UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Análise topológica do tráfego aéreo com dados do ANAC

Marzia Petrucci

Trabalho de Conclusão de Curso - MBA em Ciência de Dados (CEMEAI)

ICMC USP
SÃO CARLOS

REFERÊNCIAS

Marzia Petrucci

Topological Data Analysis for Air Traffic with ANAC Data

Final Paper submitted to the Center for Mathematical Sciences Applied to Industry of the Institute of Mathematics and Computer Sciences – USP, in partial fulfillment of the requirements for the MBA in Data Science.

Concentration Area: Data Science

Advisor: Prof. Dr. Antonio Castelo Filho

USP – São Carlos
January 2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Petrucci, Marzia
P498a Análise topológica do tráfego aéreo com dados do ANAC / Marzia Petrucci; orientador Antonio Castelo Filho. -- São Carlos, 2024.
52 p.

Trabalho de conclusão de curso (MBA em Ciência de Dados) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Análise topológica dos dados. 2. Topological Data Analysis. 3. TDA. 4. ANAC. I. Castelo Filho, Antonio , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Marzia Petrucci

Análise topológica do tráfego aéreo com dados do ANAC

Monografia apresentada ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciências de Dados.

Área de Concentração: Ciências de Dados

Orientador: Prof. Dr. Antonio Castelo Filho

USP – São Carlos
Janeiro de 2024

Agradeço à Profa. Dra. Cynthia de Oliveira Lage Ferreira pela orientação. Obrigada para me direcionar no processo de aprendizagem dessa nova técnica e suas possibilidades. Gostaria também de agradecer Thiago Ferreira pela ajuda.

Agradeço ao prof. Antonio Castelo Filho pela ajuda no projeto e sua realização.

"Complicare è facile, semplificare é difficile."

Munari

RESUMO

PETRUCCI, MARZIA. **Análise topológica do tráfego aéreo com dados do ANAC**. 2024. 52 p. Monografia (MBA em Ciências de Dados) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Nesse trabalho de final de curso vamos aplicar a análise topológica dos dados para o tráfego aéreo do ANAC. A análise topológica permite investigar as relações entre dados. Essa noção de relação é topológica, porque consideramos uma distância entre par de pontos no espaço dos atributos. Fazendo variar a distância coletamos as componentes conexas, para reconstruir as informações topológicas da variedade descrita pelos pontos. Essa variedade pode ter características diferentes como furos ou vazios. Essas características e as mudanças nelas codificam as relações e podem ser usadas no machine learning. Nos dados do ANAC analisamos os anos entre 2018 e 2023. Comparamos os meses entre eles para verificar as mudanças no tráfego nacional e internacional independentemente da sazonalidade. Sucessivamente confrontamos como o tráfego do aeroporto de Congonhas afetou a topologia, mostramos isso eliminando as rotas por esse aeroporto e analisando as mudanças .

Palavras-chave: Análise topológica dos dados, análise qualitativa, ANAC.

ABSTRACT

PETRUCCI, MARZIA. **Topological Data Analysis for Air Traffic with ANAC Data**. 2024. 52 p. Monografia (MBA em Ciências de Dados) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

In the work for the finalization of the MBA in Data Science, we apply Topological Data Analysis on the air traffic for the ANAC. Topological Data Analysis allows us to investigate the relations between the data. This notion of relations is topological because we consider a distance between a pair of points in the attributes space. Gauging the distance we collect the connected components, to reconstruct the topological information of the manifold described by the data points. This manifold can have different characteristics as holes and voids. These characteristics and their changes codify the relations and can be used in machine learning. In the ANAC data, we analyzed the years between 2018 and 2023. We compared each month to verify the changes in the topology of the National or International air traffic independently from the seasonality. Afterwards, we compared how eliminating the air traffic from Congonhas airport would affect the topology.

Keywords: Topological Data analysis, TDA, ANAC.

LIST OF FIGURES

LIST OF CHARTS

CONTENTS

CHAPTER

1

INTRODUCTION

In this monograph for the conclusion of the MBA in Data Science we apply the Topological Data Analysis for the ANAC data of air traffic in Brazil. This is an attempt to contribute to spreading the TDA tools to the scientific community. The TDA method is very recent in scientific production and represents a new way to look at data. In this chapter we are going to present the motivation and objectives, and at the end, we present the organization of the monograph.

1.1 Context

Topology is a way to look at the shapes' objects or surfaces. It is very intuitive, as our human capacity to differentiate objects has the same logic. For example, we can distinguish 0 from 8 counting the number of loops. This characteristic is invariant of small deformations, rotations and scale transformations. Whatever the environment, or the orientation of the number we can easily tell if it's a 0 or an 8 (CARLSSON, 2020). Let us see another example: a 3D surface, a torus. The torus is the surface of the donut, its icing. We can draw two different kinds of paths on the donut: one encircling the central hole, co-planar, and one around the pastry, transversal; as shown in Figure 1. These two are different kind of paths because we can not

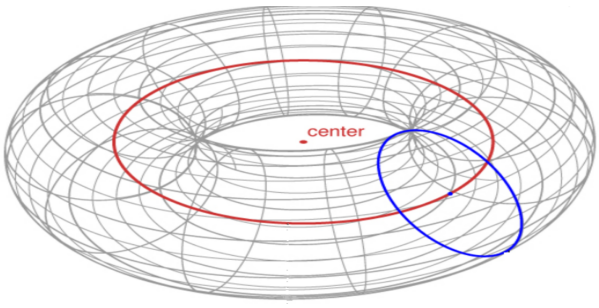
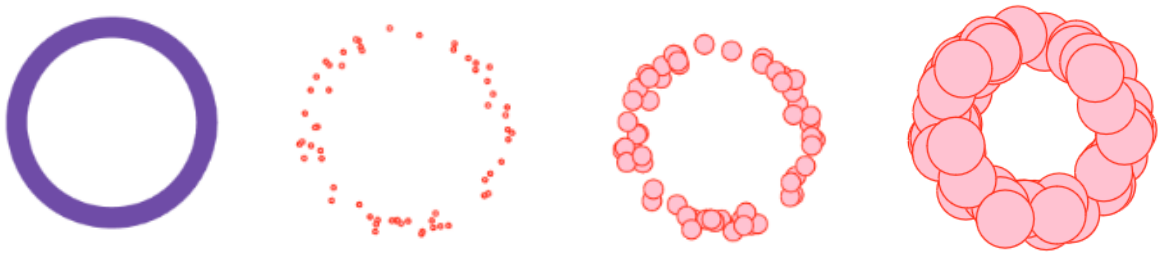


Figure 1 – Torus and two loops representing the fundamental classes of possible paths on the surface[1].

[1] https://mathmonks.com/torus. Accessed 12/16/2023

deform one into the other. They form two distinct classes. The same procedure is carried out in a mathematically rigorous way by topology and, in specific, homology which counts the number of loops, or classes. For more formal definitions please consider Carlsson's book (CARLSSON; VEJDEMO-JOHANSSON, 2021).

In the present work, we apply the knowledge of topology to Big Data with the recent method called Topological Data Analysis (TDA). Topological Data Analysis develops from the data points a graph, then a triangulation, or the equivalent in higher dimension a simplicial complex. The simplicial complex is a geometrical object with information about the relational proprieties of the points. To apply this method we look at the data in the high dimensional space of their attributes (for structured data). The insight is to consider the points as representatives of a surface in this high-dimensional space. We built the method to retrieve the surface, i.e. the continuous path between the points; to achieve this result, we use a distance metric (CARLSSON, 2009; CARLSSON, 2014).



(a) Support S. (b) $\varepsilon = 0.03$. (c) $\varepsilon = 0.10$. (d) $\varepsilon = 0.30$.

