PLOS ONE

RESEARCH ARTICLE

Topological data analysis in air traffic management: The shape of big flight data sets

Manuel Cuerno[1,2*‡], **Luis Guijarro**[2], **Rosa María Arnaldo Valdés**[3], **Fernando Gómez Comendador**[3]

1 Department of Mathematics, CUNEF University, Madrid, Spain, **2** Department of Mathematics, Universidad Autónoma de Madrid and ICMAT CSIC-UAM-UCM-UC3M, Madrid, Spain, **3** Department of Aerospace Systems, Air Transportation and Airports, E.T.S.I. Aeronáutica y del Espacio, Universidad Politécnica de Madrid, Madrid, Spain

‡ Current address: Department of Mathematics, CUNEF University, Madrid, Spain
* manuel.mellado@cunef.edu

Abstract

Analyzing flight trajectory data sets poses challenges due to the intricate interconnections among various factors and the high dimensionality of the data. Topological Data Analysis (TDA) is a way of analyzing big data sets focusing on the topological features this data sets have as point clouds in some metric space. Techniques as the ones that TDA provides are suitable for dealing with high dimensionality and intricate interconnections. This paper introduces TDA and its tools and methods as a way to derive meaningful insights from ATM data. Our focus is on employing TDA to extract valuable information related to airports. Specifically, by utilizing persistence landscapes (a potent TDA tool) we generate footprints for each airport. These footprints, obtained by averaging over a specific time period, are based on the deviation of trajectories and delays. We apply this method to the set of Spanish' airports in the Summer Season of 2018. Remarkably, our results align with the established Spanish airport classification and raise intriguing questions for further exploration. This analysis serves as a proof of concept, showcasing the potential application of TDA in the ATM field. While previous works have outlined the general applicability of TDA in aviation, this paper marks the first comprehensive application of TDA to a substantial volume of ATM data. Finally, we present conclusions and guidelines to address future challenges in the ATM domain.

Citation: Cuerno M, Guijarro L, Valdés R M A, Comendador F G. (2025) Topological data analysis in air traffic management: The shape of big flight data sets. PLoS ONE 20(2): e0318108. https://doi.org/10.1371/journal.pone.0318108

Editor: Lazaros K. Gallos, Rutgers The State University of New Jersey, UNITED STATES OF AMERICA

Received: May 29, 2024

Accepted: December 18, 2024

Published: February 27, 2025

Data availability statement: Data cannot be shared publicly because of it belongs to Eurocontrol's software NEST. This program has not public access, it is only under permission of Eurocontrol. CRIDA provided us the datasets we used for the paper.

Funding: Manuel Cuerno was supported in part by the Formación del Personal Investigador Graduate Research Grant (PRE2018-084109). He and Luis Guijarro were also supported in

Introduction

Airports and Air Traffic Management Systems form intricate sociotechnical structures, presenting challenges in analysis due to their high interdependence. Globally, these systems manage operations involving over 32,000 in-service aircraft operated by more than 1,300 airlines, facilitating over 4.1 billion passengers across 41.9 million flights and 45,000 routes at 3,700 airports [1,2].

part by research grants from the Ministerio de Economía y Competitividad de España (MTM2017-?85934-C3-2-P and PID2021-124195NB-C32). Finally, this work has also been supported by the Department of Mathematics of CUNEF University, Madrid, Spain.

Competing interests: The authors have declared that no competing interests exist.

The interconnectedness, interdependences and complexity of the system is reflected in the amount of data generated by its operation. Although flight trajectory data offer a big potential to grasp the features and behaviour of such complex system, analyzing flight trajectory data proves challenging due to its high dimensionality, continuous nature, and multi-level interactions influenced by factors such as air traffic control, weather conditions, aircraft type, and pilot behaviour [3]. The dynamic changes and multiple variables involved, like altitude, speed, heading, longitude, latitude, and time, further complicate manual processing [4–6]. Its high dimensionality can make it difficult to visualize or analyse the data without reducing its dimensionality through techniques like principal component analysis or t-SNE.

To address the high dimensionality issue, dimensionality reduction techniques such as principal component analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE) are employed to identify common traffic patterns [7–13]. However, limitations exist, including empirical selection of techniques and potential loss of relevant information, potential lack of interpretability, and issues of generalizability to different contexts. The choice between PCA or t-SNE in the analysis algorithm is often made through empirical observations, and alternative machine learning techniques might yield superior results for specific data types or traffic patterns. While PCA and t-SNE can uncover data patterns, the interpretation of these patterns poses challenges, requiring additional domain expertise for meaningful insights. The studies conducted may lack generalizability to different airports or contexts, necessitating further research to assess the applicability of these techniques in diverse settings. Moreover, some studies encounter issues related to limited sample size, data quality, or pre-processing, and there's always a risk of losing relevant information during dimensional reduction.

Clustering techniques like k-means, hierarchical clustering, and DBSCAN are alternative methods to identify common traffic patterns [10]. In [14], a trajectory clustering ensemble method is introduced, utilizing a similarity matrix and employing the Nanjing Lukou Airport terminal area as an illustrative example. In the study by Zeng et al [15], a DBSCAN clustering analysis method is employed for the detection of outliers in aircraft airborne and controllers' data, contributing to the assessment and monitoring of flight status. Furthermore, a paper at the SESAR Innovation Days 2017 [16] proposed a method for analyzing the traffic patterns of civil aviation flights using principal component analysis (PCA) and DBSCAN clustering. The method was tested on a dataset of flight trajectory data from the Beijing Capital International Airport, and the results showed that the method was able to identify common traffic patterns and distinguish them from those abnormal.

While effective, they have limitations such as sensitivity to initialization, difficulty in determining the number of clusters and handling noise, scalability and interpretability issues and outlier sensitivity. Additionally these methods are limited to Euclidean distances (k-means and hierarchical clustering), to finding geometric structures and to unsupervised learning.

Beyond the previously mentioned methods, machine learning algorithms, including decision trees, random forests, and neural networks, are also utilized for pattern identification in flight trajectory data. However, limitations include sensitivity to data availability, lack of standardization, and challenges in result interpretation [17]. As can be seen, overall, while the existing research works have made significant contributions to the field of analysing high-dimensional flight trajectory data, there are still limitations that need to be addressed in future research to improve the accuracy, reliability, and scalability of these methods.

To overcome existing limitations, this paper proposes the use of Topological Data Analysis (TDA), a mathematical framework using tools from algebraic topology. TDA offers advantages in identifying topological features, simplifying complex data, handling noise, comparing datasets, and building predictive models [18].

A key strength of TDA lies in its ability to identify underlying structures and patterns in data with enhanced resilience to noise and outliers. For instance, Casacuberta et al. [19] demonstrated this by estimating the intrinsic dimension of a point cloud in a high-dimensional space, revealing that the real dimension of the data was less than that of the ambient space. TDA also proves beneficial in mitigating the challenge of limited data availability, enabling the integration of diverse data sources and extracting meaningful insights from incomplete or noisy datasets. Furthermore, TDA offers a solution to scalability issues by facilitating the analysis of large and complex datasets through parallel computing techniques. Additionally, TDA addresses concerns related to the lack of standardization by providing a flexible and adaptable framework applicable to a broad spectrum of data formats and types, and it works no matter the initial domain is and can give conclusions only based on the data input no matter where this data comes from. This allows the researcher extract some initial conclusions excluding the domain of the data.

TDA applications include identifying topological features, simplifying complex data, handling noise, comparing datasets, and building predictive models [19,20]. In addition, other authors have also applied different topological tools for air traffic management problems. Examples include detecting anomalous trajectories and assessing delay network robustness in adverse weather conditions [17], predictive modelling [21] or more topological graph-based studies related to finding patterns in different air traffic situations [22].

Overall, TDA offers a powerful tool, either alone or combined with other existent methods, for analysing complex and high-dimensional aviation data sets. By identifying topological features and patterns, TDA can reveal hidden relationships and help airlines and airports making better decisions about flight scheduling, maintenance, and safety as well as improving airport operations and the passenger experience.

In this work, we present a proof of concept of TDA applied to the Air Traffic and Transportation Management field. In particular, we construct a point cloud based on spatiotemporal and delay flight-data of a certain airport in order to develop, using TDA techniques, a footprint of it. After this process, we compare the actual classification of the Spanish Airport net (only based on number of passengers per year) with the one we produce with the footprints and compute a *block permutation test* in order to calculate the statistical significance of our experiment. Additionally, we compare our results with ones produced by a non-TDA method based on centrality measures. Interesting similarities and discrepancies appeared on this experiment, showing how TDA can enrich the study of the ATM in future projects.

To the best of our knowledge, there have been generic works outlining the possible application of TDA in aviation [23,24] but none of them have applied any TDA technique apart from computing some simplicial complexes based on airports [24] or some sensor analysis partially related with aircraft search [23]. No rigorous work has already been conducted to apply TDA to a large volume of aircraft trajectory data for the purpose of anticipating and identifying deviations and anomalies in aircraft space/time trajectories.

Additionally, there is a lack of research aimed at inferring patterns of behaviour at different airports or classifying and characterizing airports based on the distribution of their daily flights through trajectory deviation and delay. Despite this gap in the literature, some works based on centrality measures and graph theory have been developed to globally classify networks of airports [25–28].

The application of TDA techniques, and other types of mathematical analysis algorithms applied to data series, requires close collaboration between research groups focused on mathematical modelling and analysis, and research groups focused on different fields of application. This means that certain lines identified as potentially of interest may lack continuity. In this case, other references have been identified that show how some groups, which have initiated work in these areas, maintain interest in this application and continue to develop various projects specifically [29–33]. Moreover, TDA and other topological tools are also being applied in other areas of transport, not only in aviation, that surely could improve its implementation in the ATM field [23,34–36].

This paper proposes a method for assessing the structural characteristics of air traffic situation based on TDA, providing new clues to give a more precise description of air traffic complexity, to assess the deviation from expected aircraft trajectories, to study the generation of delays, to identify structural patterns, and to detect and analyze anomalies in airport operations. In particular this method provides a natural classification of airports attending to operational characteristics. The method adjusts and captures, from a mathematical point of view, the grouping criteria that the airport provider uses operationally. It mimics those criteria. It has the potential to capture more complex operational criteria and relationships.

Furthermore, this initial work wants to encourage the future use of TDA in the aerospace field as a powerful tool to deal with the existent data constrains and enrich its combination with already known methods obtaining the optimal use of all of them.

Materials and methods

Topological Data Analysis [18,37,38] employs a set of structures, named simplicial complexes, for the purpose of data analysis. Roughly speaking, if the data is represented by a point cloud, we can endow it with a simplicial structure and study how it evolves when we vary certain scaling parameters. This is the setting of the one-parameter TDA, which is the one most commonly used, and the one we are going to implement in this work.

While multi-parameter TDA is an attractive field with promising results [39,40], it is computationally challenging and thus most approaches are theoretical. However, the study of multi-parameter TDA could be an interesting setting to further explore problems in this field allowing us to include more factors in the analysis, as we have explained in the Introduction of this paper.

The primary objective of this Section is to offer introductory concepts of one-parameter TDA and the tools and methods we use in this paper. For the sake of clarity and concision, we will not present a more formal description of simplicial complexes, homology and persistent homology. For the interested reader we recommend [18,37] and, a mathematical introduction to homology can be found in [41,42].

Persistent homology

We will consider a *point cloud* $\mathcal{V}$ as a set of points in $\mathbb{R}^d$ for some $d \geq 2$. Roughly speaking, TDA would extract the most relevant "topological features", called n-cycles, of $\mathcal{V}$. The n-cycles represent n-dimensional holes. For example, $\mathbb{S}^1 \subset \mathbb{R}^2$ has one 1-cycle. In order to obtain the information of the n-cycles in our point cloud, TDA usually computes the *persistent homology* of $\mathcal{V}$. It will measure how predominant those n-cycles are.

To extract the persistent homology of the point cloud $\mathcal{V}$, we first need to construct a *filtration* based on $\mathcal{V}$. A commonly used method for this is the *Vietoris-Rips filtration*, denoted as $VR(\mathcal{V})$. This filtration is a nested sequence of simplicial complexes ($VR_{r_0}(\mathcal{V}) \subseteq VR_{r_1}(\mathcal{V})$ for $r_0 \leq r_1$, where $A \subseteq B$ means A is a subset of B), and each complex $VR_r(\mathcal{V})$ in the filtration is parameterized by a radius $r > 0$.

The vertices (0-simplices) of these complexes correspond to the points in $\mathcal{V}$. A k-simplex, which can be understood as a k-dimensional tetrahedron, is included in $VR_r(\mathcal{V})$ if every pair of points in any $(k+1)$-tuple $v_0, \ldots, v_k \subseteq \mathcal{V}$ satisfies $\text{dist}_{\mathbb{R}^d}(v_i, v_j) \leq r$ for all i, j. In other words, the collection of points $v_0, \dots, v_k \subseteq \mathcal{V}$ will form a k-dimensional tetrahedron. So, a simplicial complex consists of simplices of various dimensions.

