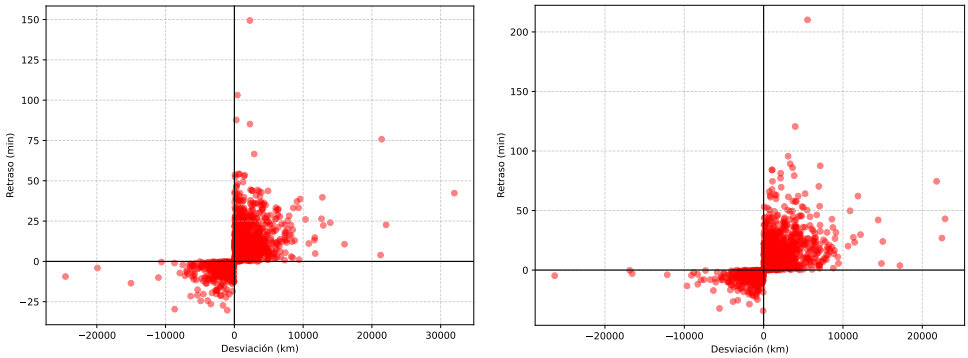

Figura 5.14: Nubes de puntos para el aeropuerto de Madrid-Barajas (LEMD) en los días 3 (izquierda) y 4 (derecha) de julio de 2018, generadas en el caso 2 de la metodología.

Posteriormente, en las Figuras 5.15 a 5.22 se muestran los diagramas de persistencia de los aeropuertos de Valladolid (LEVD), Santander (LEXJ), Ibiza (LEIB) y Madrid-Barajas (LEMD), correspondientes a los días 3 y 4 de julio de 2018 y generados a partir de las nubes de puntos mostradas previamente. Para cada aeropuerto se muestran los resultados obtenidos considerando tanto el caso 1 (distancias siempre positivas) como el caso 2 (distancias negativas cuando el retraso es negativo) de la metodología. Los diagramas han sido calculados utilizando la filtración de Vietoris–Rips en las dimensiones de homología H_0 y H_1, aunque en este TFM únicamente se analiza la información proporcionada por H_0, correspondiente a la conectividad de los datos.

En particular, las Figuras 5.15, 5.17, 5.19 y 5.21 muestran los diagramas de persistencia correspondientes al caso 1, para los aeropuertos de Valladolid, Santander, Ibiza y Madrid-Barajas, respectivamente. Por su parte, las Figuras 5.16, 5.18, 5.20 y 5.22 presentan los diagramas correspondientes al caso 2, para esos mismos aeropuertos y fechas.

Analizando los diagramas de persistencia en H_0 y H_1, se observa que los diagramas correspondientes a Ibiza (LEIB) y Madrid (LEMD) (Figuras 5.19, 5.20, 5.21 y 5.22) contienen muchos más puntos en H_0 que los de Valladolid (LEVD) y Santander (LEXJ) (Figuras 5.15, 5.16, 5.17 y 5.18), lo cual es coherente dado que Valladolid y

Santander presentan un número considerablemente menor de vuelos. En particular, solo los aeropuertos de Ibiza y Madrid presentan puntos en H_1, lo que implica la presencia de estructuras de tipo ciclo, es decir, bucles en los datos, reflejando una mayor complejidad topológica en la distribución. Esta complejidad está probablemente asociada tanto a la mayor cantidad de datos disponibles en estos aeropuertos como a una mayor variabilidad en su comportamiento operativo.

Por otra parte, si se observa el diagrama correspondiente al día 3 de julio para el aeropuerto de Valladolid (LEVD), tanto en el caso 1 (Figura 5.15) como en el caso 2 (Figura 5.16), el diagrama contiene únicamente tres puntos en H_0. Esto se debe a que la nube de puntos original para ese día incluye solamente tres vuelos (Figuras 5.7, 5.8), lo que limita considerablemente la complejidad topológica que puede capturarse mediante homología persistente.

En términos generales, al comparar los diagramas obtenidos para cada aeropuerto entre el caso 1 de la metodología, donde las distancias se representan en las nubes de puntos siempre positivas, y el caso 2 de la metodología, donde las distancias se representan en las nubes de puntos negativas si el retraso también lo es, se observa que los diagramas de persistencia no varían demasiado entre ambas configuraciones de cada aeropuerto. Esto indica que representar las distancias positivas o negativas no parece alterar sustancialmente la estructura de conectividad de las nubes de puntos, al menos en lo que respecta a la dimensión H_0.

Hay que tener en cuenta que, en el contexto del análisis topológico realizado mediante la filtración de Vietoris-Rips, el parámetro r representa un umbral de proximidad que determina qué puntos de la nube se conectan entre sí en cada etapa del proceso. Este umbral, así como las distancias calculadas entre puntos, se mide en las mismas unidades que el espacio métrico definido sobre la nube de puntos. Dado que en este TFM se emplea un espacio vectorial de cinco dimensiones en el que algunas variables están expresadas en kilómetros (métricas de distancias) y otra en minutos (como el retraso), la métrica combinada utilizada, usualmente una distancia euclídea, da lugar a una escala sin unidades físicas directas. En consecuencia, los valores del parámetro r, así como las coordenadas de nacimiento y muerte (b, d) de los diagramas de persistencia y los picos de los paisajes, deben interpretarse como medidas relativas dentro de un espacio con unidades heterogéneas. Salvo que se aplique una normalización previa de las variables, estas distancias no corresponden a una magnitud física concreta, sino a una escala interna determinada por la combinación de las variables originales y su distancia en el espacio euclidiano.

Por otra parte, en las Figuras 5.23 a 5.30 se presentan los paisajes de persistencia correspondientes a los aeropuertos de Valladolid (LEVD), Santander (LEXJ), Ibiza (LEIB) y Madrid-Barajas (LEMD), generados a partir de los diagramas de persistencia mostrados previamente (Figuras 5.15 a 5.22) y correspondientes a los días 3 y 4 de julio de 2018.

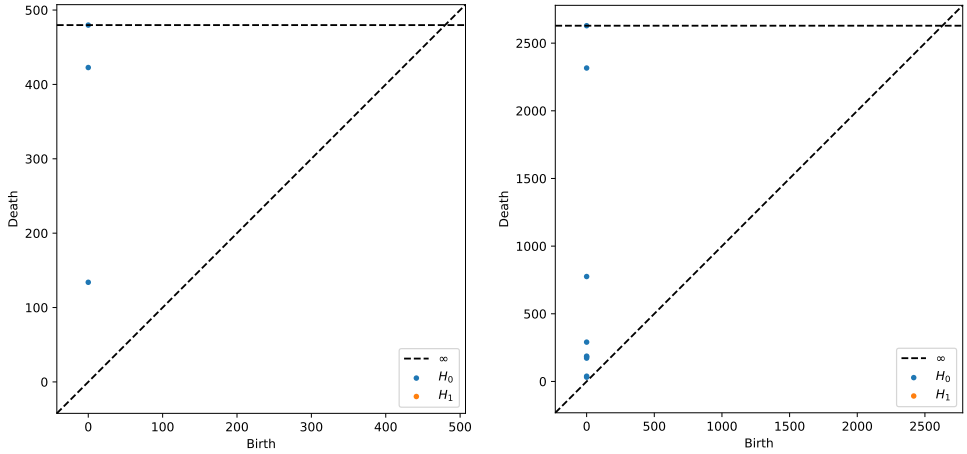

Figura 5.15: Diagramas de persistencia para el aeropuerto de Valladolid (LEVD) en los días 3 (izquierda) y 4 (derecha) de julio de 2018 y caso 1 de la metodología.

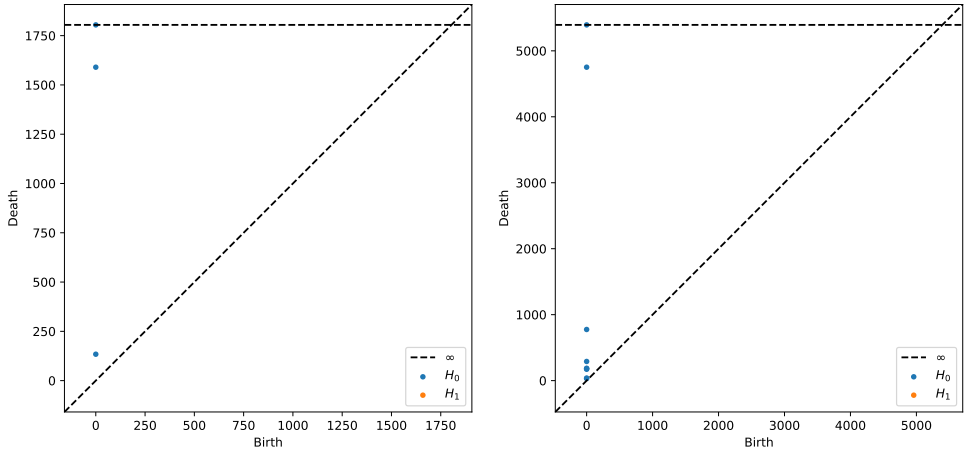

Figura 5.16: Diagramas de persistencia para el aeropuerto de Valladolid (LEVD) en los días 3 (izquierda) y 4 (derecha) de julio de 2018 y caso 2 de la metodología.