Figure 2 – The circle S 2a, then it is represented by approximately 60 data points. In 2b, 2c, and 2d a disk is grown for each point, the union of balls with radius $\varepsilon = 0.03, 0.10, 0.30$, representing S. (WASSERMAN, 2018)

Starting at each point, we grow a high dimensional sphere, so when two spheres touch we can draw a link between the two data points. This procedure of growing a sphere on each point is the scale of the metric swiping the radius from 0 to infinity. As shown in Figure 2, where the support is a circle in 2a, then it is represented by some random points, from each point a disc is grown in 2b, 2c, and 2d increasing the radius. When two spheres touch we can draw a link between the points, we are building a graph. When three links are closed in a triangle with data points as vertices, we count the surface between them. When four triangles form a tetrahedron we count the volume enclosed (WASSERMAN, 2018).

Along with the swipe of the radius we collect information about the persistence of the connected components. The number of the connected surfaces at that particular radius, which gives us the H_0 information studied in persistence homology. The connected components can further form loops on the surface created, this is encoded in the H_1 space. Considering the volumes, we can find also voids formed in the high dimensional manifold, so we can retrieve the H_2 information. This process is called the Vietoris-Rips filtration.

In the following, we list the construction of Vietoris-Rips complex (ZOMORODIAN, 2010).

With a distance metric, such as the Euclidean metric, the distance filtration assigns weights based on pairwise distances between points. The growth of the distance forms simplicial complexes: 0-complexes are links connecting points, 1-complexes are surfaces enclosed by these links and are triangles (or squares), 2-complexes are volumes carved from these surfaces and are solids, and so on... Complexes need to be sorted in ascending order of their weights; in the case of a tie, the lower complex precedes the higher one: that is the filtration.

Moreover, the simplicial complexes form a manifold in the attributes high dimensional space. We follow the construction of the manifold gauging the distance metric, in particular the features that describe topologically the manifold. Types of topological features:

- dimension 0: number of connected components (H_0)
- dimension 1: classes of cycles (H_1)
- dimension 2: classes of voids (H_2)
- ...

To continue the list of features, we have to abandon our 3- dimensional visualization capacity, but we enlist void regions of the space in that dimension. So the topology is interested in the connectivity of the region, where are the voids, and how we can make a continuous path in the manifold. The appearance and disappearance of a connected component, a void, or a circle is called its life and it is represented with a persistence diagram (or persistence barcode). If the life never ends while the metric distance goes to infinity, it has a weight infinity. Persistent homology is capable of preserving distances under random projections (SHEEHY, 2014). (WASSERMAN, 2018; SIZEMORE *et al.*, 2019; TIERNY *et al.*, 2018; ALI *et al.*, 2023; CHAZAL *et al.*, 2013)

With the persistence homology, we achieve the persistence diagrams, representing the life span of the components. The diagram has the horizontal axis as the creation of the component in the metric distance ε, while on the vertical axis its destruction, in ε scale. From the persistence diagram, we calculate the Betti Curve, which is a function mapping a persistence diagram to an integer-valued curve (RIECK; SADLO; LEITTE, 2020; HENSEL; MOOR; RIECK, 2021).

1.2 Problem

We apply the methodology explained in the previous section to the data on air traffic in Brazil. The route network for domestic flights is depicted in Figure 3.

We consider the years from 2018 until 2023. The datasets represent the information for every route: the number of passengers, the loads, the kilometers, the flights, and the fuel. For

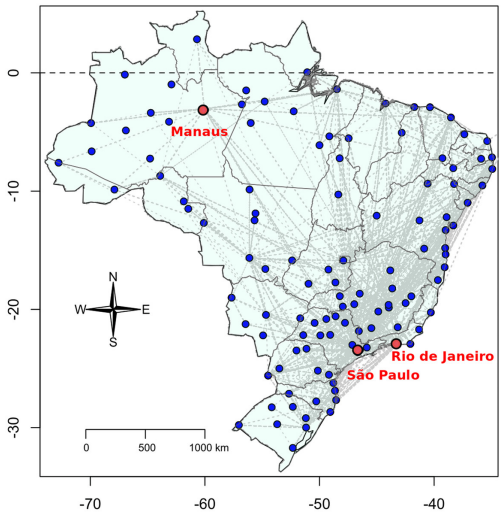


Figure 3 – Route network, domestic flights (RIBEIRO *et al.*, 2020).

each entry, all this numeric information identifies the coordinates in the high dimensional space of the attributes. Each entry is a point in this space. We work to retrieve the variations of the topology of the point clouds due to some specific occurrence.

The problems analyzed are:

- the pandemic.
- the elimination of the Congonhas airport.

The pandemic in 2020 is an event of high interest due to the almost complete stop of passenger flights for civilians. Although the routes were still active for cargo and public service officers. An interesting question is if the stop that happened in 2020 was disruptive enough to change the topology of the point clouds.

The Congonhas airport is the second most transited airport in Brazil, the first most transited for domestic flights with the most important air route São Paulo-Rio de Janeiro. The airport has limitations for the gross load of the aircraft, and number of flights per hour, and faces constant congestion both for the number of passengers and of flights. The airport is located in a central region of São Paulo city and there are many concerns about its impact on the population, air quality and car congestion in this region. Moreover due to safety reasons, as the lethal accident in 2007 has dramatically proved. A question we are investigating is the effect of the elimination of the routes connected to Congonhas on the topology.

1.2.1 Expected Results

The results expected are:

- the disruption of the topological features for the 2020 dataset,

- evaluating the importance of the Congonhas airport in the network, eliminating it from the datasets.

1.3 Organization

To answer these questions we organized this work with a review of the previous works, in Chapter 2 we present the papers and authors first introducing this method, the TDA. We also list the principal fields where the major contributions to TDA specifically were applied to machine learning.

The analysis of the problem starts in following Chapter 3, where we describe the Crisp methodology for the Data Science projects: the understanding of the problem, understanding of the data, and the model applied. The chapter continues on how the methodology is applied to our TDA project with the description of the data and the model in analysis.

In Chapter 4, we introduce the graphical results of the TDA application.

Lastly in Chapter 5, we discuss the results, the shortcomings, the advantages and the possible improvements.

CHAPTER

2

BIBLIOGRAPHIC REVIEW

Topological Data Analysis is an application in Data Science of ideas and methods of various fields of mathematics. This area of study is very recently developed in a seminal paper in 2009 (CARLSSON, 2009). The scientific applications of persistence diagrams is applied in the last years. This field is very recent, not very well known as the mathematical concept not well spread in the computational community.

We review the most important papers and authors translating the procedures for Data Science (DS) in the 2.1 section. We list some of the most recent applications of the TDA method in a plethora of sectors in 2.2. For a hands-on approach, we collect the libraries available for the implementation of the TDA analysis and some papers where can be found a comparison of the libraries in 2.3.

2.1 Theory

In structured data the effort to reduce the high dimensions of the data without losing important information has led to numerous techniques using linear and not linear methods (AYESHA; HANIF; TALIB, 2020). Moreover, the possibility to use the topological machinery for Big Data naturally simplified the task, as the values in a high dimensional space are looked at as point clouds, as points with a distance function (CARLSSON, 2009; DEY; WANG, 2022; CARLSSON; VEJDEMO-JOHANSSON, 2021). Equipped with the notion of a distance the space can be considered as continuous: so the characteristics of the manifold can be considered.

On the positive side, this approach is independent of the scale, and the coordinate system and retrieves information inherent to the data itself; it is robust against noisy and incoherent points. So, the problem then shifts to recognizing the surface represented by the point cloud data and understanding its shape (DEY; WANG, 2022). Where it is continuous, where are the holes, what are the connected components, etc... Even further the shape can be studied through

persistence homology, the analysis of the durability of the characteristics in relation to the distance considered. This information is encoded in a signature, called a barcode (CARLSSON, 2014) or in the persistence diagrams. For those interested in practical use, in this tutorial, there are the guidelines to calculate and implement the persistence homology algorithms to different types of data sets (OTTER; TILLMANN; GRINDROD, 2017).