Remark. Typically, the point clouds we deal with are finite, which means there exists some $r_{\max} \in \mathbb{R}^+$ such that $VR_r(\mathcal{V}) = VR_{r_{\max}}(\mathcal{V})$ for all $r \geq r_{\max}$.

Due to the construction of the filtration, all values of r are in $[0, r_{\max}]$ are considered. Specifically, $VR_0(\mathcal{V}) = \mathcal{V}$, and for $r > r_{\max}$, we have $VR_r(\mathcal{V}) = VR_{r_{\max}}(\mathcal{V})$. Hence, no arbitrary choice of r is made at any point in the process.

Thus, we obtain the following filtration

$$\mathcal{V} = VR_0(\mathcal{V}) \subseteq \cdots \subseteq VR_r(\mathcal{V}) \subseteq \cdots \subseteq VR_{r_0'}(\mathcal{V}) = VR_{r_{\max}}(\mathcal{V}).$$

These inclusions induce maps $H_n(VR_{r_1}(\mathcal{V})) \to H_n(VR_{r_1}(\mathcal{V}))$ for $r_0 \leq r_1$, where $H_n(VR_r(\mathcal{V}))$ is the n-homology group of $VR_r(\mathcal{V})$, i. e., the group spanned by the generators of each n–dimensional hole of $VR_r(\mathcal{V})$. The *n-persistent homology* of $\mathcal{V}$ with the Vietoris–Rips filtration measures how much the n-topological features live during the filtration. Every of those features has two coordinates (b, d) that give the following information: b (*birth parameter*) is the smallest $r > 0$ such that $H_n(VR_b(\mathcal{V}))$ has that feature and d (*death parameter*) is the biggest $r > b > 0$ such that $H_n(VR_d(\mathcal{V}))$ also has it. If d does not exists, we will assign $d = \infty$. The *persistence* or *lifetime* of an n–persistent cycle is defined as the difference d–b. For graphical intuition about these constructions, we recommend the beautiful figures made by R. Ghrist in the following paper [43].

Persistence diagrams and barcodes

There exist multiple equivalent definitions of persistence diagrams. In order to mantain it as simple as possible, we will present them as a multiset of points in the upper semispace of the first quadrant of $\mathbb{R}^2$, i. e., $\mathbb{R}^2_{\geq} = \{(x, y) \in \mathbb{R}^2 : y \geq x\}$. A *multiset* V of $\mathbb{R}^d$ is a set of points in $\mathbb{R}^d$ whose elements have multiplicity, i.e., each point in V can be repeated more than once. For example, $V_0 = \{(0,0), (1,0)\}$ and $V_1 = \{(1,0), (1,0), (-1,1), (2,2), (2,2), (2,2)\}$ are multisets of $\mathbb{R}^2$. In V_0 each element has multiplicity 1 and in V_1, $(1,0)$ has multiplicity 2 and $(2,2)$ has multiplicity 3.

With this notion we are ready to present an easy definition of the *persistence diagram* as in the spirit of [44]. A *persistence diagram* $PD_i(\mathcal{F})$ is a multiset of points $(b, d) \in \overline{\mathbb{R}}^2_{\geq 0}$ where $\overline{\mathbb{R}}^2_{\geq 0} = \{(x, y) \in \overline{\mathbb{R}} \times \overline{\mathbb{R}} : 0 \leq x < y\}$ and $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$, and whose points represent the i–persistent homology of the filtration $\mathcal{F}$. Every point $(b, d) \in PD_i(\mathcal{F})$ encodes the information of an i–dimensional hole of the filtration $\mathcal{F}$ that appears at time b and dies at time d.

Equivalently, we can introduce the definition of a *barcode* of certain persistence diagram $PD_i(\mathcal{F})$ as a set of intervals $[b, d]$, where each interval corresponds to a point $(b, d) \in PD_i(\mathcal{F})$. Fig 1 illustrate that barcode allows to show the multiplicity of points whereas persistent diagrams not.

Remark. In order to improve the readability of the following definitions and sections, we will use $PD_i(X)$ as the persistence diagram of the Vietoris–Rips filtration of the point cloud X.

Persistent diagrams and barcodes can be considered as visual tools to help in the interpretation of the persistent homology; we can combine various i–persistent homologies in the same picture, as we have done in Fig 1 with the 0 and 1–persistent homology of a point cloud of 150 points extracted randomly from the unit sphere $\mathbb{S}^2 \subset \mathbb{R}^3$.

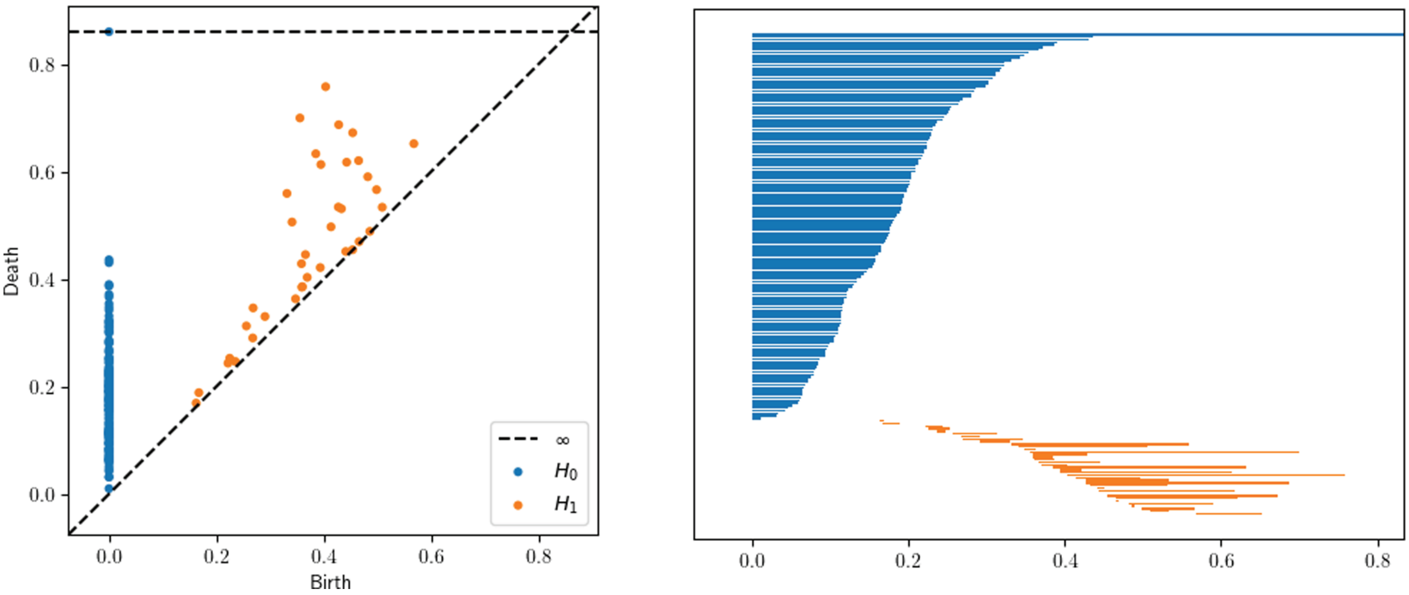

Fig 1. Persistence Diagram and Barcode of H_0 and H_1 for 150 random points on $\mathbb{S}^2 \subset \mathbb{R}^3$. Regarding the left figure, the x-axis represents the birth and the y-axis the death of each topological feature. Regarding the right figure, the x-axis represents the birth and the death of each topological feature.

https://doi.org/10.1371/journal.pone.0318108.g001

Finally, it is worth to mention that there exist different notion of distances between persistence diagrams such as the *bottleneck* or the *heat kernel* distance. All of those have their respective stability theorems which hold that small perturbations in the point clouds reflects on small perturbations on the persistence diagrams (please, consult [45]).

Persistence landscapes

As discussed, persistence diagrams and barcodes serve as powerful tools for visualizing the outcomes of persistent homology. However, one limitation of these representations is that the information they offer is not amenable to straightforward algebraic manipulations. In other words, adding two barcodes or computing averages is not a well-defined process.

In order to deal with this problem, there exist several vectorizations of the persistent diagrams (see [46]). In particular, in this paper we will use the *persistence landscapes*. Bubenik and collaborators [47–50] defined them as functions that encapsulate the information of a given persistent diagram $PD_i(X)$.

Let $PD_i(X) = \{(a_j, b_j)\}_{j \in J}$ be a persistence diagram. We define the *kth–persistence landscape* of $PD_i(X)$ via an auxiliary function: first, for $a < b$, let

$$f_{(a,b)}(t) = \max(0, \min(t - a, b - t)).$$

Then define

$$\lambda_k(t) = \max{}_k \{f_{(a_j,b_j)}(t)\}_{j \in J},$$

where $\max_k$ denotes the k-th largest element.

This process is repeated until $\lambda_l \equiv 0$ for any $l \in \mathbb{N}$. The set of functions $\lambda = \{\lambda_1, \dots, \lambda_{l-1}\}$ is denoted as the *persistence landscape* of $PD_i(X)$ (Fig 2).

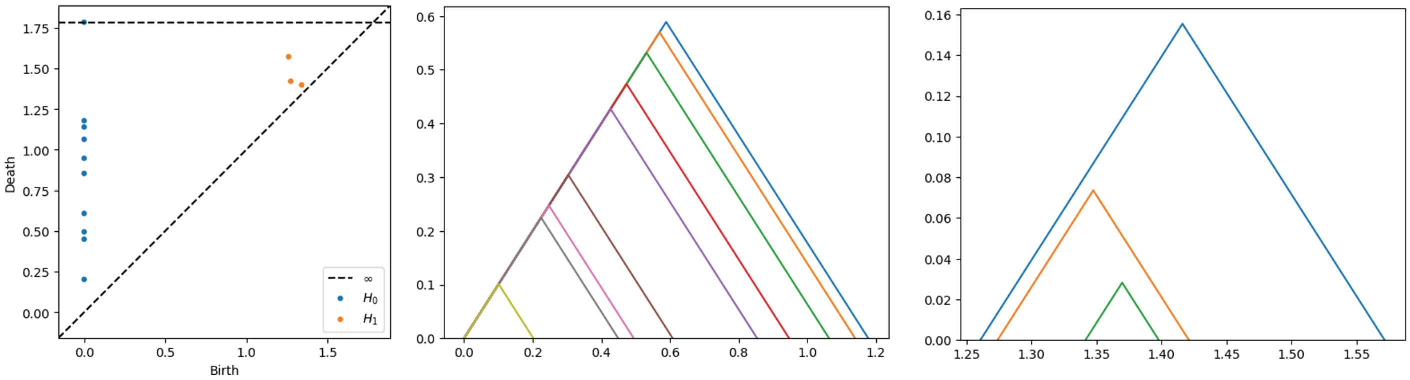

Fig 2. On the left: PD_0 and PD_1 of 10 points on $\mathbb{S}^2 \subset \mathbb{R}^3$. On the center: the landscapes of the H_0. On the right: the landscapes of the H_1.

https://doi.org/10.1371/journal.pone.0318108.g002

One of the advantages of persistence landscapes is the possibility of computing the *average persistent landscape* in the presence of a family of persistent landscapes (Fig 3). Let $\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(n)}$ be a sequence of persistence landscapes with their respective kth-persistence landscapes $\lambda^{(i)} = \{\lambda_1^{(i)}, \ldots, \lambda_l^{(i)}\}$ for $i \in \{1, \ldots, n\}$. Then we define the *average of persistence landscapes* as

$$\overline{\lambda}(k,t) = \frac{1}{n}\sum_{i=1}^{n} \lambda^{(i)}(k,t).$$

In other words, each kth-persistence landscape of the new average persistence landscape is

$$\overline{\lambda}_k(t) = \frac{1}{n}\sum_{i=1}^{n} \lambda_k^{(i)}(t).$$

Remark. From here, when we refer to distance between landscapes, we will be using the *supremum norm* or *infinity norm*,

$$\left\|\lambda(k,t) - \widetilde{\lambda}(k,t)\right\| = \sup_k \left\|\lambda_k - \widetilde{\lambda}_k\right\|_\infty = \sup_k\{\sup_t |\lambda_k(t) - \widetilde{\lambda}_k(t)|\}. \tag{1}$$

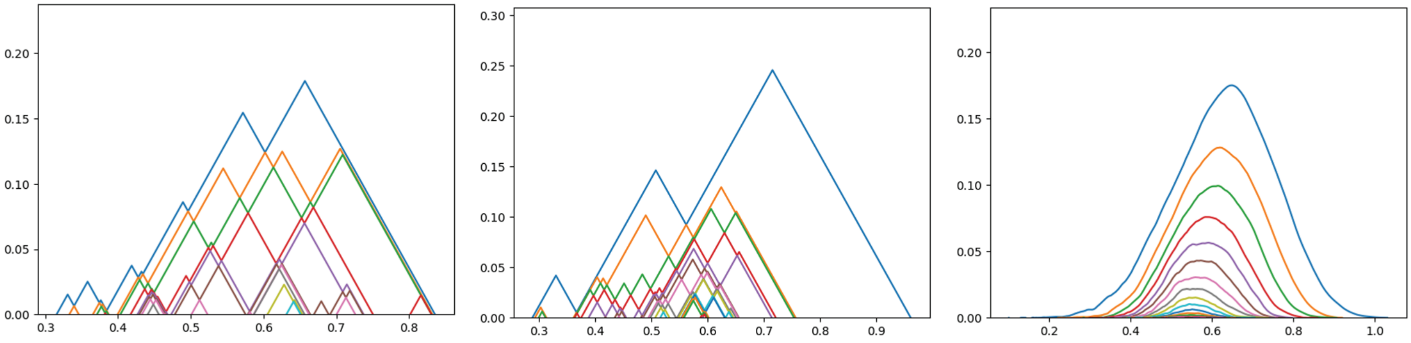

Fig 3. On the left and center: landscape number 4 and 69 respectively of the sphere experiment. On the right: the average landscape.

https://doi.org/10.1371/journal.pone.0318108.g003

Under this distance, persistence landscapes also have stable behaviour under small perturbations of the original point cloud [48,49].