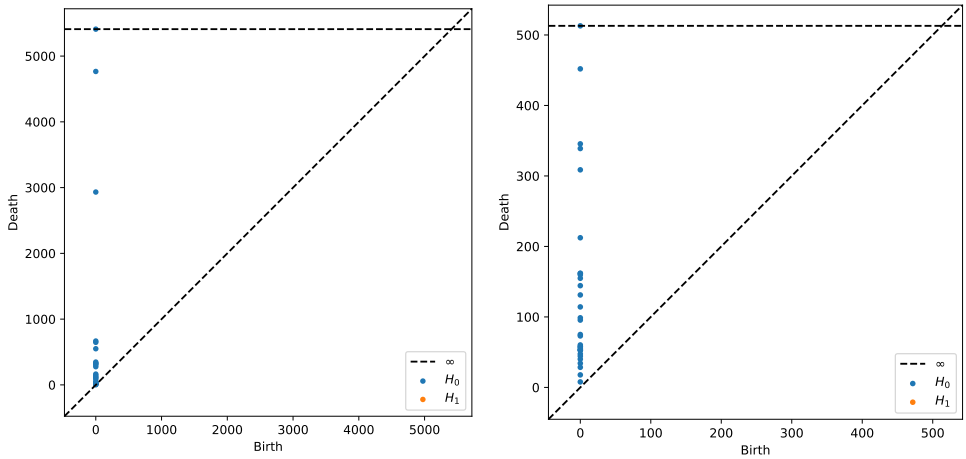

Figura 5.17: Diagramas de persistencia para el aeropuerto de Santander (LEXJ) en los días 3 (izquierda) y 4 (derecha) de julio de 2018 y caso 1 de la metodología.

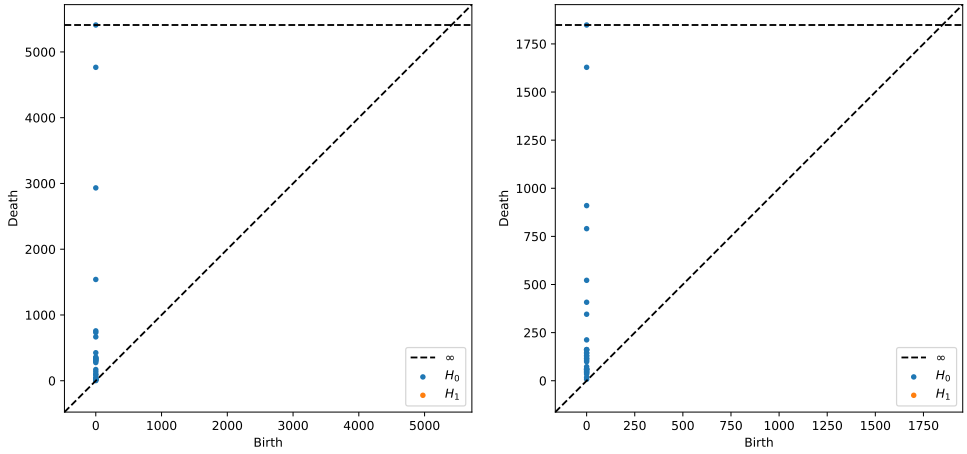

Figura 5.18: Diagramas de persistencia para el aeropuerto de Santander (LEXJ) en los días 3 (izquierda) y 4 (derecha) de julio de 2018 y caso 2 de la metodología.

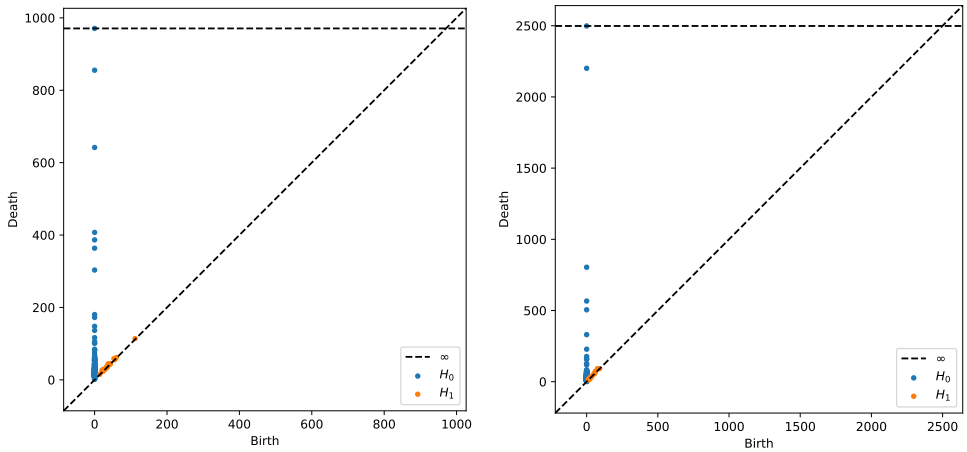

Figura 5.19: Diagramas de persistencia para el aeropuerto de Ibiza (LEIB) en los días 3 (izquierda) y 4 (derecha) de julio de 2018y caso 1 de la metodología.

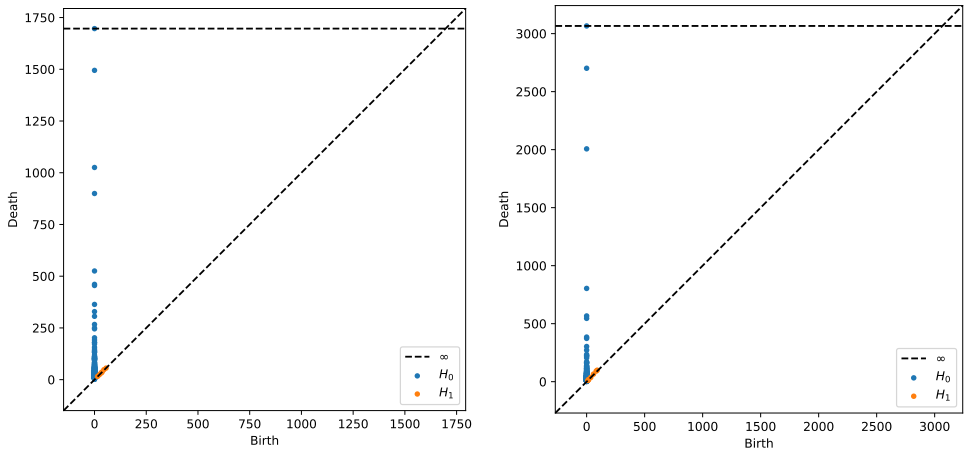

Figura 5.20: Diagramas de persistencia para el aeropuerto de Ibiza (LEIB) en los días 3 (izquierda) y 4 (derecha) de julio de 2018 y caso 2 de la metodología.

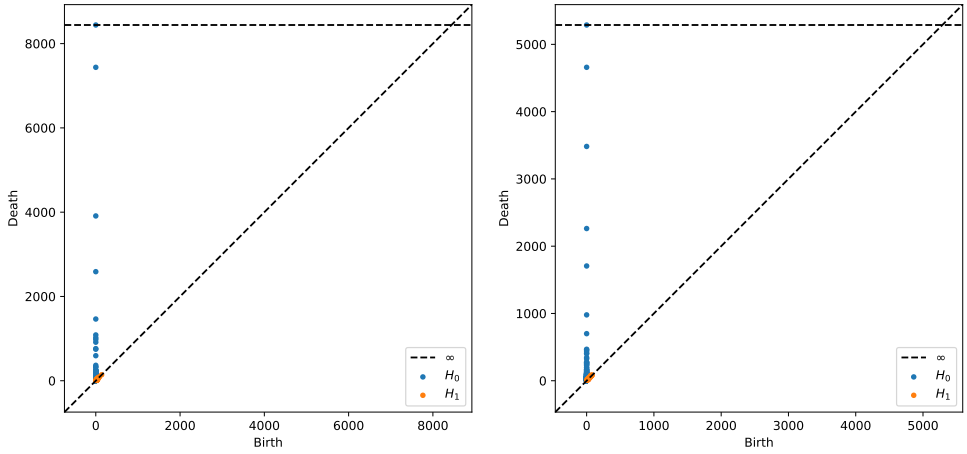

Figura 5.21: Diagramas de persistencia para el aeropuerto de Madrid-Barajas (LEMD) en los días 3 (izquierda) y 4 (derecha) de julio de 2018 y caso 1 de la metodología.

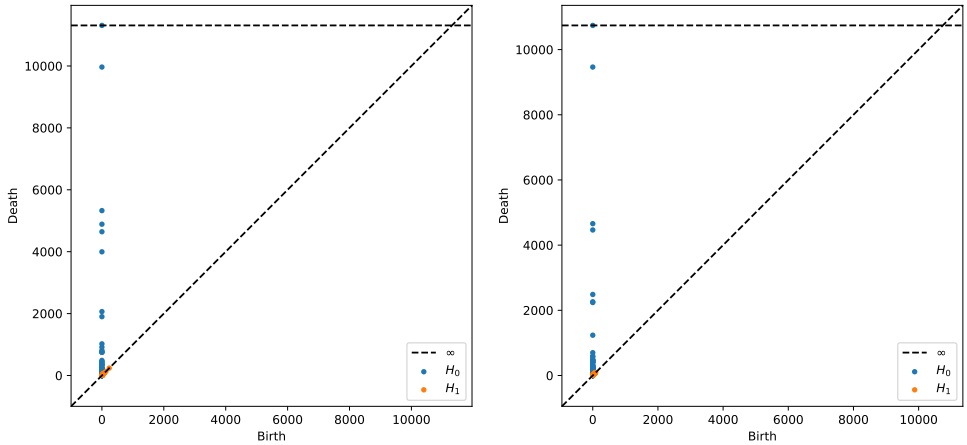

Figura 5.22: Diagramas de persistencia para el aeropuerto de Madrid-Barajas (LEMD) en los días 3 (izquierda) y 4 (derecha) de julio de 2018 y caso 2 de la metodología.

Al comparar los paisajes por aeropuerto entre ambos casos de la metodología, se observa que la estructura general de las funciones (posición y número de picos) se mantiene estable. Sin embargo, se evidencian diferencias en la altura de los picos: en el segundo caso de la metodología, donde las distancias toman valores negativos cuando hay adelanto, algunos picos son visiblemente más altos. Por ejemplo, en el aeropuerto de Valladolid (LEVD), para el día 3 de julio, el pico más alto del paisaje alcanza un valor aproximado de 800 en el segundo caso, frente a 200 en el primero. Para el día 4 de julio, esta diferencia es aún mayor: 2500 en el segundo caso frente a 1200 en el primero.