2.2 Applications

The vast applicability of this recent method is applied to structured data and unstructured ones as images. In these reviews (LEYKAM; ANGELAKIS, 2023; CHAZAL; MICHEL, 2021) the authors present the various fields of application for TDA in machine learning, the first citation especially for TDA applied to physics. In the field of extracting context from text, TDA can contextualize word embeddings and construct a network of coherent topics together with meaningful relationships between them (BYRNE *et al.*, 2022). For images, it can recognize information in environment maps with low-resolution maps (OFORI-BOATENG *et al.*, 2021). The method is useful also for time-dependent information, TDA was useful for classifying time series: it extracted the essential structure of the time series (KARAN; KAYGUN, 2021; UMEDA, 2017; RAVISHANKER; CHEN, 2019; SEVERSKY; DAVIS; BERGER, 2016). In the context of finances, financial time series, it predicted crashes (GIDEA; KATZ, 2018; YEN; CHEONG, 2021).

Using unstructured data, TDA is also used in bio-medicine (CARLSSON, 2017), the following article analyses the advantages and disadvantages of topological tools in the field (SKAF; LAUBENBACHER, 2022). For images used for cancer detection implementing persistence homology as an input for a classification algorithm in machine learning (SINGH; FARRELLY; HATHAWAY, 2023), or for gene sequencing (RABADAN; BLUMBERG, 2019). Some recent successes of TDA in oncology, specifically in predicting treatment responses and prognosis, tumor segmentation and computer-aided diagnosis, disease classification, and cellular architecture determination can be found in (BUKKURI NOEMI ANDOR, 2021; LOUGHREY *et al.*, 2021). In neuroscience, it is used to better understand the gene interaction (SIZEMORE *et al.*, 2019). Another example of gene sequencing, TDA can identify and control emerging adaptive mutations in large genomics datasets, such as of SARS-CoV-2 genomes (BLEHER *et al.*, 2022).

As pointed out by (CHAZAL; MICHEL, 2021) TDA, and in specific persistence homology, in the machine learning context, is used:

- to visualize and as a tool for data exploratory analysis, for example with the Mapper algorithm, (CARLSSON; VEJDEMO-JOHANSSON, 2021), used to extract a graph from the points cloud.

- in features engineering, the persistence diagram data can be successfully introduced into

the machine learning pipeline (ALI *et al.*, 2022).

- for the optimization of the machine learning architecture, to design, to improve and to select models.

- as a stabilization against attacks in deep learning structures with topological layers (GABRIELSSON *et al.*, 2020).

The information in this section is summarized in Chart 1, divided into macro-fields.

Chart 1 – Bibliographic review of TDA applications divided into fields (self-authorship).

FIELD	CITATIONS
word embeddings	(BYRNE *et al.*, 2022)
image	bio-medicine: (CARLSSON, 2017) (SKAF; LAUBENBACHER, 2022) cancer detection: (SINGH; FARRELLY; HATHAWAY, 2023) maps: (OFORI-BOATENG *et al.*, 2021)
physics	(LEYKAM; ANGELAKIS, 2023) (CHAZAL; MICHEL, 2021)
genomics	(BLEHER *et al.*, 2022) (RABADAN; BLUMBERG, 2019)
neuroscience	(SIZEMORE *et al.*, 2019)
oncology	(BUKKURI NOEMI ANDOR, 2021) (LOUGHREY *et al.*, 2021)
time-series	(KARAN; KAYGUN, 2021) (UMEDA, 2017) (RAVISHANKER; CHEN, 2019) (SEVERSKY; DAVIS; BERGER, 2016)
finance	(GIDEA; KATZ, 2018) (YEN; CHEONG, 2021)
machine learning	(GABRIELSSON *et al.*, 2020) (ALI *et al.*, 2022) (CHAZAL; MICHEL, 2021) (CARLSSON; VEJDEMO-JOHANSSON, 2021)

2.3 Implementation

Libraries applying the TDA method are already available in Python. The main ones are: giotto (TAUZIN *et al.*, 2021), gudhi (MARIA *et al.*, 2014), pytorch-topological (HENSEL; MOOR; RIECK, 2021) and scikit-tda.

In this reference, there is a comparison between the various implementations of the Mapper algorithm (RAY; TROVATI, 2018), while here for Persistence Homology ones (OTTER; TILLMANN; GRINDROD, 2017).

There is a database of bibliography articles on real-world applications of TDA, DONUT (GIUNTI; LAZOVSKIS; RIECK, 2023).

Let us dive into our problem, we analyze the ANAC data and apply the TDA method in the following chapter.

CHAPTER

3

METHODOLOGY

For the implementation of the current work, we used the CRISP-Data Science methodology. The methodology organizes the workflow toward minimizing the possibility of failure in Data Science projects. Every step is fundamental in the process and it can be revisited many times before the end of the project. The steps are:

- Understanding the problem 3.1;
- Understanding the data at disposal 3.2;
- Preparing the datasets 3.2.1:
  - Collecting the datasets,
  - Selecting the valuable attributes,
  - Data treatment 3.3,
- Model: Topological Data Analysis 3.4;
- Evaluation of the results for the problem proposed 5,
- Implementation.

In the context of the MBA in Data Science, we are not going to implement the model in any application. Let us adapt the other parts of the methodology to our problem.

3.1 Problem

The problem proposed is:

- to analyze the air traffic in Brazil between the years 2018 and 2023 and look at how the pandemic impacted the data,

- to analyze how the elimination of the Congonhas airport in São Paulo transformed the topology.

We recognized the seasonality of the air traffic data due to national and International festivities, therefore we will compare the months of different years.

To analyze the pandemic we will consider the 2020 data as a reference, however in 2021 and 2022 flights had still restrictions for the vaccination protocols adopted Internationally.

3.2 Material

In this specific work, we concentrate on structured datasets of air traffic in Brazil. The datasets used are from the "Agência Nacional de Aviação Civil" (ANAC) (INFRAESTRUTURA, 2023), well-known in the scientific community as in (EDUARDO OLIVEIRA, 2023; Bazzo Vieira; Vieira Braga; PEREIRA, 2022; RIBEIRO *et al.*, 2020). The datasets are divided for year, and presented in a data frame with every entry a route traveled by a company. The attributes specified are in Chart 2.

3.2.1 Datasets

The datasets downloaded from the ANAC webpage (INFRAESTRUTURA, 2023) are divided into years. The data is presented in a tabular format (.csv files), a structured dataframe. The dimensions for each dataset are:

- 2018: 39753 rows x 38 columns
- 2019: 37849 rows × 38 columns
- 2020: 26821 rows × 38 columns
- 2021: 28466 rows × 38 columns
- 2022: 35706 rows × 38 columns
- 2023: 32365 rows × 38 columns

Chart 2 – Features in the ANAC's datasets (self-authorship).