Computational details

The main code implemented for this work is based on Python language. For persistence diagrams and persistence landscape computations, we have used Ripser [51,52] and Persim (https://github.com/scikit-tda/persim). For some graphical representations we display on this paper we have also used GUDHI (https://gudhi.inria.fr) [53] and the Landscape Python Package (https://gitlab.com/kfbenjamin/pysistence-landscapes/). Finally, for the centrality measures computations we have used the Python package Networkx [54].

We have used data sets extracted from Eurocontrol's NEST provided by CRIDA (the Spanish Air Traffic Management R&D&I Reference Center). Analyzed data consisted of radar and planned trajectories of all flights with origin or destination at any of the European airports during the period of interest.

The computations performed in this paper were carried out on the servers of the Department of Mathematics at Universidad Autónoma of Madrid and the Department of Quantitative Methods at CUNEF University of Madrid. These resources provided the necessary computational power to handle the complex data analyses and simulations required for our research.

Application

As we have explained at the beginning of the paper, we want to use the deviation between the planned and the real trajectory of an airplane as well as its time delay to obtain a cloud of points of a certain area, airport or, even a country, with the purpose to apply TDA on it.

Deviation distance

On the way to attach the problem, we have to define a distance to measure the deviation between a planned trajectory and the one the airplane finally did. Each trajectory is given as a set of 4–dimensional ordered points which have the following configuration:

$$(t = \text{time}, L = \text{latitude}, l = \text{longitude}, a = \text{altitude (km)}).$$

We want to define a distance between two points with that form. For this purpose and to clarify the explanation, suppose we have two points (t_0, L_0, l_0, a_o) and (t_1, L_1, l_1, a_1) where $t_0 = t_1$.

Firstly, we consider the *haversine* distance for the latitude and longitude coordinates. This calculates the great-circle distance between two points on a sphere based on their longitudes and latitudes. Although the earth is not a perfect round sphere, the haversine distance will be a very good approximation of the real latitude-longitude distance, since we are going to compute mostly distances in Europe. For the altitude coordinate, we simply compute the absolute value of the difference between a_0 and a_1.

We define a distance between (t_0, L_0, l_0, a_o) and (t_1, L_1, l_1, a_1) where $t_0 = t_1 = T$ as

$$\text{dist}_T((L_0, l_0, a_o), (L_1, l_1, a_1)) := \sqrt{\text{dist}_H((L_0, l_0), (L_1, l_1))^2 + |a_0 - a_1|^2}, \tag{2}$$

where dist_H denotes the haversine distance. This gives a natural, Euclidean distance between the geographical coordinates.

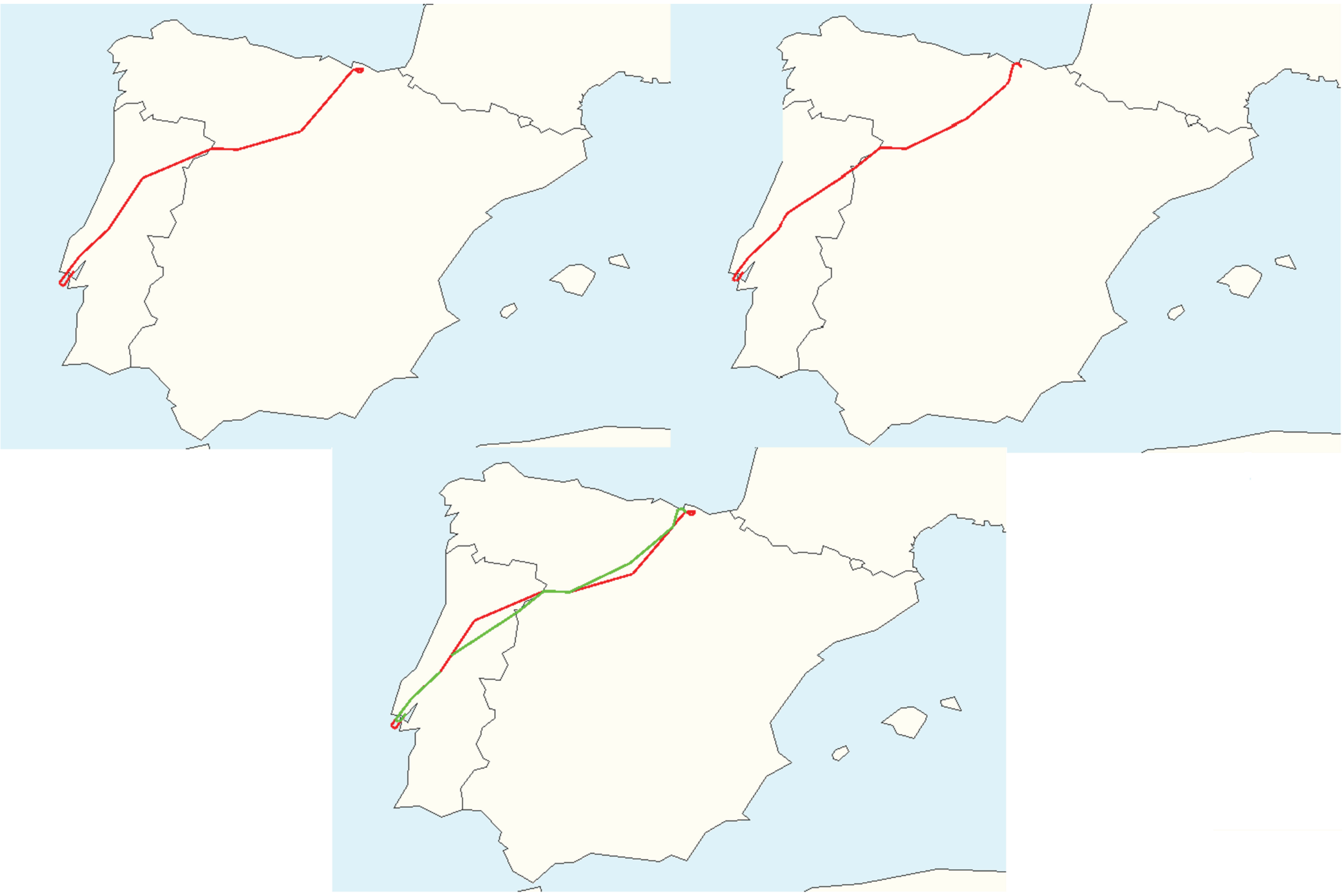

Fig 4. On the top left: The planned trajectory for a Bilbao-Lisbon flight on the 21st July 2019. On the top right: The real trajectory the plane followed. At the bottom: Both trajectories displayed con the same map.

https://doi.org/10.1371/journal.pone.0318108.g004

Let pT and rT denote the set of points of a planned and a real trajectory respectively. So, as we have mentioned above,

$$pT = \{(t_j, L_j, l_j, a_j)\}_{j \in J}$$
$$rT = \{(\widetilde{t}_i, \widetilde{L}_i, \widetilde{l}_i, \widetilde{a}_i)\}_{i \in I},$$

where I and J are sets of ordered indexes.

Remark. There are some considerations about pT and rT we want to make explicit:

1. $(L_{j_0}, l_{j_0}) = (\widetilde{L}_{i_0}, \widetilde{l}_{i_0})$ and $(L_{j_f}, l_{j_f}) = (\widetilde{L}_{i_f}, \widetilde{l}_{i_f})$, where j_0, j_f, i_0 and i_f are the initial and final indexes of each set respectively. This assertion is obvious, as the initial and final locations correspond to the initial and final destinations.
2. $a_{j_0} = a_{j_f} = a_{i_0} = a_{i_f} = 0$.
3. $\{t_j\}_{j \in J}$ and $\{\widetilde{t}_i\}_{i \in I}$ have normally different cardinality. Even $t_{j_0} \neq \widetilde{t}_{i_0}$ and $t_{j_f} \neq \widetilde{t}_{i_f}$. $\#(\{\widetilde{t}_i\}) > \#(\{t_j\})$ as we have much more points in rT than pT.

Due to the last point of the Remark and the definition (2) of our distance, for computing the distance between pT and rT we have, firstly, to redefine those sets so that they share the same temporal marks. Our choice for the new temporal sequence will be just taking the union of the temporal sequences for the planned and the real flight:

$$\{\widehat{t}_k\}_{k\in K} = \{t_j\}_{j\in J} \cup \{\widetilde{t}_i\}_{i\in I},$$

where $\widehat{t}_{k_0} = \min\{t_{j_0}, \widetilde{t}_{i_0}\}$ and $\widehat{t}_{k_f} = \max\{t_{j_f}, \widetilde{t}_{i_f}\}$, where k_0 denotes the first index (we will identify for clarity $k_0 = 0$) of K and k_f the last one.

Now, without loss of generality, suppose that $\widehat{t}_{k_l} \notin \{t_j\}_{j\in J}$ and $\widehat{t}_{k_{l-1}}, \widehat{t}_{k_{l+1}} \in \{t_j\}_{j\in J}$. We need to create a new point $p = (\widehat{t}_{k_l}, L_{k_l}, l_{k_l}, a_{k_l})$. For that purpose, we have chosen p on the segment between $(\widehat{t}_{k_{l-1}}, L_{k_{l-1}}, l_{k_{l-1}}, a_{k_{l-1}})$ and $(\widehat{t}_{k_{j+1}}, L_{k_{l+1}}, l_{k_{l+1}}, a_{k_{l+1}})$ according to the nature of our data. Furthermore, if $\widehat{t}_{k_{l-1}} \in \{t_j\}_{j\in J}$ but $\widehat{t}_{k_{l+1}} \notin \{t_j\}_{j\in J}$, we pick the first $\widehat{t}_{k_p} \in \{t_j\}_{j\in J}$ such that $k_p > k_l$ and interpolate between k_p and k_l to obtain the required information.

In addition, suppose $t_{j_0} > \widehat{t}_0$ and $t_{j_f} < \widehat{t}_{k_f}$. Let $k_{j_0}, k_{j_f} \in K$ such that $\widehat{t}_{k_{j_0}} = t_{j_0}$ and $\widehat{t}_{k_{j_f}} = t_{j_f}$; then $(\widehat{t}_{k_l}, L_{k_l}, l_{k_l}, a_{k_l}) = (\widehat{t}_{k_l}, L_{j_0}, l_{j_0}, a_{j_0})$ and $(\widehat{t}_{k_p}, L_{k_p}, l_{k_p}, a_{k_p}) = (\widehat{t}_{k_p}, L_{j_f}, l_{j_f}, a_{j_f})$, for $k_l < k_{j_0}$ and $k_p > k_{j_f}$.