Estas diferencias se deben a la manera en que se codifican las trayectorias: al introducir distancias negativas en presencia de adelantos, se incrementa artificialmente la separación entre algunos puntos de la nube. Esto provoca una mayor duración (vida) de ciertas características topológicas en la filtración de Vietoris–Rips, lo cual se traduce en intervalos más largos en los diagramas de persistencia y, en consecuencia, en picos más altos en los paisajes.

Un ejemplo adicional se observa en el aeropuerto de Santander (LEXJ), el día 4 de julio. En el caso 2 de la metodología (Figura 5.26), la función de mayor persistencia (representada en color azul) alcanza un valor significativamente superior al observado en el caso 1 (Figura 5.25): aproximadamente 800 frente a 300. Una diferencia similar se aprecia el día 3 de julio, donde en el segundo caso aparece una función de color verde que no se manifiesta en el primero.

En general, al analizar con detalle los paisajes de persistencia para cada aeropuerto y ambos casos, se encuentra que, aunque la forma general entre los paisajes de cada caso de la metodología es bastante parecida, se identifican pequeñas diferencias entre ellos, que muestra que existen diferencias entre ambas representaciones pero no muy significativas.

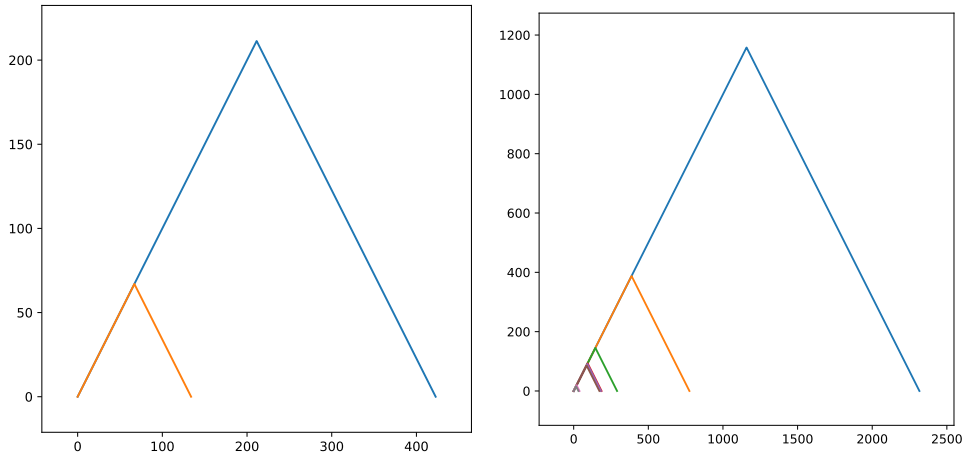

Figura 5.23: Paisajes de persistencia para el aeropuerto de Valladolid (LEVD) en los días 3 (izquierda) y 4 (derecha) de julio de 2018 y caso 1 de la metodología.

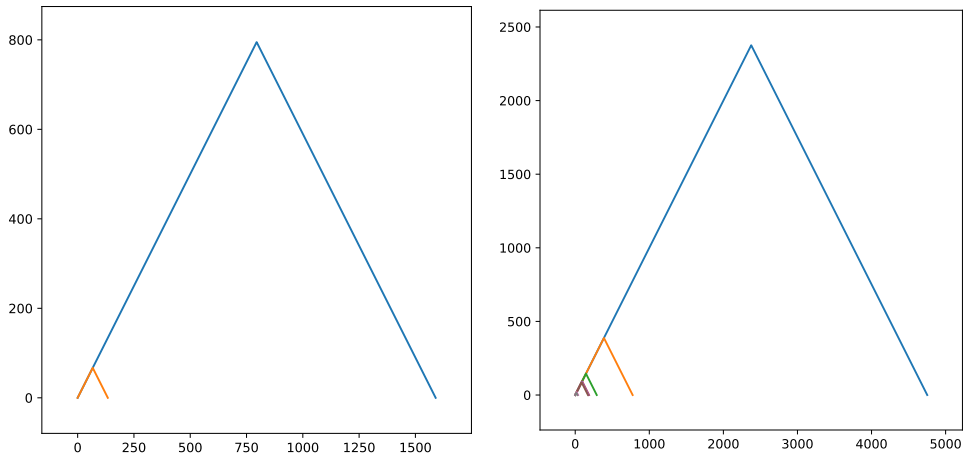

Figura 5.24: Paisajes de persistencia para el aeropuerto de Valladolid (LEVD) en los días 3 (izquierda) y 4 (derecha) de julio de 2018 y caso 2 de la metodología.

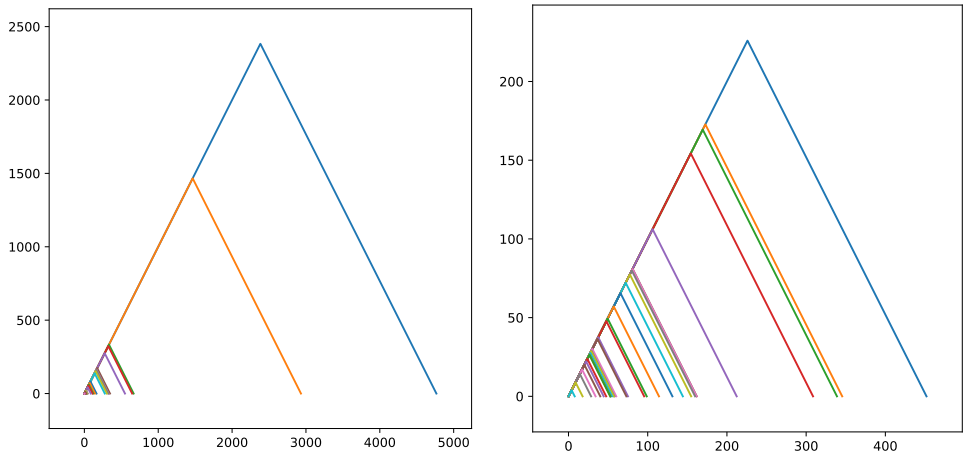

Figura 5.25: Paisajes de persistencia para el aeropuerto de Santander (LEXJ) en los días 3 (izquierda) y 4 (derecha) de julio de 2018 y caso 1 de la metodología.

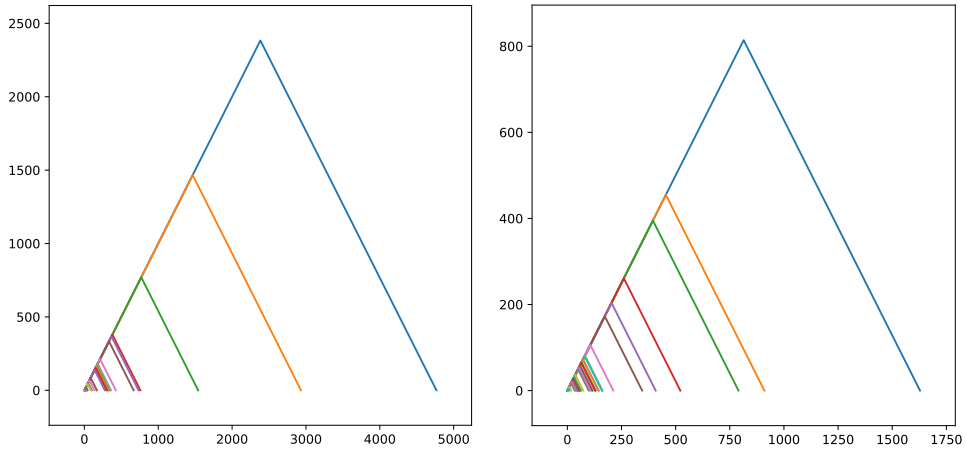

Figura 5.26: Paisajes de persistencia para el aeropuerto de Santander (LEXJ) en los días 3 (izquierda) y 4 (derecha) de julio de 2018 y caso 2 de la metodología.

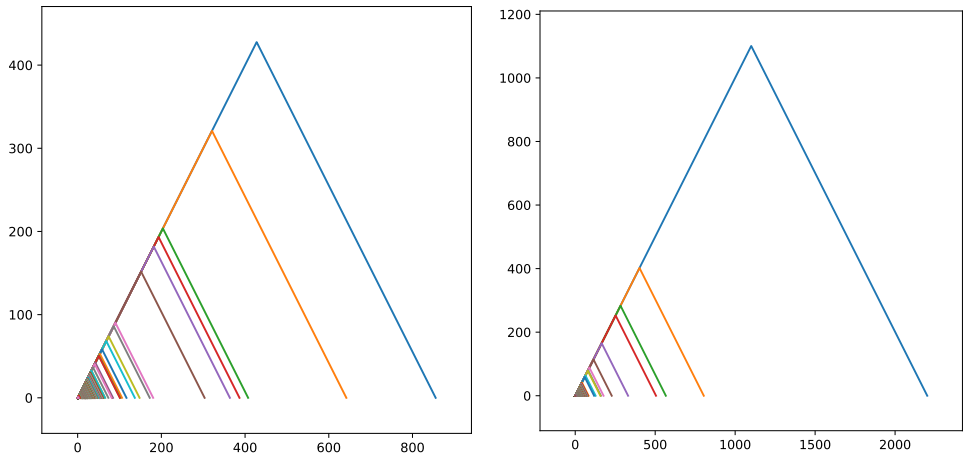

Figura 5.27: Paisajes de persistencia para el aeropuerto de Ibiza (LEIB) en los días 3 (izquierda) y 4 (derecha) de julio de 2018 y caso 1 de la metodología.