	Original	**Translation**
1	EMPRESA (SIGLA)	COMPANY (ACRONYM)
2	EMPRESA (NOME)	COMPANY (NAME)
3	EMPRESA (NACIONALIDADE)	COMPANY (NATIONALITY)
4	ANO	YEAR
5	MÊS	MONTH
6	AEROPORTO DE ORIGEM (SIGLA)	AIRPORT OF ORIGIN (ACRONYM)
7	AEROPORTO DE ORIGEM (NOME)	AIRPORT OF ORIGIN (NAME)
8	AEROPORTO DE ORIGEM (UF)	AIRPORT OF ORIGIN (STATE)
9	AEROPORTO DE ORIGEM (REGIÃO)	AIRPORT OF ORIGIN (REGION)
10	AEROPORTO DE ORIGEM (PAÍS)	AIRPORT OF ORIGIN (COUNTRY)
11	AEROPORTO DE ORIGEM (CONT.)	AIRPORT OF ORIGIN (CONT.)
12	AEROPORTO DE DESTINO (SIGLA)	DESTINATION AIRPORT (ACRONYM)
13	AEROPORTO DE DESTINO (NOME)	DESTINATION AIRPORT (NAME)
14	AEROPORTO DE DESTINO (UF)	DESTINATION AIRPORT (STATE)
15	AEROPORTO DE DESTINO (REGIÃO)	DESTINATION AIRPORT (REGION)
16	AEROPORTO DE DESTINO (PAÍS)	DESTINATION AIRPORT (COUNTRY)
17	AEROPORTO DE DESTINO (CONT.)	DESTINATION AIRPORT (CONT.)
18	NATUREZA	NATURE
19	GRUPO DE VOO	FLIGHT GROUP
20	PASSAGEIROS PAGOS	PAID PASSENGERS
21	PASSAGEIROS GRÁTIS	FREE PASSENGERS
22	CARGA PAGA (KG)	PAID CARGO (KG)
23	CARGA GRÁTIS (KG)	FREE CARGO (KG)
24	CORREIO (KG)	MAIL (KG)
25	ASK	Available Seat-Kilometers
26	RPK	Revenue Passenger-Kilometers
27	ATK	Available tonne kilometer
28	RTK	Revenue tonne kilometer
29	COMBUSTÍVEL (LITROS)	FUEL (LITERS)
30	DISTÂNCIA VOADA (KM)	DISTANCE FLIGHT (KM)
31	DECOLAGENS	TAKE-OFFS
32	CARGA PAGA KM	PAID CARGO KM
33	CARGA GRATIS KM	FREE CARGO KM
34	CORREIO KM	MAIL KM
35	ASSENTOS	SEATS
36	PAYLOAD	PAYLOAD
37	HORAS VOADAS	HOURS FLIGHT
38	BAGAGEM (KG)	BAGGAGE (KG)

3.3 Cleaning

Since in our analysis, we are not interested in specific companies but in the network formed by the air traffic we eliminated every information regarding the company, in Chart 2 we deleted the attributes in lines 1, 2, and 3.

We were interested in the information about the round trip: to this objective, we unified the route from 'AEROPORTO DE ORIGEM (SIGLA)' ('AIRPORT OF ORIGIN (ACRONYM)') and to 'AEROPORTO DE DESTINO (SIGLA)' ('DESTINATION AIRPORT (ACRONYM)') with an attribute 'COD' combining the two acronyms in alphabetic order. In this manner, we simplified the information in Chart 2 in lines from 6 to 17 in one column.

The information as 'NATUREZA'('NATURE') in line 18 of Chart 2 specifies if it is an International route or a domestic one. We discarded this attribute.

The column 'GRUPO DE VOO' means the kind of flight: 'REGULAR' (regular), 'NÃO REGULAR' (not regular) and 'IMPRODUTIVO' (unproductive); this information was not relevant in our analysis, so it has been canceled.

Attributes 20 and 21 count the passengers' number, paying and free: these attributes were summed in a unique column. The passengers' numbers were added in a unique column called 'PASSAGEIROS'.

The same procedure was applied to the information from attributes 22, 23 and 24 of kilos' load unified in 'CARGA (KG)' and the information from attributes 32 to 34 of kilometers traveled by the load unified in 'CARGA (KM)'.

Columns 25, 26, 27 and 28 are acronyms to extract the information of how many passengers per kilometer and how much cargo per kilometer on the route. As they have information that can be extracted from other attributes, these columns have been canceled.

In Chart 3 we summarized the interventions on the datasets attributes.

After the cleaning of the attributes of the dataframe are: ANO, MÊS, COD, COMBUSTÍVEL (LITROS), DISTÂNCIA VOADA (KM), DECOLAGENS, ASSENTOS, PAYLOAD, HORAS VOADAS, BAGAGEM (KG), CARGA (KG), CARGA (KM), PASSAGEIROS.

The following step was to clean and treat the entries of the dataframe. The column 'HORAS VOADAS' was of 'object' type, so we changed it for the float64. We summed all the occurrences with the same 'COD', 'ANO', and 'MES'. After the sum, the duplicates were removed from the datasets, and so were the lines with only NaN. After the reduction of the attributes, the lines with more than six NaN were also removed. The number six was chosen as above average among the ten attributes with numerical values.

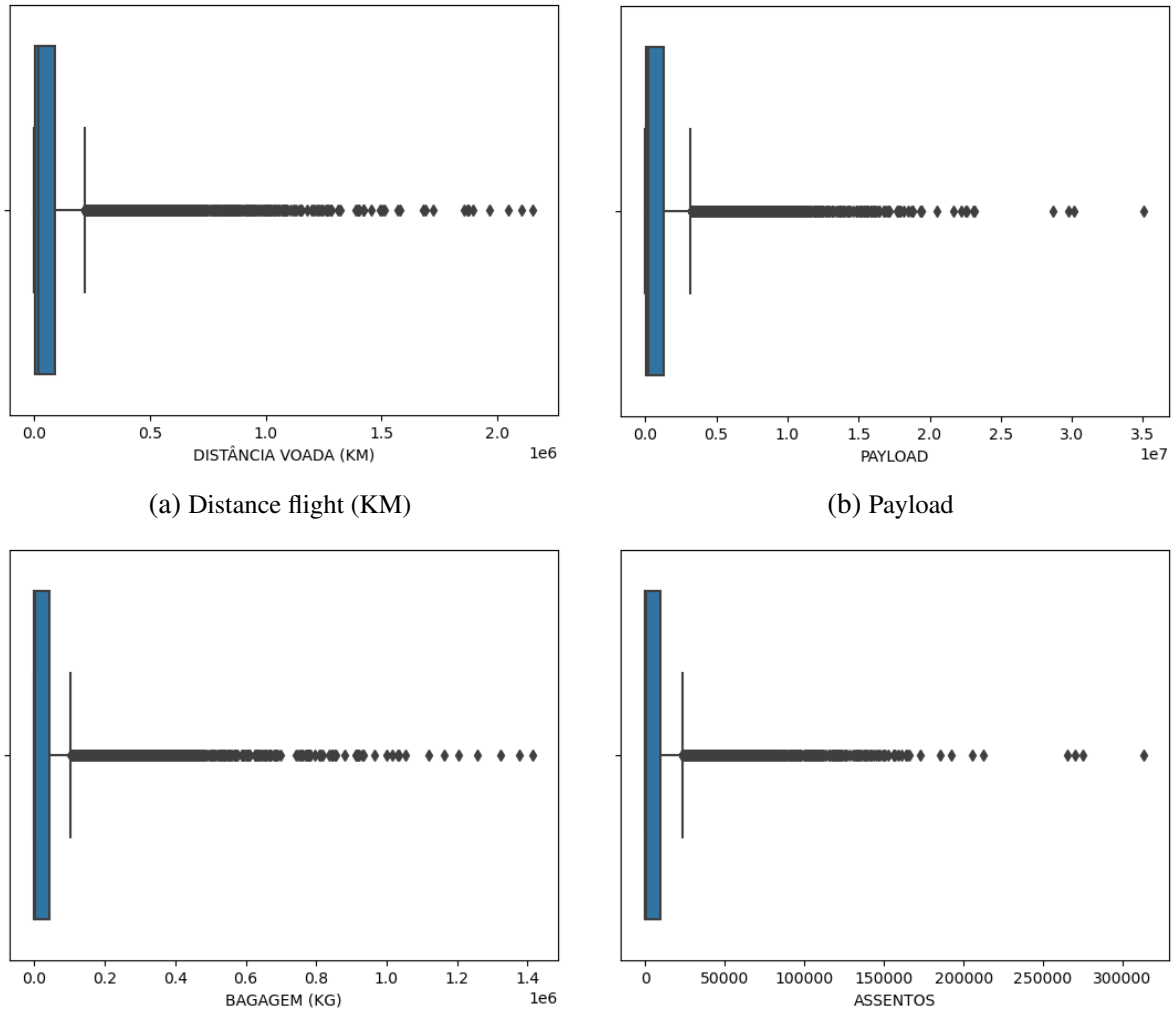
After the end of the treatment executed, the dimensions for each dataset were reduced by a factor 3 approximately:

Chart 3 – Cleaning procedure for the ANAC's datasets (self-authorship).