So, after this construction, we have redefined our trajectories and settled pT and rT on the same temporal sequence. Now, it has sense to apply the distance defined in (2) to every pair or points of $pT \times rT$. We can finally define the distance which measures the deviation (in kilometres) between a planned and a real trajectory in the following manner

$$\mathrm{dist}_{\mathrm{Dev}}(pT, rT) := \sum_{l=0}^{k_f} \mathrm{dist}_{\widehat{t}_l}\left((L_l, l_l, a_l), (\widetilde{L}_l, \widetilde{l}_l, \widetilde{a}_l)\right). \tag{3}$$

Point cloud

As shown in Materials and Methods, in order to apply TDA techniques to a given problem, we require a point cloud to work with. To better illustrate this point, let us consider a particular airport that we wish to focus on, and investigate how deviations and delays affect its performance. Suppose that this airport has ten flights per day. As discussed in the previous subsection, we compute the distances between the planned and actual trajectories of each flight, resulting in a list of ten distances denoted as $\{d_1, \dots, d_{10}\}$. We can define our point cloud as follows:

$$\mathcal{V} = \{(\widetilde{d}_i, r_i)\}_{i\in[1,10]} \subset \mathbb{R}^2,$$

where $\widetilde{d}_i = (-1)^p d_i$ with $p = 0$ if the flight arrived late or $p = 1$ if the flight arrived on time or sooner, and r_i corresponds to the quantity $\widetilde{t}_{i_f} - t_{j_f}$, i.e., the difference between the time the flight landed and the expected one. If the flight was late, we will have $r_i > 0$, and on the other hand, $r_i \leq 0$ if the flight was on time or sooner. The points will be settled in the first and third quadrant of $\mathbb{R}^2$ as $\mathcal{V}$ only has points with both positive or bot negative coordinates (Fig 5).

In this paper, we are focusing on airports in order to obtaining the point clouds. But, as this construction is generic, we can choose more than one airport, a certain region or even a country.

Persistent homology

Once we have built our point cloud we can apply TDA on it. In this case, we will work with the 0–homology as we want to extract the information of how the connected components in the Vietoris-Rips of the filtration of our point cloud $\mathcal{V}$ birth and die. With this information

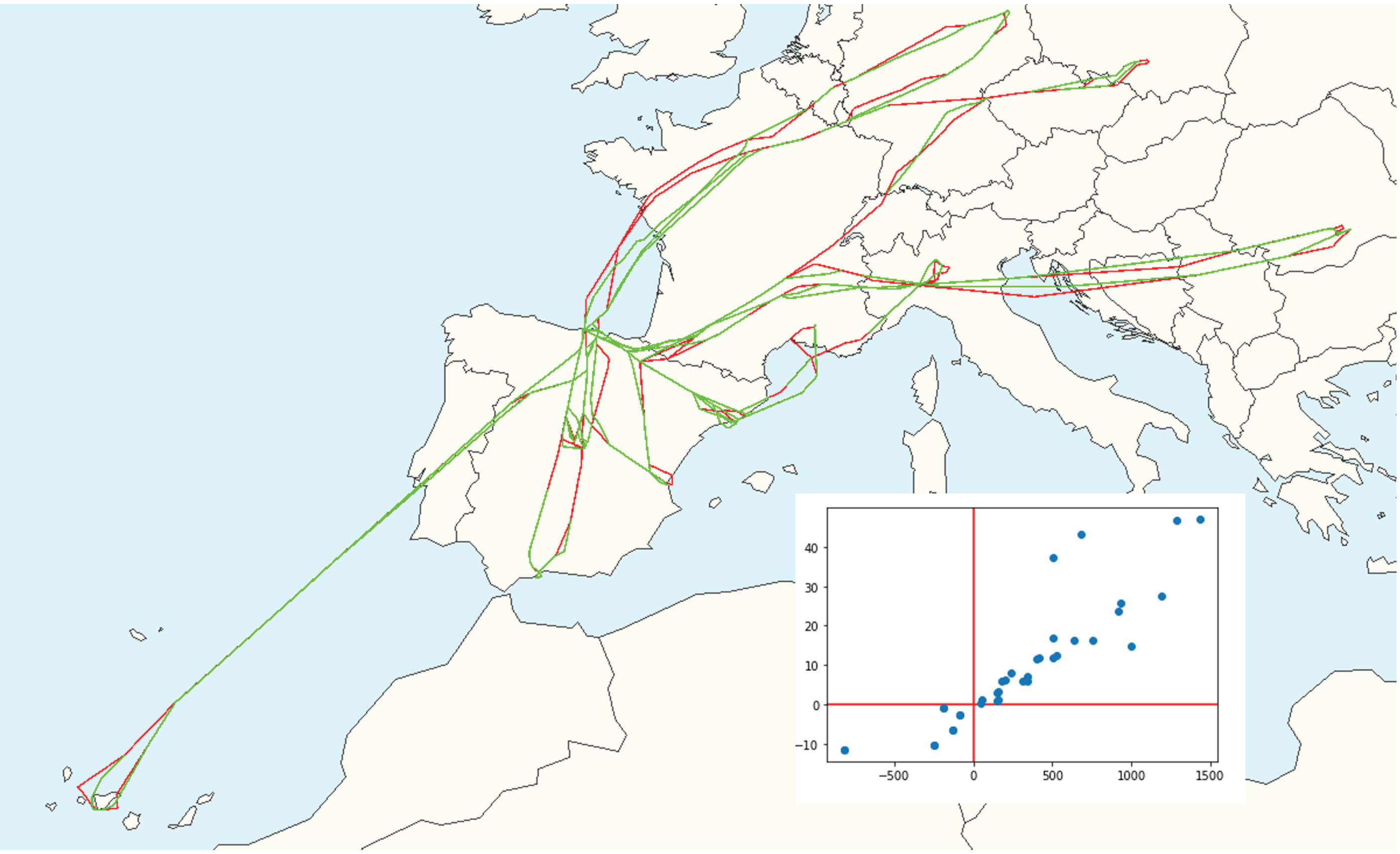

Fig 5. Point cloud of Santander's airport on the 29th June 2019.

https://doi.org/10.1371/journal.pone.0318108.g005

we recover how the points are disseminated on $\mathbb{R}^2$. Intuitively, if one flight is highly delayed, this point would be far from the origin $(0,0)$ and that phenomena would be detected by the persistent homology analysis.

For the explanation of the process we have developed, we are going to illustrate it with a particular example: Santander's airport (LEXJ) on the Summer Season (23rd March to 27th October) 2018. Although the data of 2018 can be considered old, it is not the case for the sake of this study, as current flight patterns are not significative different from those of 2018. For every day during this period of time, we are going to compute its corresponding point cloud $\mathcal{V}$ and extract from there the $PD_0(\mathcal{V})$. As the Summer Season of 2018 has 217 days, after this first step we will obtain a set $V = \{PD_0(\mathcal{V}_j)\}$ with 217 persistence diagrams; one for each day. After this, we compute the persistence landscape of each persistence diagram and compute their average persistence landscape (Fig 6).

With this average persistence landscape we obtain a particular image of every airport depending on the distribution of its daily flights via trajectory deviation and delay. One of the advantages of landscapes is that we can now compute the distance with the supremum norm we illustrated on (1) of each day with respect to the average (Fig 7). After performing these computations, the point clouds associated to the days with biggest distance from the average have clearly isolated points (Fig 8).

To summarize, we present the pseudocode of the algorithm explained in this section (Algorithm 1) to clarify the process for the rest of the paper, before presenting the results.

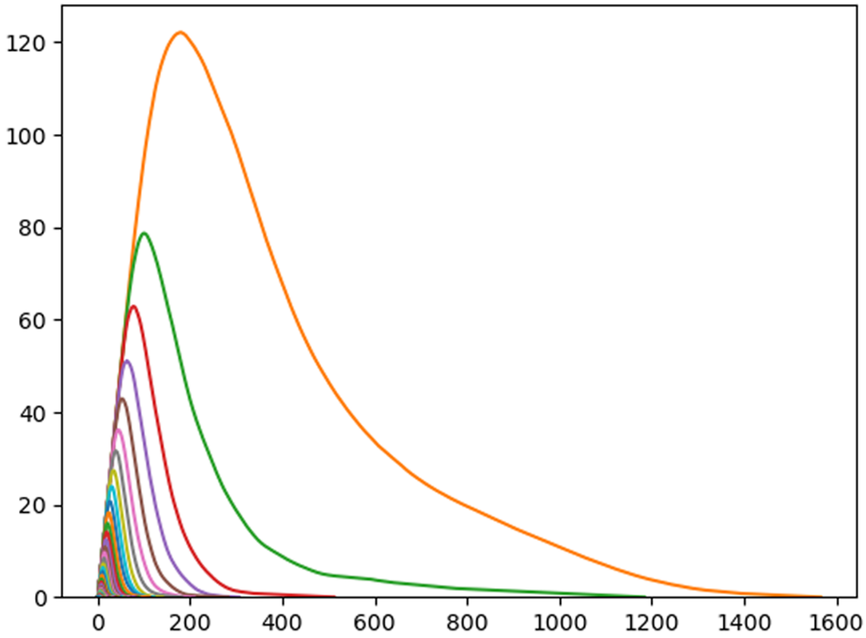

Fig 6. Average Persistence Landscape of Santander's airport in the Summer Season of 2018.

https://doi.org/10.1371/journal.pone.0318108.g006

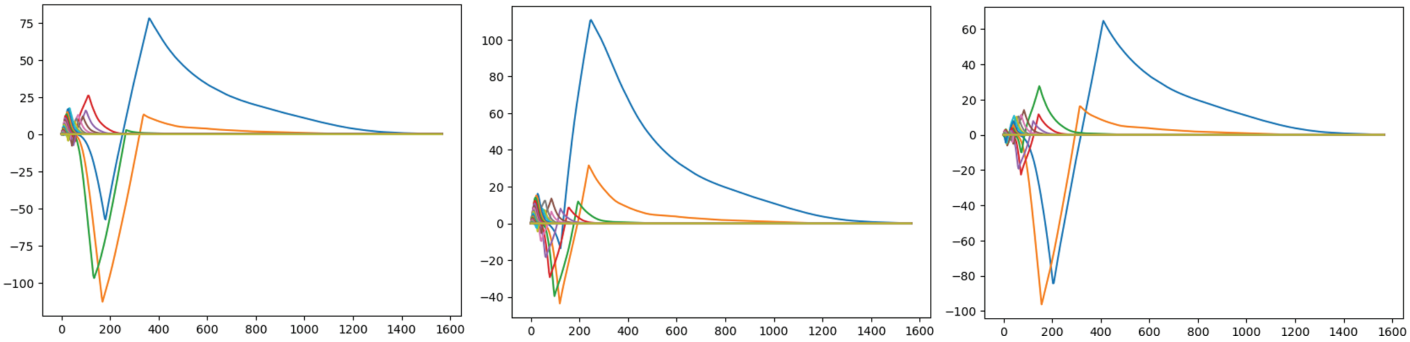

Fig 7. Three resulting differences between daily persistence landscapes and the average persistence landscape.

https://doi.org/10.1371/journal.pone.0318108.g007

Discussion and results

The purpose of this paper is to introduce the concepts of Topological Data Analysis (TDA) and explore their application in the context of Air Traffic Management (ATM) via a proof of concept presented in this section. We aim to demonstrate the efficacy of TDA through an analysis of real-world data, specifically the Spanish network of airports during the Summer Season of 2018, as classified by AENA (a Spanish public company responsible for overseeing airports of general interest in Spain).