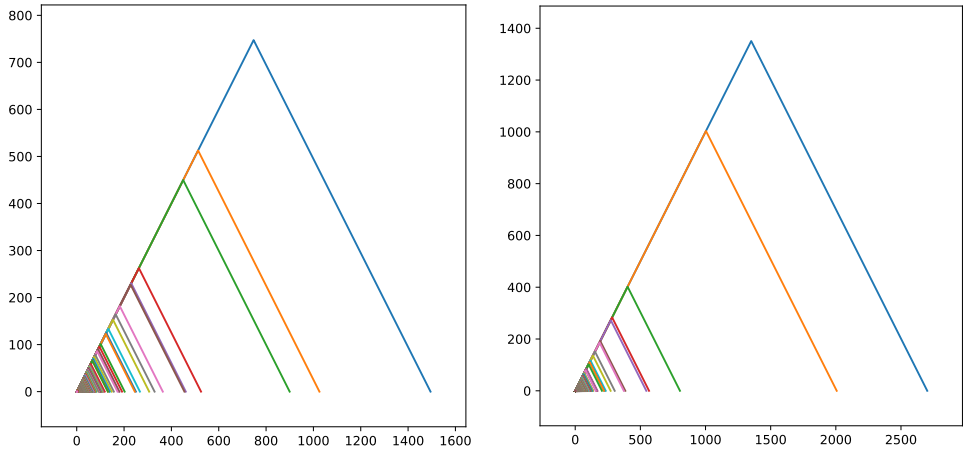

Figura 5.28: Paisajes de persistencia para el aeropuerto de Ibiza (LEIB) en los días 3 (izquierda) y 4 (derecha) de julio de 2018 y caso 2 de la metodología.

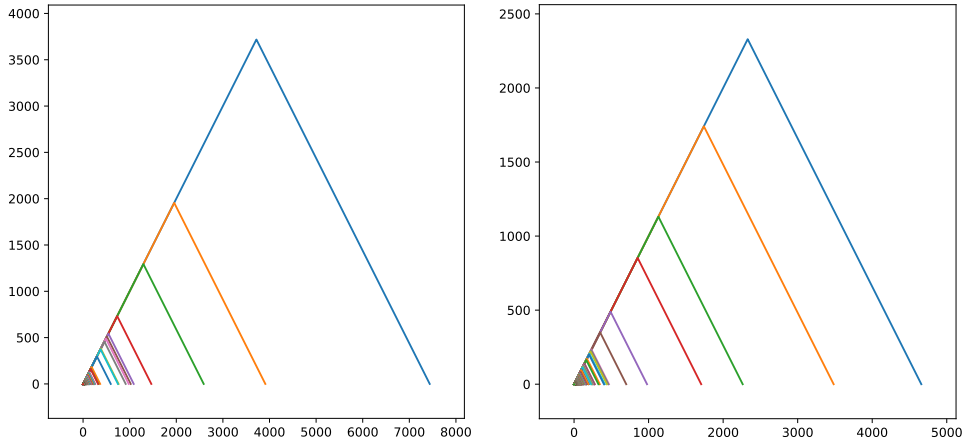

Figura 5.29: Paisajes de persistencia para el aeropuerto de Madrid-Barajas (LEMD) en los días (izquierda) y 4 (derecha) de julio de 2018 y caso 1 de la metodología.

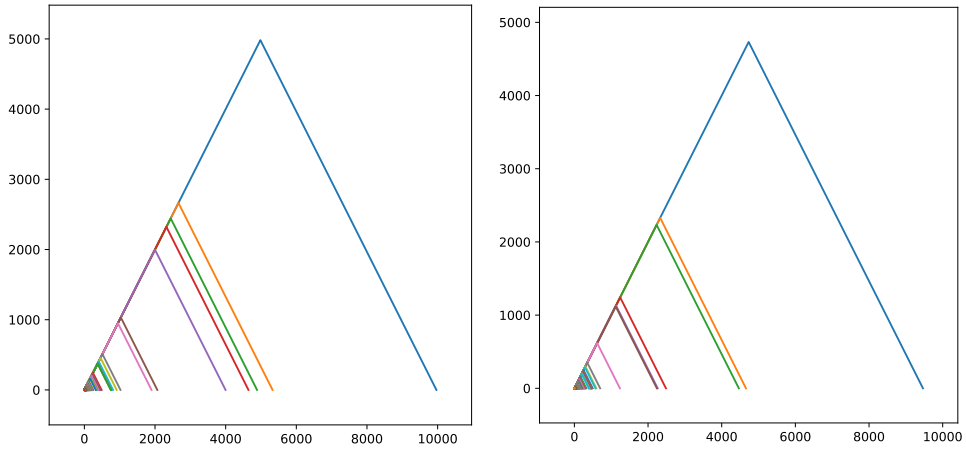

Figura 5.30: Paisajes de persistencia para el aeropuerto de Madrid-Barajas (LEMD) en los días 3 (izquierda) y 4 (derecha) de julio de 2018 y caso 2 de la metodología.

5.2.2. Distancias entre aeropuertos

Al aplicar el cálculo de homología persistente a cada nube de puntos y obtener sus correspondientes diagramas y paisajes de persistencia, se calcula la matriz de distancias promedio entre aeropuertos, tal como se explica en la Sección 4.1.3, para cada uno de los dos casos considerados en la metodología.

Para evaluar las distancias obtenidas, se extraen subtablas de dichas matrices que permiten observar las diferencias entre determinados aeropuertos y compararlas con los resultados del estudio realizado por Cuerno *et al.* (2023).

En dicho estudio, los autores comparan las distancias entre aeropuertos que pertenecen al mismo grupo operativo, según la clasificación de AENA (el gestor aeroportuario y de navegación aérea en España) correspondiente al año 2018. A continuación, se presenta dicha clasificación de aeropuertos según AENA en ese año:

- Group 3:
  - Aeropuertos de aviación general: Madrid-Cuatro Vientos (LECU) y Sabadell (LELL).
  - Bases aéreas abiertas a tráfico civil: Albacete (LEAB), Badajoz (LEBZ), León (LELN), Salamanca (LESA), Son Bonet (LESB) y Valladolid (LEVD).
  - Aeropuertos con bajo tráfico: Melilla (GEML), Burgos-Villafrá (LEBG), Córdoba (LEBA), Girona (LEGE), Huesca-Pirineos (LEHC), Logroño-Agoncillo (LERJ), Pamplona (LEPP), San Sebastián (LESO) y Vitoria (LEVT).
- Group 2: A Coruña (LECO), Almería (LEAM), Asturias (LEAS), Federico García Lorca Granada-Jaén (LEGR), Jerez (LEJR), Reus (LERS), Santander (LEXJ), Santiago (LEST), Vigo (LEVX) y Zaragoza (LEZG).
- Group 1: Alicante-Elche Miguel Hernández (LEAL), Bilbao (LEBB), Ibiza (LEIB), Málaga-Costa del Sol (LEMG), Menorca (LEMH), Sevilla (LEZL) y Valencia (LEVC).
- Canary Group: Fuerteventura (GCFV), La Gomera (GCGM), El Hierro (GCHI), La Palma (GCLA), Gran Canaria (GCLP), César Manrique Lanzarote (GCRR), Tenerife Sur (GCTS) y Tenerife Norte (GCXO).
- Special Group: Adolfo Suárez Madrid-Barajas (LEMD), Josep Tarradellas Barcelona-El Prat (LEBL) y Palma Mallorca (LEPA).

De esta manera, es posible evaluar las distancias entre los aeropuertos pertenecientes a cada grupo. En la Tabla 5.6 se muestran las distancias entre aeropuertos del grupo 2 de AENA, obtenidas a partir del primer caso de la metodología. Como se puede observar en la tabla, los aeropuertos de este grupo presentan distancias similares entre sí, con la excepción del aeropuerto de Zaragoza (LEZG), cuyas distancias con respecto al resto del grupo resultan ligeramente mayores. Esta misma diferencia ya se identifica en los resultados presentados en el estudio de Cuerno *et al.* (2023).

Tabla 5.6: Distancias entre aeropuertos pertenecientes al grupo 2 de AENA, calculada a partir del caso 1 de la metodología.

	LEAM	LEAS	LECO	LEGR	LEJR	LERS	LEST	LEVX	LEXJ	LEZG
LEAM	0,00	48.093,92	45.479,75	81.462,31	64.282,05	76.164,27	54.520,90	60.465,73	52.380,84	96.222,13
LEAS	48.093,92	0,00	25.509,00	65.550,06	48.354,91	60.911,74	38.193,86	40.110,28	33.764,51	85.528,62
LECO	45.479,75	25.509,00	0,00	61.417,73	46.504,99	56.418,19	34.770,19	39.197,82	32.923,62	81.760,27
LEGR	81.462,31	65.550,06	61.417,73	0,00	79.260,02	93.594,60	72.929,95	76.746,51	71.987,21	116.397,32
LEJR	64.282,05	48.354,91	46.504,99	79.260,02	0,00	74.246,46	55.230,84	63.094,39	53.357,01	98.871,00
LERS	76.164,27	60.911,74	56.418,19	93.594,60	74.246,46	0,00	66.255,12	70.133,64	64.421,49	112.932,90
LEST	54.520,90	38.193,86	34.770,19	72.929,95	55.230,84	66.255,12	0,00	50.494,16	43.650,94	91.047,17
LEVX	60.465,73	40.110,28	39.197,82	76.746,51	63.094,39	70.133,64	50.494,16	0,00	47.529,67	96.517,97
LEXJ	52.380,84	33.764,51	32.923,62	71.987,21	53.357,01	64.421,49	43.650,94	47.529,67	0,00	88.889,04
LEZG	96.222,13	85.528,62	81.760,27	116.397,32	98.871,00	112.932,90	91.047,17	96.517,97	88.889,04	0,00

De manera similar, en la Tabla 5.7 se muestran las distancias obtenidas entre los aeropuertos pertenecientes al grupo 2 de AENA utilizando el segundo caso de la metodología. También en este caso se observa que el aeropuerto de Zaragoza (LEZG) presenta una distancia ligeramente mayor en comparación con el resto de los aeropuertos de su grupo.