	Attributes	**Actions**
1, 2, 3	COMPANY (ACRONYM), COMPANY (NAME), COMPANY (NATIONALITY)	DELETED
4, 5	YEAR, MONTH	UNCHANGED
6, 12	AIRPORT OF ORIGIN (ACRONYM), DESTINATION AIRPORT (ACRONYM)	UNIFIED IN 'COD'
7, 8, 9, 10, 11	AIRPORT OF ORIGIN (NAME), AIRPORT OF ORIGIN (STATE), AIRPORT OF ORIGIN (REGION), AIRPORT OF ORIGIN (COUNTRY), AIRPORT OF ORIGIN (CONT.)	DELETED
13, 14, 15, 16, 17	DESTINATION AIRPORT (NAME), DESTINATION AIRPORT (STATE), DESTINATION AIRPORT (REGION), DESTINATION AIRPORT (COUNTRY), DESTINATION AIRPORT (CONT.)	DELETED
18, 19	NATURE, FLIGHT GROUP	DELETED
20, 21	PAID PASSENGERS, FREE PASSENGERS	UNIFIED IN 'PASSAGEIROS'
22, 23, 24	PAID CARGO (KG), FREE CARGO (KG), MAIL (KG)	UNIFIED IN 'CARGO (KG)'
25, 26, 27, 28	ASK, RPK, ATK , RTK	DELETED
29	FUEL (LITERS)	UNCHANGED
30	DISTANCE FLIGHT (KM)	UNCHANGED
31	TAKE-OFFS	UNCHANGED
32, 33, 34	PAID CARGO KM, FREE CARGO KM, MAIL KM	UNIFIED IN 'CARGA (KM)'
35	SEATS	UNCHANGED
36	PAYLOAD	UNCHANGED
37	HOURS FLIGHT	UNCHANGED
38	BAGGAGE (KG)	UNCHANGED

- 2018: 10860 rows x 11 columns
- 2019: 11485 rows × 11 columns
- 2020: 8843 rows × 11 columns
- 2021: 9407 rows × 11 columns
- 2022: 11207 rows × 11 columns
- 2023: 10046 rows × 11 columns

We investigate the distributions of the attributes considered after the cleaning process in Figures 4 and 5. We show only the distributions for the year 2021, as an example. The



(a) Distance flight (KM)

(b) Payload

(c) Baggage (KG)

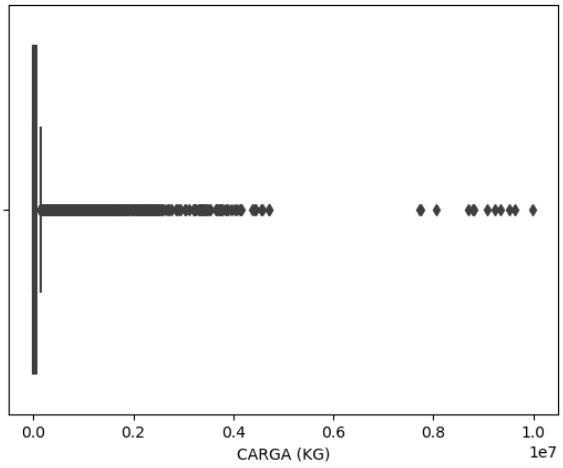
(d) Seats

Figure 4 – Distributions after the cleaning process of the dataset, year 2021 (self-authorship), first part.

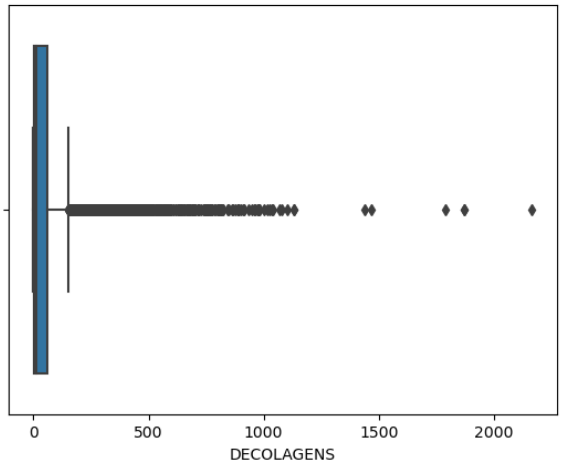
boxplot graphics show the inter-quartile intervals and the medians of the datasets, the outliers are expressed as points outside the whiskers of the inter-quartile intervals. All the usual analytics on these datasets providing quantitative information would have to compensate for the unbalanced distributions with heavy treatments of the data. However, we can consider the TDA method as the 'shape' of data is not altered if the data are scattered in an unbalanced manner.

After the cleaning process, the features were standardized with the StandardScaler() method, so the mean of every attribute is fixed to zero and it is scaled to unit variance. This procedure evens the influence of all the attributes and allows for the information retrieval of the 'shape' of the point clouds in a clean way. This choice was driven by the use of the Euclidean distance to be used in the persistence diagram calculation.

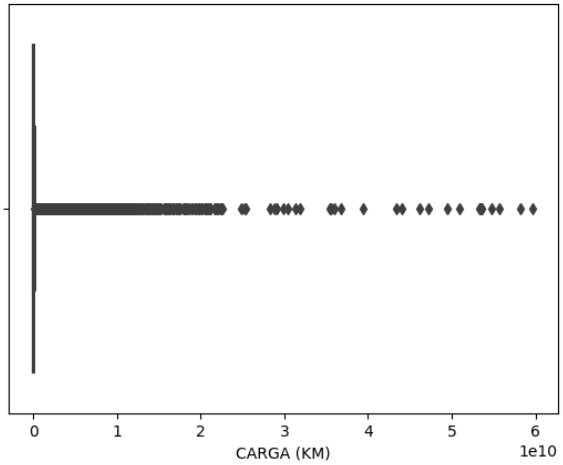
Let us see how in the following section.



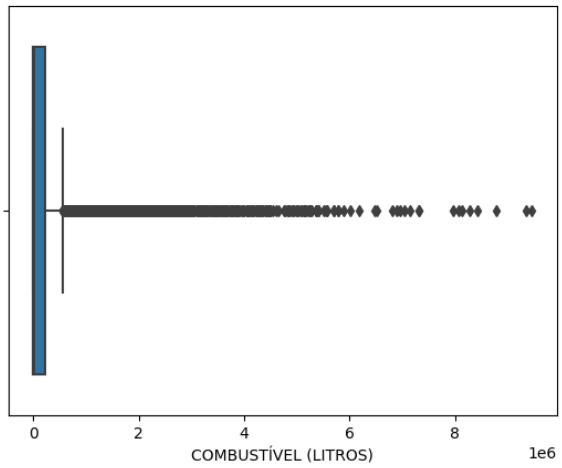
(a) Cargo (KG)