In order to facilitate our analysis, we have leveraged AENA's airport classification system for 2018, which is divided into five distinct categories. These categories include Group 3 (airports with less than half a million passengers per year), Group 2 (airports with between half a million and two million passengers per year), Group 1 (airports with more than two million passengers per year), the Canary Group (comprised of airports located in the Canary

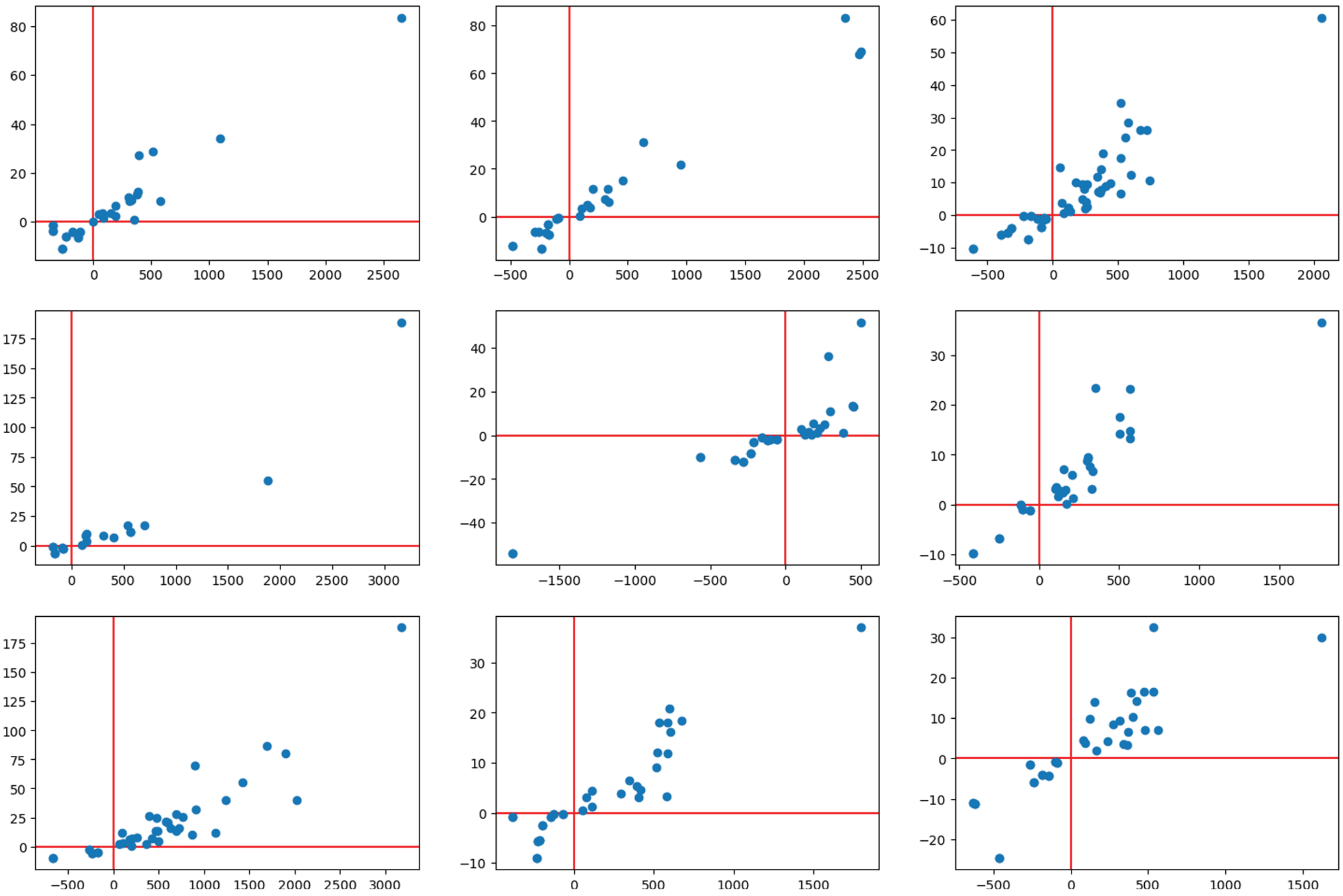

Fig 8. The nine point clouds with persistence landscapes with biggest distance with respect to the average persistence landscape of Santander's airport in the Summer Season of 2018.

https://doi.org/10.1371/journal.pone.0318108.g008

Islands), and a special group (which includes Madrid-Barajas, Barcelona-El Prat, and Palma's airport).

Remark. The upper bound imposed on Group 3 necessitated the inclusion of airports of varying types. In light of this, we have partitioned Group 3 into four distinct subgroups, namely general aviation airports, air bases open to civilian traffic, airports with low traffic, and helipads (which will be excluded from our analysis).

Similarly, the Canary Group exhibits a significant variation in the size of its airports, ranging from La Gomera with two to four flights per day to Gran Canaria with over one hundred flights per day. Despite this disparity, we have chosen to maintain all airports in the same group for our analysis.

From now on, we will denote each airport by its ICAO (International Civil Aviation Organization) code. Here is the classification given by AENA in 2018:

- *Group 3*:
  - *General aviation airports*: Madrid-C. Vientos (LECU) and Sabadell (LELL).

Algorithm 1 Algorithm for the footprint of an airport

```
Require: $N \geq 0$ and airport    ▷ N is the number of days computed in certain airport
  Create an empty list PD for the persistence diagrams
  for day in $N$ do
      Create lists with planned PT and real trajectories RT of that day
      Create an empty list PCD for the point cloud of the day
      for flight in PT and RT do
          Compute (deviation distance, delay) and add the point to PCD
      end for
      Compute the distance matrix DMD of PCD
      Compute the persistence diagram PDD of DMD and add it to PD    ▷ For this step, we use Ripser package
  end for
  Create an empty list PL for the presistence landscapes
  for pd in PD do
      Compute de persistence landscape PLD of pd and add it to PL    ▷ For this step, we use Persim package
  end for
  Finally, we computed de average persistence landscape APL (footprint) of PL    ▷ For this step, we also use Persim package
```

  - *Air bases open to civilian traffic*: Albacete (LEAB), Badajoz (LEBZ), León (LELN), Salamanca (LESA), Son Bonet (LESB) and Valladolid (LEVD).
  - *Airports with low traffic*: Melilla (GEML), Burgos-Villafría (LEBG), Córdoba (LEBA), Girona (LEGE), Huesca-Pirineos (LEHC), Logroño-Agoncillo (LERJ), Pamplona (LEPP), San Sebastián (LESO) and 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) and *Zaragoza (LEZG).*
- *Group 1*: Alicante Elche-Miguel Hernández (LEAL), Bilbao (LEBB), Ibiza (LEIB), Málaga-Costa del Sol (LEMG), Menorca (LEMH), Sevilla (LEZL) and Valencia (LEVC).
- *Canary Group*: Fuerteventura (GCFV), La Gomera (GCGM), El Hierro (GCHI), La Palma (GCLA), Gran Canaria (GCLP), César Manrique Lanzarote (GCRR), Tenerife South (GCTS) and Tenerife North (GCXO).
- *Special Group*: Adolfo Suárez Madrid-Barjas (LEMD), Josep Tarradellas Barcelona-El Prat (LEBL) and Palma Mallorca (LEPA).

Remark. We decided to highlight Zaragoza's airport (in italic) due to its uniqueness. It is the only airport in Group 2 that does not facilitate passenger aviation. We will see that this phenomena is detected by our analysis.

First, we compute the average persistence landscape of every airport corresponding to the Summer Season of 2018 obtaining a collection of average persistence landscapes $\mathcal{L} = \{\overline{\lambda}(k,t,A)\}_{A\in\mathcal{A}}$, where $\mathcal{A}$ represents the list of all of the Spanish airports. Notably, there are a wide variety of landscapes in $\mathcal{L}$. This type of analysis benefits the number of flights each airport has. So, airports with big differences in the number of passengers per year would have highly different landscapes (see Fig 9 for a comparison between Adolfo Suárez Madrid-Barajas' and Huesca-Pirineos paying attention paying attention TO the grids of the pictures).

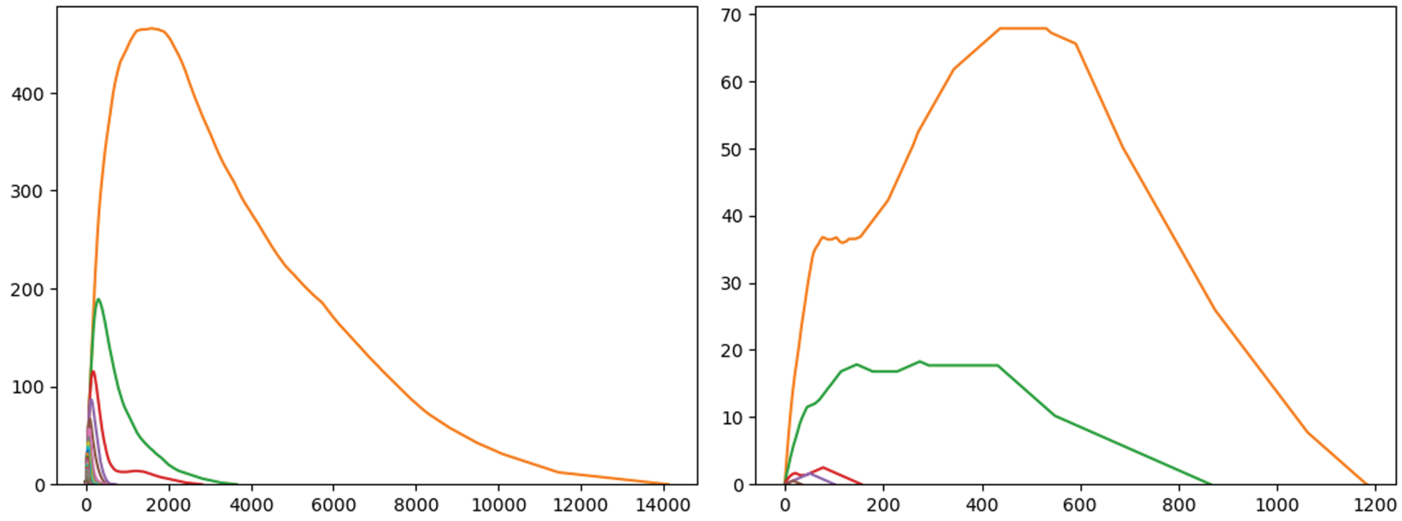

Fig 9. On the left: Average Persistence landscape of Adolfo Suárez Madrid-Barjas' airport on the Summer Season of 2018. On the right: Average Persistence landscape of Huesca-Pirineos' airport on the Summer Season of 2018. Check the difference in magnitude of the axes

https://doi.org/10.1371/journal.pone.0318108.g009

Second, we construct a distance matrix $\mathcal{D} = (d_{ij})$, where each d_{ij} is the supremum distance (explained on (1)) between the persistence average landscape of the airport *i* and the airport *j*. We have split $\mathcal{D}$ in different tables in order to clearly present the relevant information. We present the most interesting considerations regarding them. For the interested reader, the rest of the tables are displayed on S1 Text:

- On Table 1 we have highlighted four airports: GEML (red), LEGE (blue), LEPP (blue) and LEVT (blue). On Table 2 we have higlighted three airports: LEGE, LEPP and LEVT (all in blue).

  Melilla's airport (GEML) seems very different from the ones in its group and the ones in Group 2. It seems that its location affects it in comparison to the rest of the airports in the Iberica's Peninsula.

  The airports highlighted in blue seems to work better with Group 2 than with Group 3 - *Airports with low traffic* in the Summer Season 2018. A few years later, Girona's airport (LEGE) was indeed included in Group 2.
- Table 3 represents the symmetric part of $\mathcal{D}$ corresponding to Group 2.

Table 1. Distance matrix corresponding to the Group 3 - *Airports with low traffic*.

	GEML	LEBA	LEBG	LEGE	LEHC	LEPP	LERJ	LESO	LEVT
GEML	0.00	93.78	69.05	117.80	67.61	105.05	62.34	66.97	109.09
LEBA	93.78	0.00	30.87	66.68	65.14	42.23	31.76	32.52	61.30
LEBG	69.05	30.87	0.00	77.77	43.86	53.81	30.38	40.65	72.06
LEGE	117.80	66.68	77.77	0.00	86.07	35.63	67.72	57.29	23.91
LEHC	67.61	65.14	43.86	86.07	0.00	70.16	34.20	45.35	80.90
LEPP	105.05	42.23	53.81	35.63	70.16	0.00	46.40	45.63	25.94
LERJ	62.34	31.76	30.38	67.72	34.20	46.40	0.00	36.44	63.69
LESO	66.97	32.52	40.65	57.29	45.35	45.63	36.44	0.00	45.05
LEVT	109.09	61.30	72.06	23.91	80.90	25.94	63.69	45.05	0.00

https://doi.org/10.1371/journal.pone.0318108.t001

Table 2. Distance matrix corresponding to the Group 2 (rows) and Group 3 - *Airports with low traffic* (columns) where we have remarked two columns: Girona's and Vitorias's airport respectively.

	GEML	LEBA	LEBG	LEGE	LEHC	LEPP	LERJ	LESO	LEVT
LEAM	116.56	68.71	79.58	35.88	87.21	37.49	70.24	53.22	14.89
LEAS	101.64	59.92	68.69	44.71	75.33	26.39	61.26	39.34	22.59
LECO	94.07	47.88	55.34	64.26	67.45	41.22	49.55	27.13	40.81
LEGR	102.80	58.45	66.47	56.37	75.31	34.86	59.47	35.85	32.92
LEJR	120.01	63.52	74.16	24.02	87.82	31.04	64.35	57.16	13.78
LERS	116.04	69.39	80.49	36.89	86.45	36.90	70.73	55.17	15.85
LEST	113.49	60.70	68.37	27.82	81.63	29.77	60.70	49.35	11.10
LEVX	101.06	54.88	61.78	45.00	70.76	24.40	55.77	36.73	21.64
LEXJ	113.84	61.08	68.05	44.10	83.30	29.01	60.72	48.26	20.35
LEZG	242.82	202.76	207.87	157.17	181.63	185.35	188.05	199.45	174.77

https://doi.org/10.1371/journal.pone.0318108.t002

As we mentioned before, Zaragoza's airport (LEZG) is the only airport in Group 2 that does not facilitate passenger aviation. That pathological behaviour is clearly detected by the TDA analysis.