Tabla 5.7: Distancias entre aeropuertos pertenecientes al grupo 2 de AENA, calculada a partir del caso 2 de la metodología.

	LEAM	LEAS	LECO	LEGR	LEJR	LERS	LEST	LEVX	LEXJ	LEZG
LEAM	0,00	73.227,43	71.563,61	105.436,33	86.979,74	100.445,75	78.114,90	86.835,41	75.088,16	119.803,82
LEAS	73.227,43	0,00	35.832,58	76.485,32	61.152,90	74.985,87	46.145,43	53.050,70	43.781,73	99.881,86
LECO	71.563,61	35.832,58	0,00	72.665,77	58.963,37	71.262,39	43.149,50	52.453,26	42.587,90	96.678,65
LEGR	105.436,33	76.485,32	72.665,77	0,00	96.635,08	109.077,50	81.442,88	90.845,61	82.379,71	133.006,75
LEJR	86.979,74	61.152,90	58.963,37	96.635,08	0,00	90.257,12	66.592,00	78.284,79	64.578,07	113.588,43
LERS	100.445,75	74.985,87	71.262,39	109.077,50	90.257,12	0,00	80.433,54	85.813,52	77.451,42	128.230,30
LEST	78.114,90	46.145,43	43.149,50	81.442,88	66.592,00	80.433,54	0,00	60.448,83	52.931,92	104.254,29
LEVX	86.835,41	53.050,70	52.453,26	90.845,61	78.284,79	85.813,52	60.448,83	0,00	61.251,75	113.307,74
LEXJ	75.088,16	43.781,73	42.587,90	82.379,71	64.578,07	77.451,42	52.931,92	61.251,75	0,00	99.963,04
LEZG	119.803,82	99.881,86	96.678,65	133.006,75	113.588,43	128.230,30	104.254,29	113.307,74	99.963,04	0,00

A su vez, en las Tablas 5.8 y 5.9 se presentan las distancias obtenidas entre los aeropuertos del grupo 1 y los aeropuertos del grupo especial de AENA, calculadas a partir del primer y segundo caso de la metodología.

Tanto en la Tabla 5.8 como en la Tabla 5.9, se observa que los aeropuertos de Barcelona (LEBL) y Palma de Mallorca (LEPA) presentan distancias menores respecto al resto de aeropuertos del grupo 1, en comparación con el aeropuerto de Madrid-Barajas (LEMD), cuya distancia es considerablemente mayor. En el estudio de Cuerno *et al.* (2023), el aeropuerto de Palma de Mallorca también se identificaba como el más cercano al grupo 1, mientras que Barcelona y Madrid-Barajas mostraban distancias significativamente más altas.

No obstante, en el presente análisis se aprecia una diferencia importante: Barcelona, al igual que Palma de Mallorca, ahora muestra una distancia reducida respecto al resto del grupo 1. Esta variación se explica por los ajustes introducidos en el cálculo de las distancias en la tercera versión de la metodología. En la primera versión, las distancias entre las trayectorias planificadas y reales se calculan de la misma manera que en el estudio, lo que resultaba en una mayor distancia para Barcelona en comparación con Palma de Mallorca.

Tabla 5.8: Distancias entre aeropuertos del grupo 1 y los aeropuertos del grupo especial de AENA, calculada según el caso 1 de la metodología.

	LEBL	LEMD	LEPA
LEAL	176.911,02	326.881,21	156.257,10
LEBB	187.231,40	334.460,91	149.534,17
LEIB	184.699,31	338.265,44	151.154,77
LEMG	205.669,34	346.159,70	185.485,22
LEMH	153.717,12	310.865,19	126.134,67
LEST	159.935,55	313.631,86	127.034,96
LEVC	171.117,52	313.944,23	146.166,67
LEZL	170.488,32	319.023,80	136.792,73

Tabla 5.9: Distancias entre aeropuertos del grupo 1 y los aeropuertos del grupo especial de AENA, calculada según el caso 2 de la metodología.

	LEBL	LEMD	LEPA
LEAL	212.763,66	382.312,74	196.294,30
LEBB	218.169,18	399.758,39	186.776,89
LEIB	213.233,38	407.319,18	179.153,42
LEMG	247.831,76	416.051,24	233.899,56
LEMH	180.384,67	379.524,32	160.829,31
LEST	189.867,25	381.604,74	162.538,50
LEVC	200.528,89	380.248,66	185.039,84
LEZL	201.652,19	385.722,24	173.940,90

5.2.3. Clasificación de aeropuertos

5.2.3.1. Cálculo del Silhouette Score

Como se explicó previamente en la Sección 4.1.4, para llevar a cabo la agrupación de aeropuertos en clústeres (o grupos) se emplea el algoritmo *Agglomerative Clustering*. Por otro lado, para determinar el número óptimo de clústeres (k) que mejor representan la estructura de los datos, se utiliza el índice de *Silhouette Score*.

En este análisis, el algoritmo se ejecuta repetidamente, probando diferentes cantidades de clústeres, desde 2 hasta 9. Para cada configuración, se calcula el correspondiente *Silhouette Score*. Este procedimiento se aplica a las dos matrices de distancia generadas mediante la tercera versión de la metodología. Una de estas matrices se construye a partir de la nube de puntos que considera únicamente distancias positivas, mientras que la otra incluye también distancias negativas cuando el retraso es negativo. Además, se cuenta con una matriz adicional proporcionada por los autores del estudio de Cuerno *et al.* (2023), obtenida a partir de sus propias estimaciones de distancias entre aeropuertos, sobre la cual también se calcula el *Silhouette Score* siguiendo el mismo enfoque.

A continuación, se presentan los valores de *Silhouette Score* obtenidos para cada una de las tres matrices de distancias. En concreto, las Tablas 5.10 y 5.11, junto con las Figuras 5.31 y 5.32, muestran los valores y gráficas correspondientes al aplicar el *Silhouette Score* a cada una de las matrices obtenidas mediante cada caso de la metodología. Por su parte, la Tabla 5.12 y la Figura 5.33 recogen los valores y la gráfica obtenida a partir de la matriz de distancias del estudio de Cuerno *et al.* (2023).

Los resultados revelan ciertas diferencias en el comportamiento del *Silhouette Score* entre las distintas matrices. Aunque en todos los casos el score más alto se alcanza con 2 clústeres (0.7217 para la matriz con distancias positivas y 0.7052 para la matriz con distancias negativas), los valores obtenidos para 4 y 5 clústeres también

son considerablemente aceptables. En la matriz de distancias positivas los valores para 4 y 5 clústeres son 0.3881 y 0.3460, respectivamente (Tabla 5.10 y Figura 5.31), mientras que en la matriz negativa son 0.4716 y 0.4359 (Tabla 5.11 y Figura 5.32). Esto sugiere que existe una estructura de agrupamiento razonable incluso con un mayor número de clústeres.

En el caso de la matriz de distancias correspondiente al estudio realizado por Cuerno *et al.* (2023) (Tabla 5.12 y Figura 5.33), los valores más altos del Silhouette Score también se obtienen con 2 (0.7877) y 3 clústeres (0.7313), aunque nuevamente se observan valores aceptables para 4 (0.4075) y 5 clústeres (0.3608).

Considerando estos resultados, y con el objetivo de obtener una clasificación más detallada y útil de los aeropuertos, se ha optado por aplicar el algoritmo de *Agglomerative Clustering* con 5 clústeres. Aunque esta opción no maximiza el *Silhouette Score*, ofrece valores satisfactorios (0.34 y 0.43, según la matriz considerada) y permite una segmentación más rica y diferenciada que la agrupación en 2 o 3 clústeres.

Tabla 5.10: Silhouette Score para matriz de distancia calculada a partir del primer caso de la metodología.

Clusters	Silhouette Score
2	0.7217
3	0.4114
4	0.3881
5	0.3460
6	0.3547
7	0.3035
8	0.2975
9	0.2205

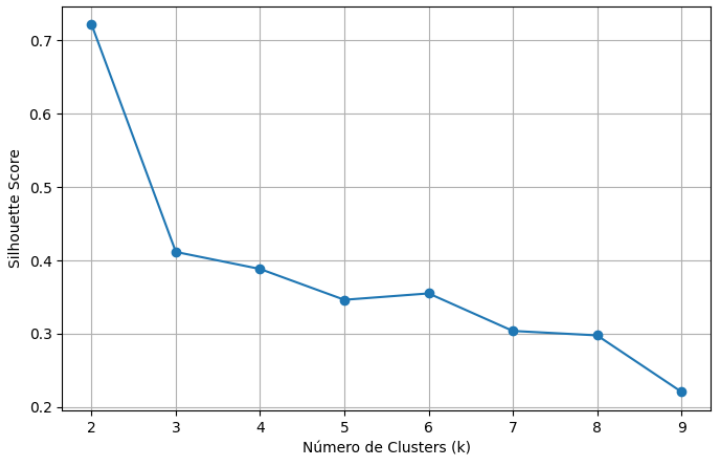

Figura 5.31: Gráfica Silhouette Score para matriz de distancia calculada a partir del primer caso de la metodología.

Tabla 5.11: Silhouette Score para matriz de distancia calculada a partir del segundo caso de la metodología.

Clusters	Silhouette Score
2	0.7052
3	0.4701
4	0.4716
5	0.4359
6	0.4312
7	0.3753
8	0.2234
9	0.2111

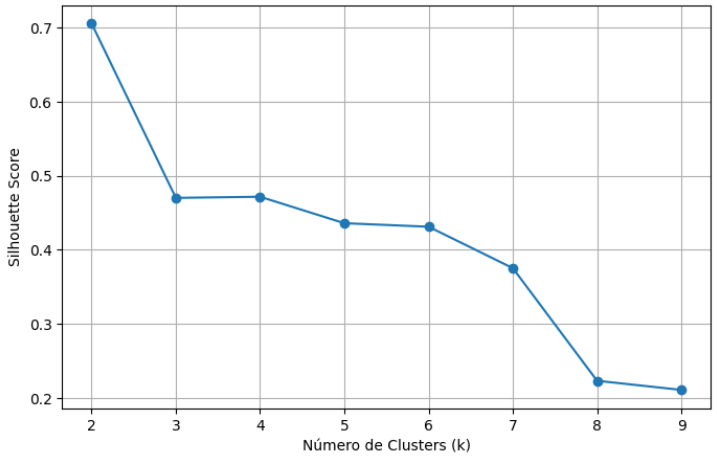

Figura 5.32: Gráfica Silhouette Score para matriz de distancia calculada a partir del segundo caso de la metodología.