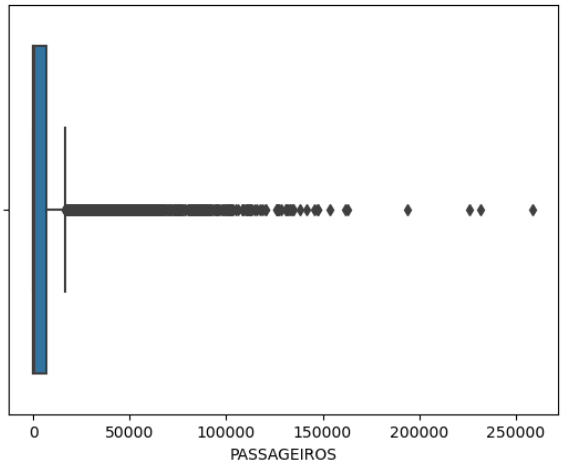
(b) Take-offs



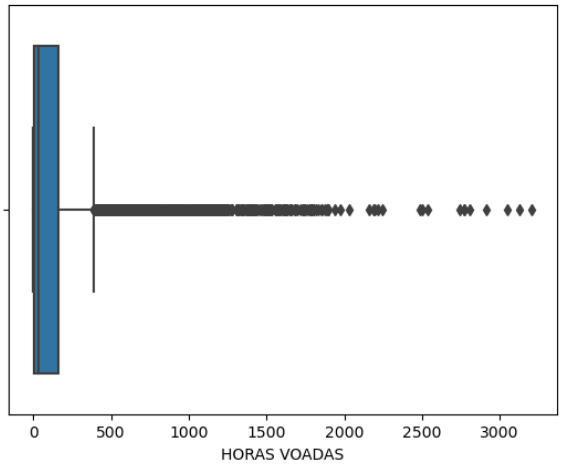
(c) Cargo (KM)



(d) Fuel (Liters)



(e) Passengers



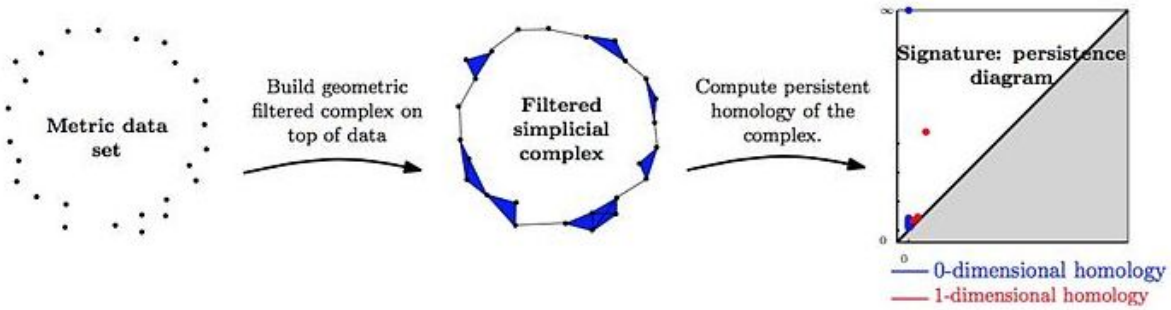
(f) Hours Flight

Figure 5 – Distributions after the cleaning process of the dataset, year 2021 (self-authorship), second part.

3.4 Topological Data Analysis

The Topological Data Analysis (TDA) utilizes the tools from algebraic topology to investigate the relations between high volumes of data. In the TDA we are going to build the Vietoris-Rips complex, which is the topological space from the point clouds of the raw data. The topological space is built from the distance (Euclidean metric). The construction defines the persistence diagram, which shows the life of the homological spaces: H_0, H_1, and H_2, as shown in Figure 6.

Figure 6 – Flow of work with TDA. (CHAZAL *et al.*, 2013)



The attributes express the information about the load, passengers, kilometers, fuel, etc... The values assigned to every entry, which are the air route traveled, form a point in the space of the attributes, a high dimensional space. All the entries are looked at as forming a points cloud, they represent an underlined surface. The Vietoris-Ripser filtration builds a 'stylized' version of the surface building the simplicial complex from the points cloud.

After the cleaning of the datasets and the separation into months of the data, we calculated the Vietoris-Ripser filtration and its persistence diagram for each separated set of data.

The persistence diagrams represent the life of each topological feature of H_0, H_1, and H_2 (connected components, circles, and voids). To compare the different months in the years considered we collected information calculating the Betti Curve of each diagram. The Betti Curve is a representation of the persistence diagram into an integer-valued curve (RIECK; SADLO; LEITTE, 2020). To better visualize the behavior of the 2020 Betti Curve, we calculated the residual of the other curves with respect to it, as a reference.

The TDA implementation chosen 2.3 for the calculations was the giotto-tda (TAUZIN *et al.*, 2021), having the most rapid calculation of the Vietoris-Ripser filtrations. Another advantage was the simplicity and beauty of the graphics.

Let us see our findings in the following chapter.

CHAPTER

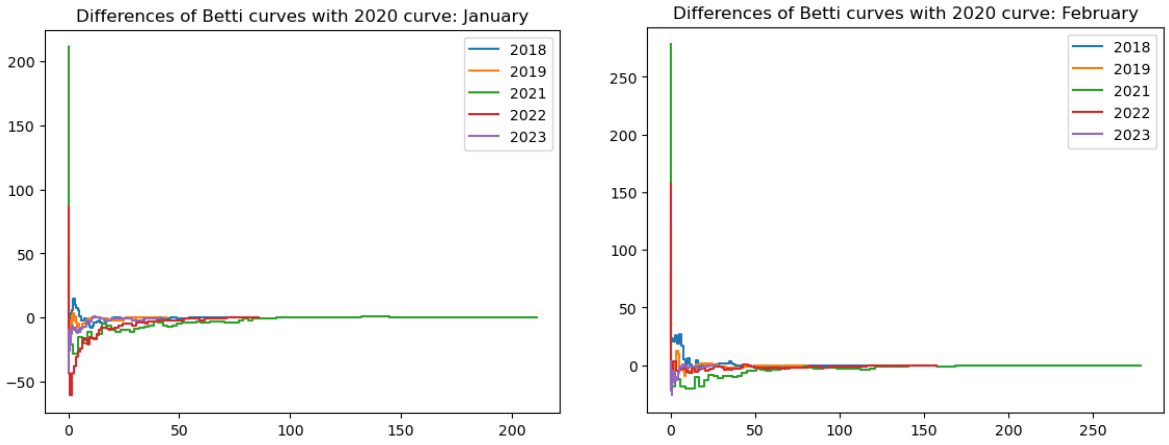
4

RESULTS

In Chapter 3, we introduced the TDA process to calculate the Persistence diagrams; let us see in the following the graphical results.

4.1 Graphics

The results of our analysis are presented by the residuals' Betti Curves, allowing a graphical confrontation of the different persistence diagrams information. The residuals are the differences of the Betti curves with the 2020 curve, which became our zero line. The graphics are divided into months, respecting the seasonality of the analysis. We show the residuals for the connected components, the H_0 information.

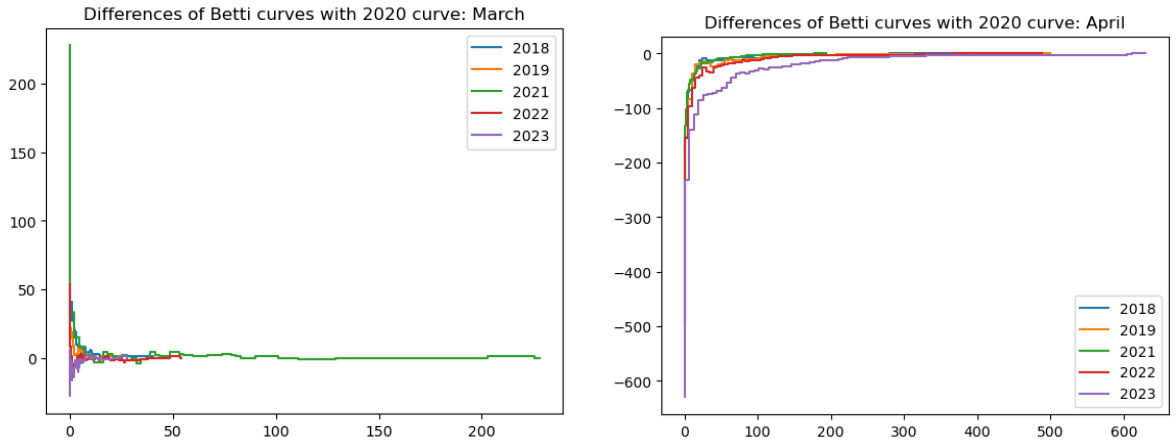


(a) January.