Furthermore, we have observed that airports located in close proximity tend to have similar geometric properties, as it is the case with Asturias (LEAS), Santander (LEXJ), Santiago (LEST), and Vigo (LEVX). This phenomenon suggests that the point cloud we create based on deviation of trajectories and delays, encodes information regarding geographical location of the airports and origins and destination of its flights. TDA analysis is capable of capturing that as we can check in the distances of airports with those similar characteristics.

- On Table 4 we show the differences between Group 2 and Group 1.

  We can see that Menorca's airport (LEMH) is close to the airports of Group 2. Moreover, its distance to Santander's airport (LEXJ) is one of the closest in $\mathcal{D}$.

  Nowadays, Santiago's airport (LEST) is in Group 1. We can check how in 2018 this aiport was close to some airports (Bilbao (LEBB), Menorca (LEMH), Sevila (LEZL)) of Group 1.

  Finally, we want to highlight the similarity between Sevilla's (LEZL) and Jerez's (LEJR) airport (here it is another example of two airports that are very close geographically) and how Zaragoza's aiport (LEZG) due to its uniqueness it is still very different from airports of Group 1.

  Accordingly to the information displayed above, on Table 5 we can see the symmetric part of $\mathcal{D}$ related to Group 1, and how Menorca's airport (LEMH) has bigger distances with its own group than with Group 2.

Table 3. Distance matrix corresponding to the Group 2 airports where we have remarked the Zaragoza's airport column and some close distances.

	LEAM	LEAS	LECO	LEGR	LEJR	LERS	LEST	LEVX	LEXJ	LEZG
LEAM	0.00	20.27	34.82	26.39	21.01	15.06	18.62	29.54	15.59	178.62
LEAS	20.27	0.00	22.89	15.39	23.99	19.35	24.05	11.82	13.71	195.42
LECO	34.82	22.89	0.00	14.46	42.92	36.61	37.09	21.62	24.81	212.38
LEGR	26.39	15.39	14.46	0.00	36.74	30.78	29.65	13.62	20.32	204.78
LEJR	21.01	23.99	42.92	36.74	0.00	13.80	16.74	29.64	20.73	175.38
LERS	15.06	19.35	36.61	30.78	13.80	0.00	19.84	31.14	16.62	188.94
LEST	18.62	24.05	37.09	29.65	16.74	19.84	0.00	19.25	16.69	177.77
LEVX	29.54	11.82	21.62	13.62	29.64	31.14	19.25	0.00	16.14	191.58
LEXJ	15.59	13.71	24.81	20.32	20.73	16.62	16.69	16.14	0.00	193.29
LEZG	178.62	195.42	212.38	204.78	175.38	188.94	177.77	191.58	193.29	0.00

https://doi.org/10.1371/journal.pone.0318108.t003

Table 4. Distance matrix corresponding to the Group 2 (columns) and Group 1 (rows) with different marks.

	LEAM	LEAS	LECO	LEGR	LEIR	LERS	LEST	LEVX	LEXI	LEZG
LEAL	65.50	75.44	94.91	87.04	54.71	68.10	58.41	75.54	74.80	128.03
LEBB	29.20	41.97	59.93	52.40	22.80	35.39	24.14	39.74	40.26	154.06
LEIB	31.66	48.49	65.37	57.83	29.56	42.30	31.02	44.62	46.30	147.48
LEMG	101.62	113.41	132.27	124.55	92.45	105.89	95.73	112.71	112.38	98.78
LEMH	22.24	13.10	23.64	18.86	22.87	16.99	24.12	18.83	9.37	197.85
LEVC	58.15	64.32	83.97	76.14	43.39	56.81	47.25	64.78	63.47	140.87
LEZL	28.20	33.84	52.66	44.80	15.41	27.62	22.85	33.74	31.92	161.61

https://doi.org/10.1371/journal.pone.0318108.t004

- Canary Group is interesting due to two important factors: its geographical location faraway from the rest airports of Spain and the diversity on the size of the airports in the group.

  La Gomera's and El Hierro's airport are the smallest of that group (for example, La Gomera's airport has four flights per day and El Hierro approximately ten). Its mutual distance is very small, as we can see on Table 6, in comparison with the rest of the Canary airports.

  It is very interesting to check how the Canary Group behaves with the rest of Spanish airport net. It seems that its geographical location and the length of its flights (on longer flights the deviation between trajectories increases) are an important factor to characterize this group.

- On Table 7 we displayed the symmetric part of $\mathcal{D}$ corresponding to the Special Group.

  Due to the uniqueness of this airports, the distance between them are huge and disparate. The information arose from this analysis suggests that Palma's airport (LEPA) is more similar to the airports from the existent groups than to Madrid or Barcelona. This can be checked in the tables on S1 Text, but for the clarity of the remark, we displayed on Table 8 the part of $\mathcal{D}$ corresponding to the Special Group with Group 1.

Table 5. Distance matrix corresponding to the Group 1 with Menorca's airport highlighted.

	LEAL	LEBB	LEIB	LEMG	LEMH	LEVC	LEZL
LEAL	0.00	37.29	37.18	38.29	77.34	17.64	42.97
LEBB	37.29	0.00	19.61	73.43	44.16	29.01	26.18
LEIB	37.18	19.61	0.00	73.16	51.19	29.37	21.39
LEMG	38.29	73.43	73.16	0.00	115.33	52.71	80.47
LEMH	77.34	44.16	51.19	115.33	0.00	66.20	36.46
LEVC	17.64	29.01	29.37	52.71	66.20	0.00	31.66
LEZL	42.97	26.18	21.39	80.47	36.46	31.66	0.00

https://doi.org/10.1371/journal.pone.0318108.t005

Table 6. Distance matrix corresponding to the Canary Group with different marks.

	GCFV	GCGM	GCHI	GCLA	GCLP	GCRR	GCTS	GCXO
GCFV	0.00	153.65	156.43	31.55	37.29	27.18	78.66	80.91
GCGM	153.65	0.00	11.09	163.19	144.32	141.84	183.45	102.31
GCHI	156.43	11.09	0.00	166.15	147.30	144.71	185.81	105.02
GCLA	31.55	163.19	166.15	0.00	63.54	23.46	110.20	72.97
GCLP	37.29	144.32	147.30	63.54	0.00	53.73	64.41	69.07
GCRR	27.18	141.84	144.71	23.46	53.73	0.00	88.74	62.67
GCTS	78.66	183.45	185.81	110.20	64.41	88.74	0.00	126.77
GCXO	80.91	102.31	105.02	72.97	69.07	62.67	126.77	0.00

https://doi.org/10.1371/journal.pone.0318108.t006

Table 7. Distance matrix corresponding to the Special Group.

	LEBL	LEMD	LEPA
LEBL	0.00	251.13	337.09
LEMD	251.13	0.00	463.99
LEPA	337.09	463.99	0.00

https://doi.org/10.1371/journal.pone.0318108.t007

This detailed study of the distance matrix $\mathcal{D}$ has clearly illustrated how much intricate information TDA is able to extract from an apparently very easy point cloud based on deviation of trajectories and delays. We found that, apart for the number of passengers per year in each airport on which the Spanish classification is now constructed on, geographical properties, origins and destinations and typology of the airport also influence and appears on the classification that TDA shows.

We also studied the statistical significance of the distances obtained in $\mathcal{D}$. To do this, and following the statistical tests performed by Bubenik and Dłotko in [48,50], we decided to compute a *block permutation test* [55] between each airport. Instead of permuting all persistence landscapes of airport *i* with those of airport *j*, we permuted blocks of consecutive landscapes of airport *i* with blocks of consecutive landscapes of airport *j*, considering the dependencies between consecutive days in air traffic. Due to computational limits, we calculated 400 permutations for each pair of airports.

Although the complete results of the *p*-values can be found in S2 Text, we present some interesting conclusions extracted from them. As we previously pointed out, Palma's airport (LEPA) would fit better in other groups such as Group 1 or 2 (either some big *p*-values with airports of Group 3), at least if we base our classification on trajectories and delays. Additionally, the square parts of S19 Table (S2 Text) regarding Group 2 and Group 1 respectively, reinforce our conclusions about the similarities among them in terms of air traffic, and also between groups. Finally, the variability in capacity and number of flights per day in airports of Group 3 (some of them, like LEHC, had only 31 flights in the entire 2018 Summer Season) is reflected in their *p*-values. They also follow the pattern of having higher *p*-values with airports of the same group, but some anomalies are detected when comparing them with Group 1. We assume that the difference in capacity and number of flights per day could be the reason for these discrepancies.

A final step regarding matrix $\mathcal{D}$ would be trying to replicate a point cloud in some Euclidean space with a distance matrix $\widetilde{\mathcal{D}}$ similar to $\mathcal{D}$. There are a lot of different and successful methods such as Isomap [56]. Instead of doing that, we consider that applying TDA as well to $\mathcal{D}$ (Ripser allows a distance matrix as input for computations) is suitable for this kind of

Table 8. Distance matrix corresponding to the Special Group (columns) and the Group 1 (rows) with Palma's airport (LEPA) highlighted.

	LEBL	LEMD	LEPA
LEAL	306.89	448.50	36.36
LEBB	324.83	460.19	21.29
LEIB	317.48	440.60	23.69
LEMG	281.08	438.53	71.78
LEMH	360.30	465.55	47.73
LEST	336.21	459.76	32.27
LEVC	313.41	453.90	23.77
LEZL	329.62	461.41	17.11

https://doi.org/10.1371/journal.pone.0318108.t008

experiments. As we explained, TDA tries to deal with high dimensionality and interconnectedness, so a natural question will be how the points that realize $\mathcal{D}$ as distance matrix behave. In future works, we hope to explore and present this part of the research, trying to provide some classification to the whole European airport net. Meanwhile, on Fig 10 the reader can see the 0- and 1-persistence diagram produced by the Spanish airport network during the Summer Season of 2018.

Based on the conclusions extracted from $\mathcal{D}$ and the p-value computations, we see that improving the data used to classify airports (the Spanish Airport Network classification currently relies on the number of passengers per year) would provide a more complex and detailed picture of the network. This, in turn, would help in taking similar actions at airports of the same class and reducing redundant operations. We firmly believe that adding additional information, such as the order of flights per day or even their capacities, would enrich the experiment we have presented in this paper. Currently, we are working in that direction.

Centrality measures

Although we have identified some pros and cons of using TDA in this particular experiment, a reasonable next step would be to compare our results with those obtained using different methods. For that purpose, in line with the recent literature on centrality measures presented in the Introduction [25–28], we have computed four centrality metrics in order to also obtain information of the Spanish Airport Network:

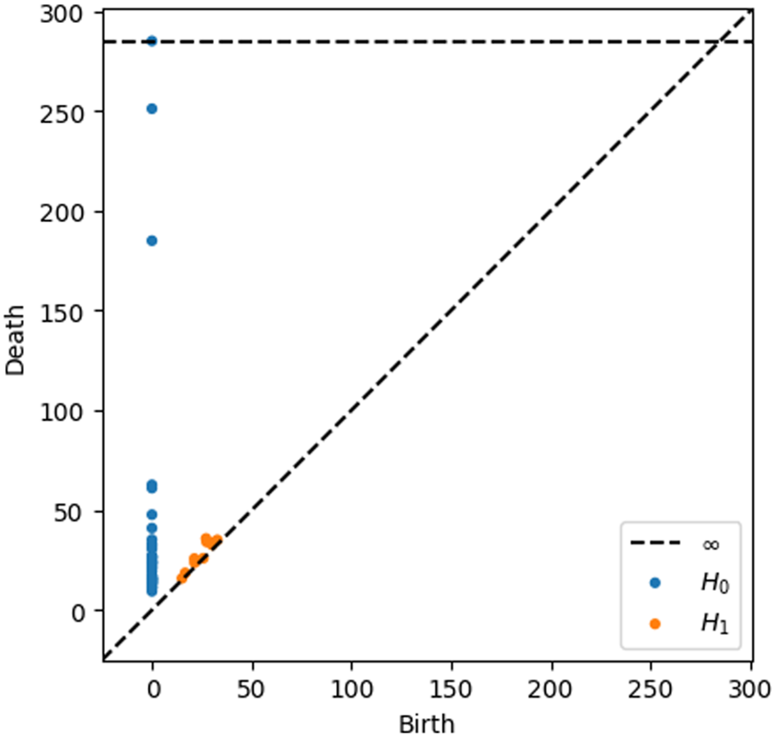

Fig 10. 0- and 1-persistence diagrams of the net of Spanish airports in the Summer Season of 2018.

https://doi.org/10.1371/journal.pone.0318108.g010

1. *Betweenness centrality:* It measures how often a vertex appears on the shortest paths between other vertices. Vertices with high betweenness can play a key role in a network by influencing the flow of information between other vertices.
2. *Degree centrality:* It quantifies the ratio between the number of edges links of certain node by the total number of connections of the network.
3. *Closeness centrality:* It computes the average shortest path from a given vertex to all other vertices in the network.
4. *Local Reaching centrality:* It indicates how effectively a node can reach others within its local neighborhood. It evaluates the proportion of nodes that can be accessed from the given node within a certain radius, and ultimately selects the maximum value.