Tabla 5.12: Silhouette Score para matriz de distancia obtenida en el estudio.

Clusters	Silhouette Score
2	0.7877
3	0.7313
4	0.4075
5	0.3608
6	0.3892
7	0.3866
8	0.3454
9	0.2886

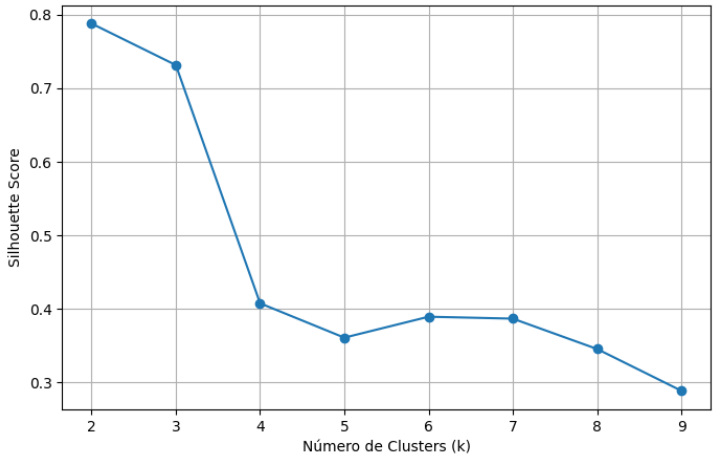

Figura 5.33: Gráfica Silhouette Score para matriz de distancia obtenida en el estudio.

5.2.3.2. Clasificación

En las Tablas 5.13 y 5.14 se presentan las clasificaciones obtenidas al aplicar el algoritmo *Agglomerative Clustering* a las matrices de distancias calculadas mediante ambos casos de la tercera versión de la metodología. En general, como se puede apreciar en ambas tablas, los aeropuertos de mayor tamaño o con mayor tráfico, como Madrid-Barajas (LEMD) y Barcelona-El Prat (LEBL), tienden a agruparse por separado, formando conjuntos diferenciados del resto. Esto resulta coherente, ya que el elevado volumen de vuelos que manejan, tanto nacionales como internacionales, incrementa la complejidad de su espacio aéreo y, en consecuencia, la probabilidad de desviaciones y retrasos debido a la saturación.

Por otro lado, aeropuertos como Ibiza (LEIB), Tenerife Sur (GCTS), Gran Canaria (GCLP) y Palma de Mallorca (LEPA) también tienden a agruparse en conjuntos propios. Esto indica que presentan un comportamiento diferenciado respecto al resto, ya que son aeropuertos con un volumen de tráfico elevado, especialmente durante los meses de verano o vacaciones. Al tratarse de destinos turísticos, es razonable esperar una mayor variabilidad en sus trayectorias y, en consecuencia, un incremento en las desviaciones y retrasos.

No obstante, un caso particular es el del aeropuerto de Zaragoza (LEZG), que, a pesar de no contar con un volumen de tráfico especialmente alto, tiende a clasificarse aparte de los aeropuertos más pequeños. Esto podría estar relacionado con condiciones geográficas específicas que afectan las trayectorias de llegada o salida, generando desviaciones más pronunciadas.

Los aeropuertos de menor tamaño (por ejemplo, Sabadell, León o Huesca Pirineos) se agrupan con frecuencia en conjuntos comunes, lo cual es consistente con su bajo volumen de operaciones y una menor complejidad operativa, generalmente menos afectada por la congestión.

En resumen, se observa una tendencia clara en la clasificación: los aeropuertos pequeños tienden a agruparse entre sí; algunos aeropuertos turísticos se agrupan juntos en grupos separados del resto; el aeropuerto de Zaragoza, aunque no es ni turístico ni congestionado, también presenta una diferencia con respecto al resto de

aeropuertos pequeños; y aeropuertos como Madrid-Barajas y Barcelona-El Prat aparecen frecuentemente de forma aislada, reflejando su carácter singular en cuanto a volumen y complejidad operativa.

Por otra parte, la clasificación obtenida utilizando la matriz de distancias calculada según los resultados del estudio realizado por Cuerno *et al.* 2023 es la que se puede apreciar en la Tabla 5.15.

Al comparar las clasificaciones obtenidas en este TFM (Tablas 5.13 y 5.14) con la clasificación derivada de las distancias del artículo original (Tabla 5.15), se observa que las diferencias, aunque presentes, no son significativamente elevadas en relación con el total de 45 aeropuertos analizados.

En el caso de la clasificación basada en la matriz de distancias construida a partir de la nube de puntos con únicamente valores positivos (Tabla 5.13), se identifican discrepancias en 10 aeropuertos respecto a la clasificación del artículo. Estos aeropuertos son: Gran Canaria (GCLP), Ibiza (LEIB), Albacete (LEAB), León (LELN), Alicante (LEAL), Málaga (LEMG), Valencia (LEVC), Fuerteventura (GCFV), La Palma (GCLA) y Lanzarote (GCRR). Esta diferencia representa aproximadamente el 22 % del total de aeropuertos.

Por otro lado, la clasificación obtenida a partir de la matriz de distancias que incorpora valores negativos cuando el retraso también es negativo (Tabla 5.14) difiere del artículo en 11 aeropuertos: Gran Canaria (GCLP), Palma de Mallorca (LEPA), Zaragoza (LEZG), Albacete (LEAB), León (LELN), Alicante (LEAL), Málaga (LEMG), Valencia (LEVC), Fuerteventura (GCFV), La Palma (GCLA) y Lanzarote (GCRR). Esto equivale al 24 % del total de aeropuertos.

Estas diferencias son esperables y coherentes con los ajustes introducidos en esta tercera versión de la metodología, donde se perfeccionó el cálculo de las distancias entre trayectorias planificadas y reales. Este refinamiento busca una representación más precisa del comportamiento operativo de los aeropuertos, en contraste con el enfoque más simplificado del estudio original. A pesar de estas variaciones, se puede concluir que la clasificación general se mantiene bastante consistente.

Adicionalmente, al comparar las dos clasificaciones obtenidas por cada caso de la metodología (Tablas 5.13 y 5.14), se observa que las diferencias son mínimas. En particular, solo se identifican discrepancias en cuatro aeropuertos: Gran Canaria (GCLP), Palma de Mallorca (LEPA), Ibiza (LEIB) y Zaragoza (LEZG). Esto sugiere que la forma de representar la nube de puntos, ya sea considerando únicamente las distancias positivas o incluyendo también aquellas negativas cuando el retraso es negativo, no influye de manera significativa en los resultados de la clasificación.

Por último, en las Figuras 5.34 y 5.35 se muestran los grafos obtenidos mediante el algoritmo *Mapper* combinado con *DBSCAN* y *HDBSCAN*, respectivamente, extraídos del estudio de Serrano (2025). Estas representaciones permiten observar de forma global cómo se agrupan los aeropuertos españoles según su comportamiento. Al comparar estos grafos con las clasificaciones obtenidas en este TFM (ver Tablas 5.13 y 5.14), se puede ver que, en el caso del grafo con *DBSCAN* (Figura 5.34), aeropuertos como Madrid-Barajas (LEMD), Palma de Mallorca (LEPA), Barcelona-El Prat (LEBL), Ibiza (LEIB) o Zaragoza (LEZG) aparecen agrupados en nodos juntos, lo que coincide con los resultados de este TFM, donde suelen diferenciarse del resto por su actividad más intensa o particular.

Tabla 5.13: Clasificación de aeropuertos obtenida a partir de la matriz de distancias calculada según el primer caso de la metodología.

Cluster	Aeropuerto	Código	Grupo AENA
0	Madrid-Cuatro Vientos	LECU	Group 3
0	Sabadell	LELL	Group 3
0	Albacete	LEAB	Group 3
0	Badajoz	LEBZ	Group 3
0	León	LELN	Group 3
0	Salamanca	LESA	Group 3
0	Son Bonet	LESB	Group 3
0	Valladolid	LEVD	Group 3
0	Melilla	GEML	Group 3
0	Burgos-Villafría	LEBG	Group 3
0	Córdoba	LEBA	Group 3
0	Girona	LEGE	Group 3
0	Huesca-Pirineos	LEHC	Group 3
0	Logroño-Agoncillo	LERJ	Group 3
0	Pamplona	LEPP	Group 3
0	San Sebastián	LESO	Group 3
0	Vitoria	LEVT	Group 3
0	A Coruña	LECO	Group 2
0	Almería	LEAM	Group 2
0	Asturias	LEAS	Group 2
0	Granada-Jaén	LEGR	Group 2
0	Jerez	LEJR	Group 2
0	Reus	LERS	Group 2
0	Santander	LEXJ	Group 2
0	Santiago	LEST	Group 2
0	Vigo	LEVX	Group 2
0	Alicante-Elche-Miguel Hernández	LEAL	Group 1
0	Bilbao	LEBB	Group 1
0	Málaga-Costa del Sol	LEMG	Group 1
0	Menorca	LEMH	Group 1
0	Sevilla	LEZL	Group 1
0	Valencia	LEVC	Group 1
0	Fuerteventura	GCFV	Canary Group
0	La Gomera	GCGM	Canary Group
0	El Hierro	GCHI	Canary Group
0	La Palma	GCLA	Canary Group
0	César Manrique Lanzarote	GCRR	Canary Group
0	Tenerife North	GCXO	Canary Group
0	Palma Mallorca	LEPA	Special Group
2	Zaragoza	LEZG	Group 2
2	Ibiza	LEIB	Group 1
2	Tenerife South	GCTS	Canary Group
4	Gran Canaria	GCLP	Canary Group
3	Madrid-Barajas	LEMD	Special Group
1	Barcelona-El Prat	LEBL	Special Group

Tabla 5.14: Clasificación de aeropuertos obtenida a partir de la matriz de distancias calculada según el segundo caso de la metodología.