(b) February.

Figure 7 – Residuals of H_0 Betti Curves for January and February in comparison to the 2020 Betti Curve (self-authorship).

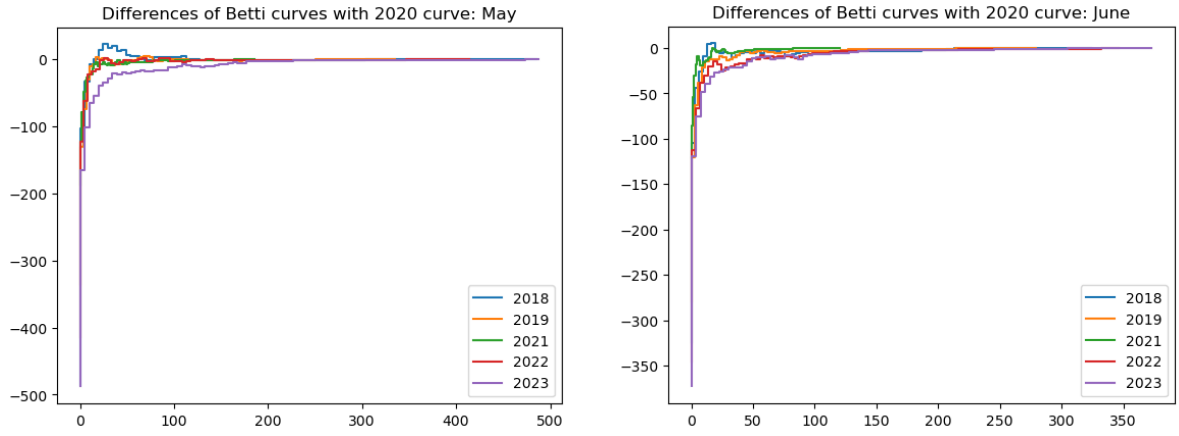
The January and February, in Figure 7 residuals are spread around zero so there is not a clear behaviour respect the 2020 Betti curve.



(a) March (b) April

Figure 8 – Residuals of H_0 Betti Curves for March and April in comparison to the 2020 Betti Curve (self-authorship).

In Figure 8, the March graphic shows the same pattern as the previous months, shown in Figure 7; but the April graphic shows that all the residuals are negative, indicating a clear difference in the behaviour.



(a) May (b) June

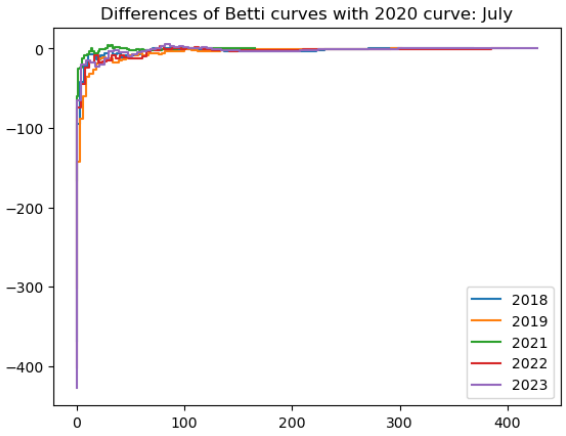
Figure 9 – Residuals of H_0 Betti Curves for May and June in comparison to the 2020 Betti Curve (self-authorship).

In Figure 9 about the residual for May and June, we can also recognize the same pattern, all the residuals are negatives.

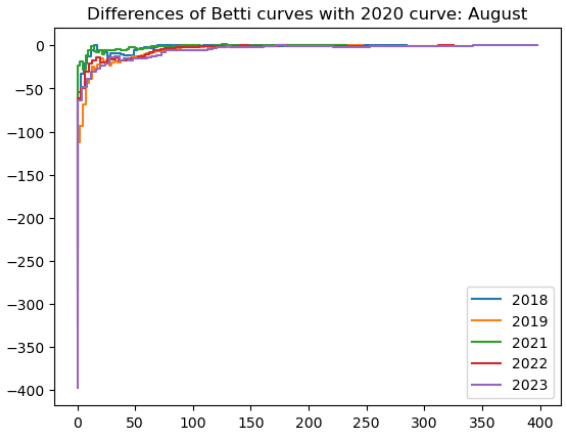
In Figure 10 for July and August, the same behaviour is present, all the residuals are negative.

In Figure 11, for the September graphic, we can see the same behaviour, as all residuals are negative. While in October, November, and December the 2018 and 2019 Betti curve have positive residuals at the beginning of the graphic.

Summarizing the H_0 curves in Figures 7, 8, 9, 10, and 11, we presented that:

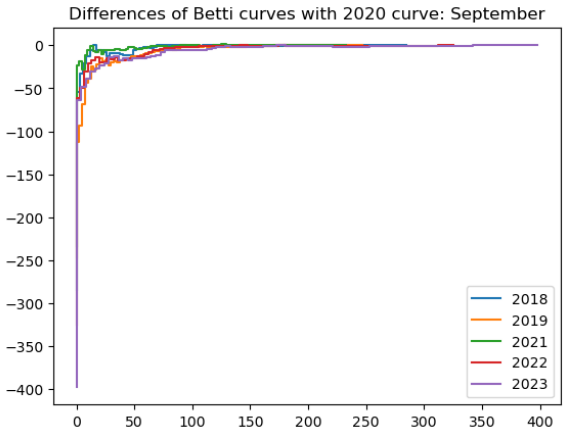


(a) July

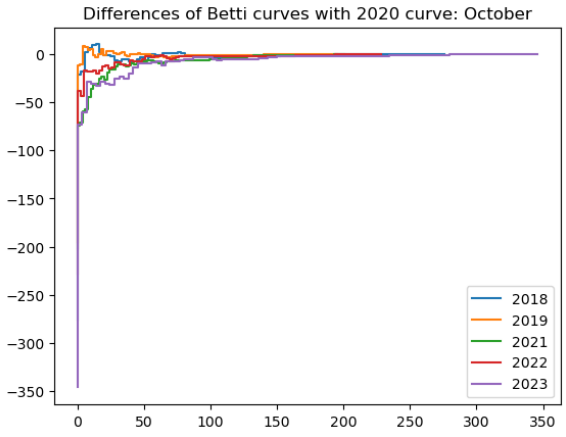


(b) August

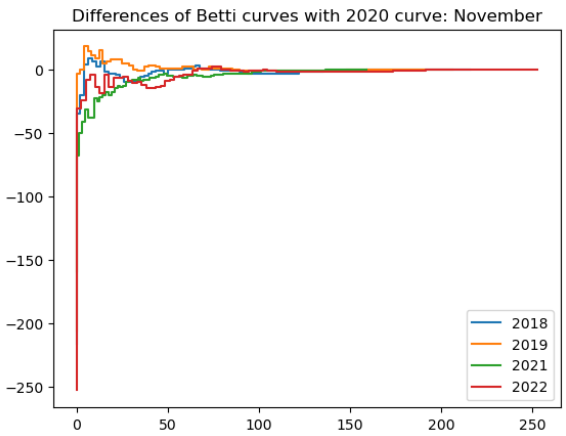
Figure 10 – Residuals of H_0 Betti Curves for July and August in comparison to the 2020 Betti Curve (self-authorship).