For readers interested in deeper definitions and implications of these metrics, we provide the following references: [57,58].

We have calculated these metrics for the whole European Airport Network for every day in the Summer Season 2018 (217 days), building a directed and weighted graph for each day: nodes are the airports, direct edges point out that a flight goes from airport v_1 to airport v_2, and the weight of each edge is the number of flights in that day between airport v_1 and v_2. Once we built the 217 graphs, we calculated the centrality metrics of each day and, finally the average of each metric, extracting only the Spanish airports results and sorted them accordingly. The whole table with these results can be found in S3 Text. Additionally, Table 9 provides a summary of the number of edges and nodes in these networks.

The first clear advantage of using these centrality methods over TDA is the computation time, as TDA is much slower. However, these two methods approach the problem in completely different ways. Centrality measures, based on graph theory, provide straightforward descriptive insights without involving complex analysis or drawing inferences, whereas TDA tries to focus on interconectedness and searching patterns under different perspectives we previously presented. Although graphs can be enriched with a lot of information and can be weighted in many ways, in this particular study, the information they encapsulate is simpler than the data uniquely created for the footprints of the airports. For example, the geographical relation discovered for LEAS, LEXJ, LEST and LEVX airports are lost with centrality measures. Moreover, as we previously mentioned, the point cloud used for the persistence diagrams could be made more complex by including additional information, such as economic data of the flights, their order during the day, or even their capacity. In Limitations' Section, we explore in more detail these possible upgrades.

S21 Table (S3 Text) presents valuable insights. First, Local reaching and Betweenness columns are practically the same. Although we compute four different aspects of our network, finally we are only basing our new classification over three of them. Second, in terms of the sorting, we also see the Special Group (LEMD, LEBL and LEPA) in the three first positions of the table but, as well as our TDA experiment has detected, Palma's airport has close Degree and Closeness metrics with airports of other groups. Furthermore, for the other atypical group, the Canary's one, we can see in the table that its variability is almost perfectly captured (except for La Palma's airport (GCLA)). Big airports such as Gran Canaria (GCLP) and

Table 9. Summary of the network used to compute centrality measures: the number of vertices remains constant each day, while the second and third columns provide the average and standard deviation of the number of edges in the directed graph of the Spanish airport network.

Num. vertices	Average Edges	Standard Deviation Edges
1247	2160.80	185.91

https://doi.org/10.1371/journal.pone.0318108.t009

Tenerife South (GCTS) are in high positions, the big/medium ones are spared around the whole ranking and the smallest ones (La Gomera's (GCGM) and GCHI (GCHI)), with a few flights per week, are the last two of the table.

Additionally, we observe that some high-traffic airports, such as Bilbao (LEBB) and La Palma (GCLA), are far from similar airports. Also, Zaragoza's airport (LEZG) which not operates flights with passengers, is in position thirteen of the ranking, but its uniqueness is not detected as it was in the footprint study.

Regarding small airports (Group 3), Girona's airport (LEGE) is placed in top positions reflecting its high number of connections and the improvement in number of flights that suffers during Summer seasons. TDA did also reflect this pattern. The remaining airports are in the low part of the table, also agreeing with the actual network classification.

Overall, the classification provided by the centrality measures is largely similar to the one from AENA, with only a few cases suggesting new areas for study that could enhance the classification with more data-driven insights.

We believe that centrality measures are also very powerful and promising for ATM. In our experiment, we were not able to add as much information to the graph as the point clouds used for TDA, but the computation time is considerably shorter than that required for computing the distance matrix $\mathcal{D}$. Both methods have similarities with the current Spanish Airport Network classification and present nuances that enrich it, being the one provided by centrality measures closer to the actual. We consider that a combination of TDA techniques and centrality measures could be more powerful than using them alone. We are currently working in this direction for future studies.

To sum up, in this section we presented a proof of concept demonstrating how TDA can enhance existing classifications. As we have shown, experiments like the one we conducted can lead to exploring new ways of classifying airports in a more data-driven manner, optimizing their potential and resources. Moreover, TDA can open new perspectives for addressing ATM problems and challenges due to its numerous advantages and its capability to successfully combine with existing techniques.

Limitations

This study applied Topological Data Analysis (TDA) techniques to analyze the operational behavior of various airports in Spain using data from a single season in 2018. While the results identified clusters of airports with similar behaviors, several limitations affect the generalizability of the findings and suggest potential improvements for future research.

One of the main limitations of this study is that it relies solely on data from a single operational season. This makes it difficult to extrapolate the results to other time periods, as the observed patterns may be influenced by seasonality and interannual variations in airport operations. To draw more representative conclusions, it would be advisable to extend the analysis to include data from multiple years and seasons.

Another aspect that has not been considered is the potential presence of exceptional events on certain days, such as technical, meteorological, or social incidents, which may have altered the usual behavior of the airports. These events introduce variability into the data, which could skew the results if not appropriately filtered. Therefore, filtering out atypical days would be beneficial in reducing such bias.

Another factor missing from the analysis is contextual information on operational conditions, such as weather, capacity restrictions, and flight demand. Including this data would allow for a more granular analysis and better distinctions between days with similar conditions, enhancing the precision of the findings.

Furthermore, the study did not explore how airports adapt to varying operational conditions. Classifying airports based on their responses to high-demand days or adverse weather could reveal deeper insights into their adaptability and operational resilience.

The analysis would benefit from expanding the dataset to cover multiple years and refining the filtering process to group days with similar characteristics. This would reduce unexplained heterogeneity and improve the quality of the inferences drawn.

Finally, the computational speed of TDA tools is slower compared to other techniques, such as centrality measures. Combining different methods and perspectives, and improving the parallelization of the code would be beneficial for the manageability and development of the process. Although time performance is always important, since we are focused on classification, it is acceptable to sacrifice some computational time in order to capture more nuances.

In conclusion, extending the analysis to include multiple seasons and years, filtering out atypical events, and incorporating supplementary data on airport conditions, as well as improving the speed of computation of the proposal, would enhance the robustness and depth of the study. Classifying airports based on operational scenarios would also provide a more detailed understanding of their behavior, improving the overall accuracy and reliability of the findings.

Conclusions

Airports and Air Traffic Management Systems are complex sociotechnical structures whose interconnectedness, interdependencies, and complexity generate a huge amount of flight data. Despite their large potential, flight data are complex, high-dimensional, and sparse datasets, and they are affected by inconsistencies, errors, high levels of variability, multilevel interactions, dynamic changes and high dimensionality, making them very difficult to analyse and exploit.

While existing research works have made significant contributions to the field of analyzing high-dimensional flight trajectory data, there are still limitations that need to be addressed in future research to improve the accuracy, reliability and scalability of these methods.

To overcome these difficulties, this study addresses the use of Topological Data Analysis (TDA) to analyse flight trajectory data. In particular, as a proof of concept, we try to determine relationships between different variables involved in the spatial and temporal flight trajectory and delays to identify common patterns and anomalies in airport operation, extracting a *footprint* of each airport as a way to identify it. In future studies, it could help to recognize the underlying causes of delays and develop more effective strategies for reducing them and produce, combined to other methods and techniques, a more exhaustive classification of groups of airports.

Real-world data from the Spanish network of airports in the Summer Season of 2018 were used for the assessment. For each set of airports, the average persistence landscape has been computed for each airport in that group, based on which we construct a distance matrix $\mathcal{D}$, where each element d_{ij} is the supremum distance between the persistence average landscapes of airport i and airport j. After this, we performed a block permutation test in order to check the statistical significance of our experiment. Finally, we studied either $\mathcal{D}$ and the p-values in order to extract some conclusions and compare them with the current Spanish Airport Network classification and another non-TDA method based on centrality measures.

The accomplished analysis points out several conclusions regarding the grouping of airports attending to operational criteria, and regarding the identification of patterns and anomalies:

- The results of the analysis show how different airport groups follow a sort of cluster, even though there are few in number in each group. The analysis show that the imposed upper limit on Group 3 compelled the incorporation of airports of diverse types. Consequently, Group 3 has been subdivided into four specific subgroups, namely general aviation airports, air bases accessible to civilian traffic, airports with low traffic, and helipads. Likewise, the Canary Group presents a notable variation in the sizes of its airports, ranging from La Gomera with two to four flights per day to Gran Canaria with over one hundred flights per day.
- It also allows the identification of airports that clearly differ from their preassigned group (such as Zaragoza's airport, which differs from the rest of the airports of Group 2). It highlight the Zaragoza's airport uniqueness, as the only airport in Group 2 that does not facilitate passenger aviation.
- It also helps detecting when an airport is isolated and far from any other airport in the network, such as Adolfo Suarez Madrid-Barajas and Josep Tarradellas Barcelona-El Prat's airports are from any other airport in Spain. The Average Persistence landscape of these airports also exhibit a big difference in magnitude of the axes.
- Moreover, the diversity in a category of airports is also reflected as in the case of the Canary Airports Group. It is very interesting to check using distance matrix information how the Canary Group behaves with the rest of Spanish airport net. It seems that its geographical location and the length of its flights (on longer flights the deviation between trajectories increases) are an important factor to characterize this group.
- It illustrated also that point cloud create based on deviation of trajectories and delays, encodes information regarding geographical location of the airports and origins and destination of its flights. This explain that airports located in close proximity tend to have similar geometric properties, as it is the case with Asturias (LEAS), Santander (LEXJ), Santiago (LEST), and Vigo (LEVX). TDA analysis is capable of capturing that as it can check in the distances of airports with those similar characteristics.
- Lastly, the p-values shown in S2 Text reinforce some of our conclusions and highlight weaknesses in our analysis, indicating future directions for improving our classification.

This study has become a proof of concept of how TDA can become a powerful analytical technique to help overcome some of the limitations of existing research in the field of analyzing high-dimensional flight trajectory data for the identification of common traffic patterns in airports. In particular, during the study, the following advantages of TDA were proposed:

- It can assist in identifying underlying structures and patterns within the data in a manner that is more robust to noise and outliers.
- It can also help address the issue of limited data availability by allowing for the integration of different data sources and the extraction of meaningful insights from incomplete or noisy data. This can be a crucial tool for solving the problem of scalability by allowing the analysis of large and complex datasets using parallel computing techniques.
- It can attempt to address the lack of standardization by offering a flexible and adaptable framework that is applicable to a diverse array of data formats and types.

To the best of our knowledge, apart from some purely introductory contributions [20,24], this paper presents the first effective application of TDA to the aerospace field. In particular, no rigorous work had been performed by applying this method to a large amount of aircraft trajectory data in an attempt to anticipate and identify anomalies in aircraft space/time trajectories, infer patterns of behavior at different airports, and classify and characterize airports depending on the distribution of their daily flights via trajectory deviation and delay. This pioneering effort not only expands the scope of TDA but also lays the groundwork for future research in leveraging this method to enhance our understanding of aircraft dynamics, ultimately contributing to improved aviation safety and efficiency.

Supporting information

S1 Text. Tables.
(PDF)

S2 Text: Statistical significance: p-values.
(PDF)

S3 Text: Centrality measures.
(PDF)

Acknowledgments

The authors want to thank Telmo Pérez Izquierdo and Juan Antonio Pichardo for their comments and ideas to the manuscript. They have undoubtedly improved the quality of the material we present.

Author contributions

Conceptualization: Manuel Cuerno, Luis Guijarro, Rosa María Arnaldo Valdés, Fernando Gómez Comendador.

Data curation: Manuel Cuerno, Luis Guijarro, Rosa María Arnaldo Valdés, Fernando Gómez Comendador.

Formal analysis: Manuel Cuerno, Luis Guijarro, Rosa María Arnaldo Valdés, Fernando Gómez Comendador.

Funding acquisition: Manuel Cuerno, Luis Guijarro, Rosa María Arnaldo Valdés, Fernando Gómez Comendador.

Investigation: Manuel Cuerno, Luis Guijarro, Rosa María Arnaldo Valdés, Fernando Gómez Comendador.

Methodology: Manuel Cuerno, Luis Guijarro, Rosa María Arnaldo Valdés, Fernando Gómez Comendador.

Project administration: Manuel Cuerno, Luis Guijarro, Rosa María Arnaldo Valdés, Fernando Gómez Comendador.

Resources: Manuel Cuerno, Luis Guijarro, Rosa María Arnaldo Valdés, Fernando Gómez Comendador.

Software: Manuel Cuerno, Luis Guijarro, Rosa María Arnaldo Valdés, Fernando Gómez Comendador.