Cluster	Aeropuerto	Código	Grupo AENA
2	Madrid-Cuatro Vientos	LECU	Group 3
2	Sabadell	LELL	Group 3
2	Albacete	LEAB	Group 3
2	Badajoz	LEBZ	Group 3
2	León	LELN	Group 3
2	Salamanca	LESA	Group 3
2	Son Bonet	LESB	Group 3
2	Valladolid	LEVD	Group 3
2	Melilla	GEML	Group 3
2	Burgos-Villafría	LEBG	Group 3
2	Córdoba	LEBA	Group 3
2	Girona	LEGE	Group 3
2	Huesca-Pirineos	LEHC	Group 3
2	Logroño-Agoncillo	LERJ	Group 3
2	Pamplona	LEPP	Group 3
2	San Sebastián	LESO	Group 3
2	Vitoria	LEVT	Group 3
2	A Coruña	LECO	Group 2
2	Almería	LEAM	Group 2
2	Asturias	LEAS	Group 2
2	Granada-Jaén	LEGR	Group 2
2	Jerez	LEJR	Group 2
2	Reus	LERS	Group 2
2	Santander	LEXJ	Group 2
2	Santiago	LEST	Group 2
2	Vigo	LEVX	Group 2
2	Zaragoza	LEZG	Group 2
2	Alicante-Elche-Miguel Hernández	LEAL	Group 1
2	Bilbao	LEBB	Group 1
2	Ibiza	LEIB	Group 1
2	Málaga-Costa del Sol	LEMG	Group 1
2	Menorca	LEMH	Group 1
2	Sevilla	LEZL	Group 1
2	Valencia	LEVC	Group 1
2	Fuerteventura	GCFV	Canary Group
2	La Gomera	GCGM	Canary Group
2	El Hierro	GCHI	Canary Group
2	La Palma	GCLA	Canary Group
2	César Manrique Lanzarote	GCRR	Canary Group
2	Tenerife North	GCXO	Canary Group
0	Gran Canaria	GCLP	Canary Group
0	Tenerife South	GCTS	Canary Group
3	Madrid-Barajas	LEMD	Special Group
4	Barcelona-El Prat	LEBL	Special Group
1	Palma Mallorca	LEPA	Special Group

Tabla 5.15: Clasificación de los aeropuertos según la matriz de distancias obtenida en el estudio.

Cluster	Aeropuerto	Código	Grupo AENA
0	Madrid-Cuatro Vientos	LECU	Group 3
0	Sabadell	LELL	Group 3
0	Badajoz	LEBZ	Group 3
0	Salamanca	LESA	Group 3
0	Son Bonet	LESB	Group 3
0	Valladolid	LEVD	Group 3
0	Melilla	GEML	Group 3
0	Burgos-Villafría	LEBG	Group 3
0	Córdoba	LEBA	Group 3
0	Girona	LEGE	Group 3
0	Huesca-Pirineos	LEHC	Group 3
0	Logroño-Agoncillo	LERJ	Group 3
0	Pamplona	LEPP	Group 3
0	San Sebastián	LESO	Group 3
0	Vitoria	LEVT	Group 3
0	A Coruña	LECO	Group 2
0	Almería	LEAM	Group 2
0	Asturias	LEAS	Group 2
0	Granada-Jaén	LEGR	Group 2
0	Jerez	LEJR	Group 2
0	Reus	LERS	Group 2
0	Santander	LEXJ	Group 2
0	Santiago	LEST	Group 2
0	Vigo	LEVX	Group 2
0	Bilbao	LEBB	Group 1
0	Ibiza	LEIB	Group 1
0	Menorca	LEMH	Group 1
0	Sevilla	LEZL	Group 1
0	La Gomera	GCGM	Canary Group
0	El Hierro	GCHI	Canary Group
0	Tenerife North	GCXO	Canary Group
0	Palma Mallorca	LEPA	Special Group
1	Albacete	LEAB	Group 3
1	León	LELN	Group 3
1	Alicante-Elche	LEAL	Group 1
1	Málaga-Costa del Sol	LEMG	Group 1
1	Valencia	LEVC	Group 1
1	Fuerteventura	GCFV	Canary Group
1	La Palma	GCLA	Canary Group
1	Gran Canaria	GCLP	Canary Group
1	Lanzarote	GCRR	Canary Group
2	Madrid-Barajas	LEMD	Special Group
3	Barcelona-El Prat	LEBL	Special Group
4	Zaragoza	LEZG	Group 2
4	Tenerife South	GCTS	Canary Group

En el grafo obtenido con *HDBSCAN* (Figura 5.35), también se puede ver que Madrid, Barcelona, y Palma de Mallorca forman parte de grupos pequeños o conectados entre sí con otros dos aeropuertos.

Los aeropuertos más pequeños y con menos tráfico, como El Hierro (GCHI), Melilla (GEML), La Gomera (GCGM), Córdoba (LEBA) o Huesca (LEHC), aparecen como nodos aislados o periféricos en el grafo con *DBSCAN*, mientras que en el grafo con *HDBSCAN* muchos de ellos ni siquiera aparecen representados, ya que el algoritmo no los incluye si los detecta como un nodo solo. Esto tiene sentido, ya que estos aeropuertos muestran un comportamiento muy diferente al de los grandes como Madrid, Palma o Barcelona, y esto también se refleja en las clasificaciones hechas en este TFM.

Por otro lado, hay algunas diferencias. En ambos grafos, sobre todo en el de *HDBSCAN*, se ve que aeropuertos de tamaño medio como Bilbao, Santander, Asturias, Valencia, Vigo o Sevilla aparecen separados en grupos propios. Esto no se observa tan claramente en las clasificaciones obtenidas en este TFM, por lo que los grafos *Mapper* podrían estar captando relaciones diferentes entre estos aeropuertos, quizá por su ubicación o estacionalidad.

Sin embargo, se puede observar que, en general, tanto en el estudio realizado por Cuerno *et al.* (2023), como en el de Serrano (2025), así como en el presente trabajo, se presenta una idea o comportamiento similar: los aeropuertos como Madrid-Barajas, Barcelona, Palma de Mallorca, Tenerife Sur y Zaragoza suelen estar claramente diferenciados del resto de aeropuertos de menor tamaño, como Huesca, La Gomera, Córdoba, Sabadell, entre otros. Esto demuestra que, aunque las representaciones o clasificaciones no son exactamente iguales, los estudios comparten una estructura similar y capturan la misma idea en cuanto al comportamiento de los aeropuertos según sus vuelos.

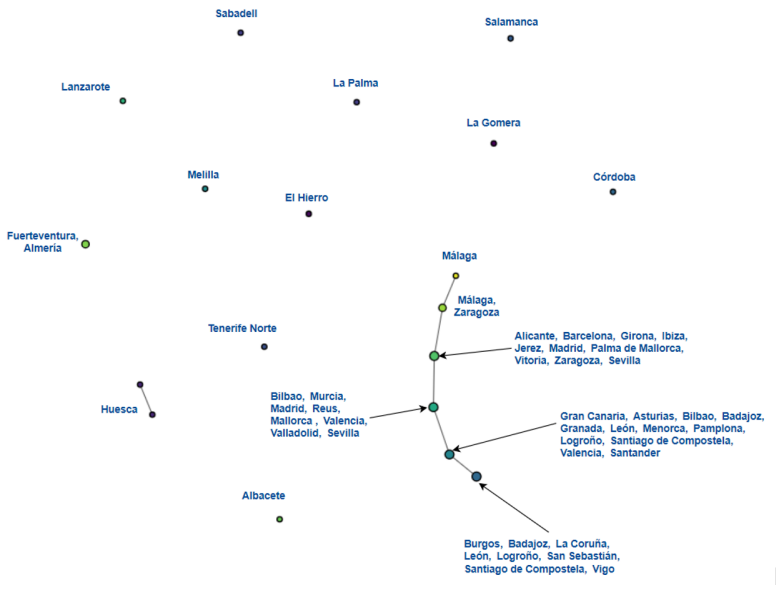

Figura 5.34: Grafo resultante del algoritmo *Mapper* junto el algoritmo *DBSCAN*.

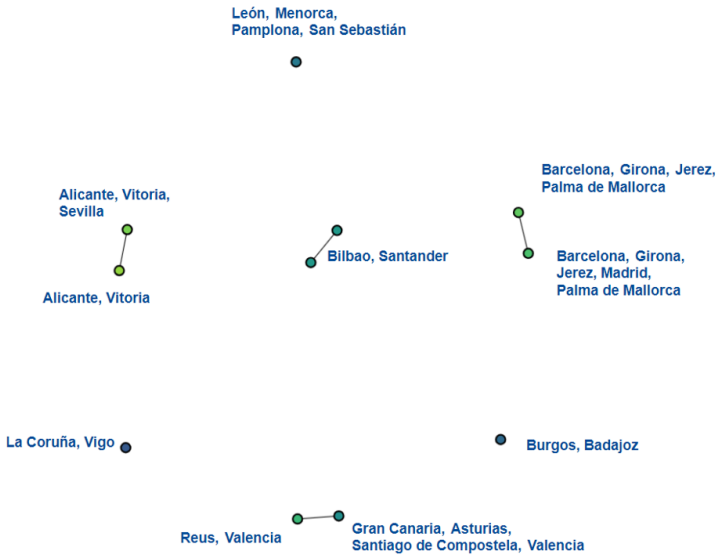

Figura 5.35: Grafo resultante del algoritmo *Mapper* junto el algoritmo *HDBSCAN*.

Capítulo 6

Conclusión

El presente Trabajo de Fin de Máster (TFM) ha explorado la aplicación del análisis topológico de datos (*Topological Data Analysis*, TDA), en particular la homología persistente, como herramienta para clasificar aeropuertos españoles en función del comportamiento de sus vuelos. A través de la adaptación de la metodología realizada por otros estudios, su refinamiento y combinación con técnicas de aprendizaje automático, se ha implementado una metodología para analizar, representar y comparar el comportamiento operativo de los aeropuertos, considerando tanto los retrasos de llegada como las distancias entre las trayectorias planificadas y las reales presentes en sus vuelos.

Además, a partir del análisis realizado a los datos, se han introducido mejoras significativas en la metodología, en especial en el cálculo de la distancia entre trayectorias mediante un ajuste temporal más preciso. Esto ha permitido obtener una representación más detallada de las distancias entre las trayectorias de los vuelos correspondientes a cada aeropuerto, mejorando así la calidad de la información sobre la cual se realiza el análisis topológico.

Por otra parte, el trabajo ha mostrado que el uso del análisis topológico permite capturar estructuras globales y patrones recurrentes en los datos de vuelo que no son fácilmente detectables mediante enfoques tradicionales. En particular, se ha observado que las clasificaciones obtenidas difieren de las establecidas oficialmente por organismos como AENA, lo cual sugiere que el enfoque basado en TDA aporta una perspectiva complementaria, centrada en el comportamiento real y funcional del tráfico aéreo. Aeropuertos como Madrid, Barcelona, Palma de Mallorca, Zaragoza, Gran Canaria, Tenerife Sur e Ibiza han sido identificados como instalaciones con un comportamiento operativo significativamente diferente al resto.

Asimismo, se ha observado que, a pesar de las modificaciones introducidas, la clasificación obtenida difiere muy poco de la clasificación derivada a partir de las distancias utilizadas en el estudio de Cuerno *et al.* (2023), lo cual valida los resultados y refuerza la robustez de la metodología propuesta.

6.1. Limitaciones

No obstante, durante el desarrollo de este TFM se han presentado diversas limitaciones. En primer lugar, una de las principales restricciones es que solo se ha utilizado

un algoritmo de clustering para la clasificación de aeropuertos. Esto se debe a que se requería un algoritmo que aceptara matrices de distancia como entrada y que permitiera establecer como parámetro el número de clústeres a obtener, lo cual reduce el número de algoritmos disponibles para su uso.

Por otra parte, se valoró la posibilidad de aplicar escalado multidimensional (*Multidimensional Scaling*, MDS) para transformar la matriz de distancias en un espacio euclídeo bidimensional, de modo que cada aeropuerto pudiera representarse como un punto y las distancias entre ellos fueran proporcionales a las de la matriz original. Esta transformación habría permitido emplear algoritmos que operan directamente sobre espacios vectoriales, como *K-Means*. Sin embargo, el uso de MDS puede implicar una pérdida significativa de información topológica, ya que la reducción de dimensionalidad no siempre conserva fielmente la estructura de distancias original, lo cual afectaría negativamente la calidad de la clasificación.

Por último, también existen limitaciones inherentes a la calidad y completitud de los datos disponibles. El proceso de preprocesamiento, interpolación de trayectorias y eliminación de vuelos con información errónea ha sido necesario para garantizar la coherencia del análisis, pero estas acciones pueden haber introducido sesgos o provocado la pérdida de información relevante, afectando potencialmente a la representatividad de los resultados.

6.2. Líneas de trabajo futuro

En cuanto a posibles líneas de trabajo futuro, una de las más prometedoras sería el uso de distintos tipos de filtrados para la aplicación del análisis topológico, como el *Alpha Complex*. Aunque esta opción es más costosa computacionalmente, ofrece una representación más precisa de la geometría de los datos y podría aportar diferencias significativas en comparación con el complejo de Vietoris–Rips. Evaluar cómo varía la clasificación de aeropuertos al cambiar el tipo de filtración permitiría validar la sensibilidad del modelo ante este tipo de decisiones metodológicas.

Otra línea futura sería explorar otras técnicas de análisis topológico de datos, como el uso de imágenes de persistencia (*Persistence Images*), una forma alternativa de representar diagramas de persistencia en un espacio vectorial. Estas representaciones permitirían emplear modelos adicionales de aprendizaje automático y podrían contribuir a mejorar el rendimiento de la clasificación.

Además, se propone continuar investigando el uso de técnicas topológicas en el análisis de vuelos y en el ámbito de la aviación en general. El análisis topológico de datos ha demostrado ser una herramienta útil para capturar patrones complejos, y su potencial en este dominio aún no ha sido explorado en profundidad. Ampliar su aplicación a otros contextos, como la predicción de congestiones o el estudio de rutas alternativas, podría abrir nuevas oportunidades de investigación tanto teóricas como aplicadas.

Bibliografía

Airports Council International (ACI). (2024). Annual World Airport Traffic Forecasts 2024–2053 [Montreal: Airports Council International].

Bhattacharya, S., Ghrist, R., y Kumar, V. (2014). Multi-robot coverage and exploration on Riemannian manifolds with boundaries. *The International Journal of Robotics Research*, *33*(1), 113-137.

Bhattacharya, S., Ghrist, R., y Kumar, V. (2015). Persistent homology for path planning in uncertain environments. *IEEE Transactions on Robotics*, *31*(3), 578-590.

Bhattacharya, S., Lipsky, D., Ghrist, R., y Kumar, V. (2013). Invariants for homology classes with application to optimal search and planning problem in robotics. *Annals of Mathematics and Artificial Intelligence*, *67*(3), 251-281.

Bian, J., Tian, D., Tang, Y., y Tao, D. (2018). A survey on trajectory clustering analysis. *arXiv preprint arXiv:1802.06971*.

Bolić, T., Castelli, L., De Lorenzo, A., y Vascotto, F. (2022). Trajectory clustering for air traffic categorisation. *Aerospace*, *9*(5), 227.

Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, *16*(1), 77-102.

Burmester, G., Ma, H., Steinmetz, D., y Hartmannn, S. (2018). Big data and data analytics in aviation. *Advances in Aeronautical Informatics: Technologies Towards Flight 4.0*, 55-65.

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, *46*(2), 255-308.

Chazal, F., y Michel, B. (2021). An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, *4*, 667963.

Cuerno, M. (2025). Data and information for TDA in ATM [Dataset]. https://doi.org/10.6084/m9.figshare.28232714.v1

Cuerno, M., Guijarro, L., Valdés, R. M. A., y Comendador, F. G. (2023). Topological Data Analysis in ATM: the shape of big flight data sets. *arXiv preprint arXiv:2304.08906*.

Domínguez Monterroza, A., Mateos Caballero, A., y Jiménez Martín, A. (2023). Topological Data Analysis to Characterize Fluctuations in the Latin American Integrated Market. *Workshop on Engineering Applications*, 195-203.

Edelsbrunner, H., Letscher, D., y Zomorodian, A. (2002). Topological persistence and simplification. *Discrete & Computational Geometry*, *28*, 511-533.

Edelsbrunner, H., y Morozov, D. (2013). *Persistent homology: theory and practice*. eScholarship, University of California.

Gariel, M., Srivastava, A. N., y Feron, E. (2011). Trajectory clustering and an application to airspace monitoring. *IEEE Transactions on Intelligent Transportation Systems*, *12*(4), 1511-1524.

Ghrist, R. (2008). Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, *45*(1), 61-75.

Giusti, C., Ghrist, R., y Bassett, D. S. (2016). Two's company, three (or more) is a simplex: Algebraic-topological tools for understanding higher-order structure in neural data. *Journal of Computational Neuroscience*, *41*, 1-14.

Huang, L., Zhang, S., Zhang, Y., Zhang, Y., y Yin, Y. (2024). Aircraft landing time prediction with deep learning on trajectory images. *arXiv preprint arXiv:2401.01083*.

Li, M. Z., Ryerson, M. S., y Balakrishnan, H. (2019). Topological data analysis for aviation applications. *Transportation Research Part E: Logistics and Transportation Review*, *128*, 149-174.

Li, Q., y Jing, R. (2021). Characterization of delay propagation in the air traffic network. *Journal of Air Transport Management*, *94*, 102075.

Liu, Y., y Hansen, M. (2018). Predicting aircraft trajectories: A deep generative convolutional recurrent neural networks approach. *arXiv preprint arXiv:1812.11670*.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65.

Serrano, R. (2025). *Aprendizaje Topológico para Analizar Trayectorias de Aeronaves* [Trabajo Fin de Máster]. Escuela Técnica Superior de Ingenieros Informáticos, Universidad Politécnica de Madrid.

Singh, G., Mémoli, F., y Carlsson, G. (2007). Topological methods for the analysis of high dimensional data sets and 3d object recognition. *Eurographics Symposium on Point-Based Graphics*, *2*, 091-100.

Wang, Z., Liang, M., y Delahaye, D. (2017). Short-term 4d trajectory prediction using machine learning methods. *SID 2017, 7th SESAR Innovation Days*.

Yoon, S., y Lee, K. (2025). Aircraft Trajectory Dataset Augmentation in Latent Space. *arXiv preprint arXiv:2506.07585*.

Zeng, W., Xu, Z., Cai, Z., Chu, X., y Lu, X. (2021). Aircraft trajectory clustering in terminal airspace based on deep autoencoder and gaussian mixture model. *Aerospace*, *8*(9), 266.

Zhou, Y., Wang, J., y Huang, G. Q. (2019). Efficiency and robustness of weighted air transport networks. *Transportation Research Part E: Logistics and Transportation Review*, *122*, 14-26.