Supervision: Manuel Cuerno, Luis Guijarro, Rosa María Arnaldo Valdés, Fernando Gómez Comendador.

Validation: Manuel Cuerno, Luis Guijarro, Rosa María Arnaldo Valdés, Fernando Gómez Comendador.

Visualization: Manuel Cuerno, Luis Guijarro, Rosa María Arnaldo Valdés, Fernando Gómez Comendador.

Writing – original draft: Manuel Cuerno, Luis Guijarro, Rosa María Arnaldo Valdés, Fernando Gómez Comendador.

Writing – review & editing: Manuel Cuerno, Luis Guijarro, Rosa María Arnaldo Valdés, Fernando Gómez Comendador.

References

1. Group ATA. Aviation: Benefits Beyond Borders; 2018. Available from: https://aviationbenefits.org/media/166711/abbb18_full-report_web.pdf
2. Li MZ, Ryerson MS. Reviewing the DATAS of aviation research data: diversity, availability, tractability, applicability, and sources. J Air Transp Manag. 2019;75:111–30. https://doi.org/10.1016/j.jairtraman.2018.12.004
3. Ahmed N, Matsushima K, Nemoto K, Kondo F. Identification of inheritance and genetic loci responsible for wrinkled fruit surface phenotype in chili pepper (Capsicum annuum) by quantitative trait locus analysis. Mol Breed. 2024;45(1):5. https://doi.org/10.1007/s11032-024-01528-y PMID: 39734933
4. Bian J, Tian D, Tang Y, Tao D. A survey on trajectory clustering analysis; 2018.
5. Bolić T, Castelli L, De Lorenzo A, Vascotto F. Trajectory clustering for air traffic categorisation. Aerospace. 2022;9(5):227. https://doi.org/10.3390/aerospace9050227
6. Jasra S, Gauci J, Muscat A, Valentino G, Zammit-Mangion D, Camilleri R. Literature review of machine learning techniques to analyse flight data. 2018.
7. Agrawal E, Navdeti C. Improving the prediction of traffic flow in the airport system using machine learning. In: Chanda CK, Szymanski JR, Sikander A, Mondal PK, Acharjee D, editors. Advanced energy and control systems. Singapore: Springer; 2022. p. 275–86.
8. Altinok A, Kiran R, Bue B, Bilimoria KD. Modeling key predictors of airport runway configurations using learning algorithms. 2018. Available from: https://arc.aiaa.org/doi/abs/10.2514/6.2018-3673
9. Andrienko G, Andrienko N, Fuchs G, Garcia JMC. Clustering trajectories by relevant parts for air traffic analysis. IEEE Trans Vis Comput Graph. 2018;24(1):34–44. https://doi.org/10.1109/TVCG.2017.2744322 PMID: 28866540
10. Ma J, Zhou J, Liang M, Delahaye D. Data-driven trajectory-based analysis and optimization of airport surface movement. Transp Res Part C: Emerg Technol. 2022;145:103902. https://doi.org/10.1016/j.trc.2022.103902
11. Olive X, Morio J. Trajectory clustering of air traffic flows around airports. Aerosp Sci Technol. 2019;84:776–81. https://doi.org/10.1016/j.ast.2018.11.031
12. Rehm F, Klawonn F, Russ G, Kruse R. Modern data visualization for air traffic management. In: 2007 Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS 2007). 2007. p. 19–24. https://doi.org/10.1109/nafips.2007.383804
13. Zeng W, Xu Z, Cai Z, Chu X, Lu X. Aircraft trajectory clustering in terminal airspace based on deep autoencoder and gaussian mixture model. Aerospace. 2021;8(9):266. https://doi.org/10.3390/aerospace8090266
14. Chu X, Tan X, Zeng W. A clustering ensemble method of aircraft trajectory based on the similarity matrix. Aerospace. 2022;9(5):269. https://doi.org/10.3390/aerospace9050269
15. Zeng C, Wang R, Zuo Q. Analysis of abnormal flight and controllers data based on DBSCAN method. Secur Commun Netw. 2022;2022:1–8. https://doi.org/10.1155/2022/7474270
16. Wang Z, Liang M, Delahaye D. Short-term 4D trajectory prediction using machine learning methods; 2017.
17. Zhou A, Maletić S, Zhao Y. Robustness and percolation of holes in complex networks. Phys A: Statist Mech Appl. 2018;502:459–68. https://doi.org/10.1016/j.physa.2018.02.149
18. Chazal F, Michel B. An introduction to topological data analysis: fundamental and practical aspects for data scientists. Front Artif Intell. 2021;4:667963. https://doi.org/10.3389/frai.2021.667963 PMID: 34661095
19. Ferrà A, Cecchini G, Nobbe Fisas F-P, Casacuberta C, Cos I. A topological classifier to characterize brain states: when shape matters more than variance. PLoS One. 2023;18(10):e0292049. https://doi.org/10.1371/journal.pone.0292049 PMID: 37782651

20. Li MZ. Detection of individual anomalous arrival trajectories within the terminal airspace using persistent homology. In: 8th International Conference for Research in Air Transportation; 2018. p. 4.

21. Cho J, Yoon Y. How to assess the capacity of urban airspace: a topological approach using keep-in and keep-out geofence. Transp Res Part C: Emerg Technol. 2018;92:137–49. https://doi.org/10.1016/j.trc.2018.05.001

22. Hongyong W, Ruiying W, Yifei Z. Analysis of topological characteristics in air traffic situation networks. Proc Inst Mech Eng Part G: J Aerosp Eng. 2015;229(13):2497–505. https://doi.org/10.1177/0954410015578482

23. Duponchel L. When remote sensing meets topological data analysis. J Spect Imaging. 2018;7.

24. Li MZ, Ryerson MS, Balakrishnan H. Topological data analysis for aviation applications. Transp Res Part E: Logist Transport Rev. 2019;128:149–74. https://doi.org/10.1016/j.tre.2019.05.017

25. Huynh HN, Ng KL, Toh R, Feng L. Understanding the impact of network structure on air travel pattern at different scales. PLoS One. 2024;19(3):e0299897. https://doi.org/10.1371/journal.pone.0299897 PMID: 38457398

26. Nikolaou P, Dimitriou L. Identification of critical airports for controlling global infectious disease outbreaks: stress-tests focusing in Europe. J Air Transp Manag. 2020;85:101819. https://doi.org/10.1016/j.jairtraman.2020.101819 PMID: 32501381

27. Song MG, Yeo GT. Analysis of the air transport network characteristics of major airports. Asian J Ship Logist. 2017;33(3):117–25. https://doi.org/10.1016/j.ajsl.2017.09.002

28. Sun L-L, Hu Y-P, Zhu C-P. Centrality anomalies for the domestic air transportation networks in the USA: an empirical benchmark. Eur Phys J Plus. 2023;138(5):383. https://doi.org/10.1140/epjp/s13360-023-04003-3 PMID: 37192841

29. Frewin A, Li MZ, Taylor CP, Weitz LA. Aggregate network model with resilience considerations for air traffic flow management. In: AIAA SCITECH 2023 Forum; 2023. p. 0733.

30. Holdren SS, Li MZ, Hoffman J. Adaptable graph networks for air traffic analysis applications. In: 2023 Integrated Communication, Navigation and Surveillance Conference (ICNS). IEEE; 2023. vol. 80. p. 1–8. https://doi.org/10.1109/icns58246.2023.10124325

31. Li M, Gopalakrishnan K, Zhu X, Nandi A, Balakrishnan H, Marla L. Identification and prediction of disruptions in airline networks. Available at SSRN 3816677. 2021.

32. Li M, Gopalakrishnan K, Balakrishnan H, Shin S, Jalan D, Nandi A. Dynamics of disruption and recovery in air transportation networks. CEAS Aeronaut J. 2021:1–11.

33. Li MZ, Gopalakrishnan K, Pantoja K, Balakrishnan H. Graph signal processing techniques for analyzing aviation disruptions. Transport Sci. 2021;55(3):553–73. https://doi.org/10.1287/trsc.2020.1026

34. Beutenmüller F, Dierolf B, Keckeisen M, Pausinger F, Vaudrevange P. Topological data analysis in automotive industry. In: International Stuttgart Symposium. Springer; 2023. p. 44–56.

35. Macarasig RJ, Soria G, Aycardo JE, Garcia A, Nable J, Go CK. Topological data analysis of collective behavior in public transportation. In: AIP Conference Proceedings. vol. 2895. AIP Publishing; 2024.

36. Zhang SY, Stumpf MP, Needham T, Barbensi A. Topological optimal transport for geometric cycle matching. arXiv preprint arXiv:240319097. 2024.

37. Carlsson G. Topology and data. Bull Amer Math Soc. 2009;46(2):255–308. https://doi.org/10.1090/s0273-0979-09-01249-x

38. Otter N, Porter MA, Tillmann U, Grindrod P, Harrington HA. A roadmap for the computation of persistent homology. EPJ Data Sci. 2017;6(1):17. https://doi.org/10.1140/epjds/s13688-017-0109-5 PMID: 32025466

39. Botnan MB, Lesnick M. An introduction to multiparameter persistence; 2022. Available from: https://arxiv.org/abs/2203.14289

40. Harrington HA, Otter N, Schenck H, Tillmann U. Stratifying Multiparameter Persistent Homology. SIAM J Appl Algebra Geometry. 2019;3(3):439–71. https://doi.org/10.1137/18m1224350

41. Hatcher A. Algebraic topology. Cambridge: Cambridge University Press; 2002.

42. Weibel CA. History of homological algebra. Hist Topol. 1999:797–836. https://doi.org/10.1016/b978-044482375-5/50029-8

43. Ghrist R. Barcodes: the persistent topology of data. Bull Amer Math Soc. 2007;45(01):61–76. https://doi.org/10.1090/s0273-0979-07-01191-3

44. Che M, Galaz-Garcıa F, Guijarro L, Solis IM. Metric geometry of spaces of persistence diagrams; 2021. Available from: https://arxiv.org/abs/2109.14697

45. Chazal F, Cohen-Steiner D, Guibas LJ, Mémoli F, Oudot SY. Gromov-Hausdorff stable signatures for shapes using persistence. Comput Graph Forum. 2009;28(5):1393–403. https://doi.org/10.1111/j.1467-8659.2009.01516.x

46. Ali D, Asaad A, Jimenez M-J, Nanda V, Paluzo-Hidalgo E, Soriano-Trigueros M. A survey of vectorization methods in topological data analysis. IEEE Trans Pattern Anal Mach Intell. 2023;45(12):14069–80. https://doi.org/10.1109/TPAMI.2023.3308391 PMID: 37647183

47. Betthauser L, Bubenik P, Edwards PB. Graded persistence diagrams and persistence landscapes. Discrete Comput Geom. 2021;67(1):203–30. https://doi.org/10.1007/s00454-021-00316-1

48. Bubenik P. Statistical topological data analysis using persistence landscapes. J Mach Learn Res. 2015;16:77–102.

49. Bubenik P. The persistence landscape and some of its properties. In: Topological data analysis—the Abel Symposium 2018. vol. 15. Cham: Springer; 2020. p. 97–117. Available from: https://doi.org/10.1007/978-3-030-43408-3_4.

50. Bubenik P, Dłotko P. A persistence landscapes toolbox for topological statistics. J Symbol Comput. 2017;78:91–114. https://doi.org/10.1016/j.jsc.2016.03.009

51. Bauer U. Ripser: efficient computation of Vietoris–Rips persistence barcodes. J Appl and Comput Topol. 2021;5(3):391–423. https://doi.org/10.1007/s41468-021-00071-5

52. Tralie C, Saul N, Bar-On R. Ripser.py: a lean persistent homology library for python. JOSS. 2018;3(29):925. https://doi.org/10.21105/joss.00925

53. Maria C, Boissonnat JD, Glisse M, Yvinec M. The Gudhi library: simplicial complexes and persistent homology. In: Hong H, Yap C, editors. Mathematical software – ICMS 2014. Berlin Heidelberg: Springer; 2014. p. 167–174.

54. Hagberg A, Swart PJ, Schult DA. Exploring network structure, dynamics, and function using NetworkX. Los Alamos, NM (United States): Los Alamos National Laboratory (LANL); 2008.

55. Lahiri SN. Resampling methods for dependent data. Springer; 2013.

56. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science. 2000;290(5500):2319–23. https://doi.org/10.1126/science.290.5500.2319 PMID: 11125149

57. Mones E, Vicsek L, Vicsek T. Hierarchy measure for complex networks. PLoS One. 2012;7(3):e33799. https://doi.org/10.1371/journal.pone.0033799 PMID: 22470477

58. Wasserman S, Faust K. Social network analysis: methods and applications. Cambridge University Press; 1994.