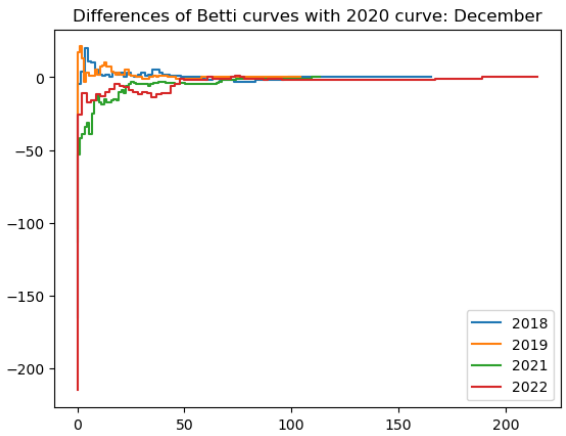
(a) September



(b) October



(c) November



(d) December

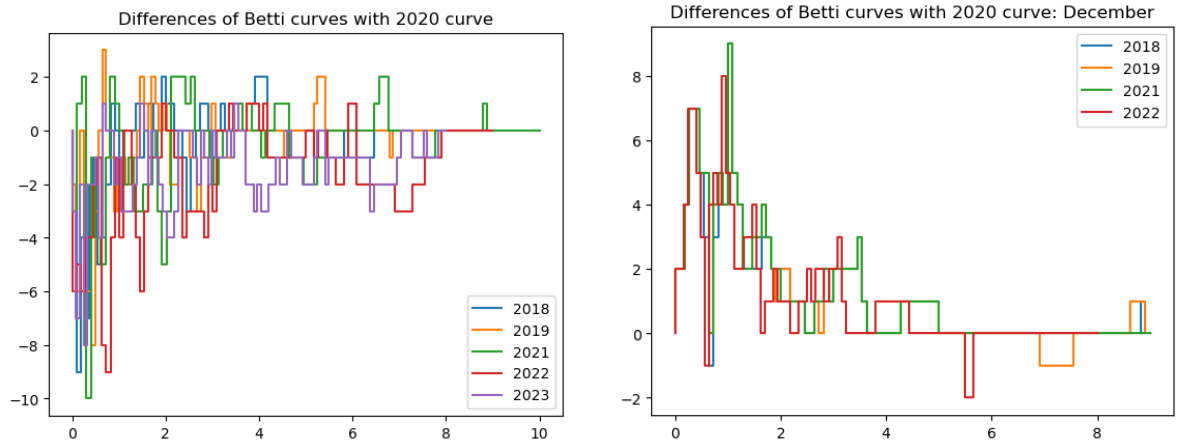
Figure 11 – Residuals of H_0's Betti Curves for September, October, November, and December in comparison with the 2020 curve (self-authorship).

- from January until March, the curves have not a distinguishable behavior from one another, they all spread around the zero, the 2020 curve (7a, 7b, and 8a),

- from April until December, the 2020 curve is well below all the others (from 8b until 11d). All the residuals in these graphics are negatives leading to this consideration.
- in November and December, i.e. 11c and 11d, the 2023 data were not available at time of consultation (November 2023).

The asymmetry of the residuals is relevant to our analysis, showing that the number of connected components in 2020 collapses rapidly from April onward.

The same analysis was executed for H_1, the number of classes of fundamental circles and for H_2, the number of classes of voids. Only the residuals for the H_1, the number of classes of loops, in April 12a and for the H_2, the number of classes of voids, in November 12b were unbalanced. We present these partial results in Figure 12. The November graphic 12b is almost always positive, meaning that the 2020 Betti Curve is above all the others. The residuals were spread evenly around zero in all other graphics.



(a) April, residuals for the H_1.

(b) November, residuals for the H_2.

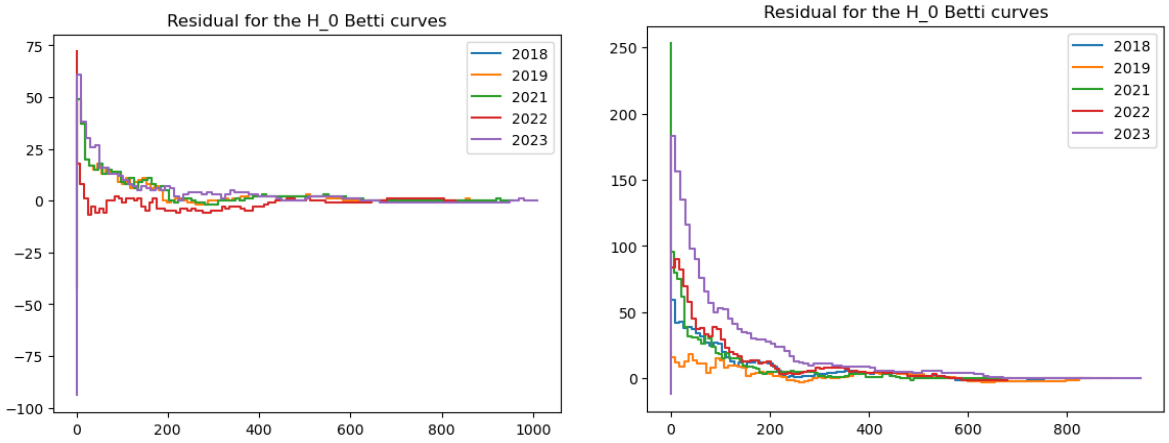
Figure 12 – Residuals for the Betti Curve of the classes of loops of April, and for the classes of voids of November (self-authorship).

We show now the results for our second problem: the analysis of the topology without the routes connected to the Congonhas airport. As in the previous analysis, we considered the 2020 Betti curve as a reference and we calculated the residuals. In this manner we can easily compare the graphics to retrive the influence of the Congonhas airport in the general topology of the years considered.

In Figure 13, the January residuals curves present a spread behaviour around zero; while the February graphic show that the residuals are only positive.

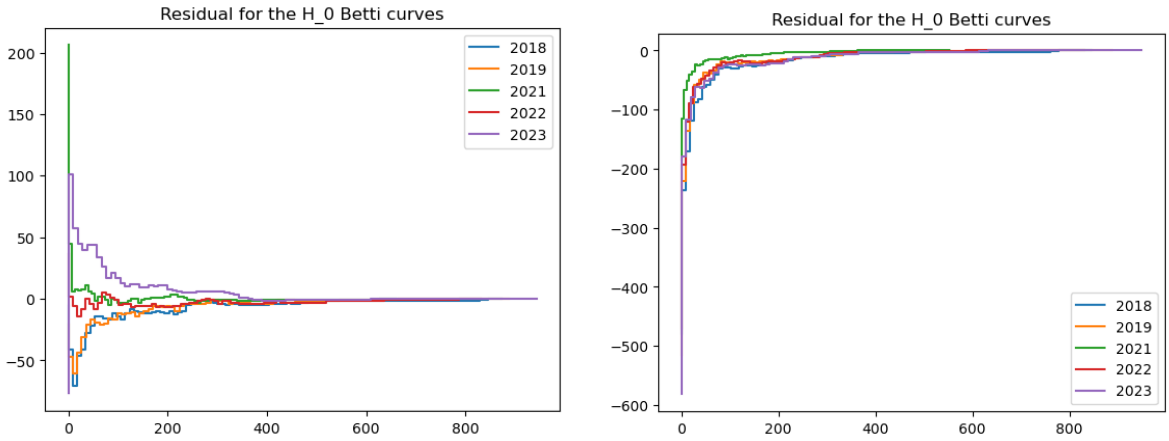
In Figure 14, the residuals curve are spread around zero in March; while in April are al negative as in the case of Figure 8a, with the Congonhas airport routes.

In Figure 15, the May and June graphics show the same behaviour as in Figure 9, all the residuals are negative.



(a) January (b) February

Figure 13 – Residuals of H_0 Betti Curves for January and February in comparison to the 2020 Betti Curve without the Congonhas airport (self-authorship).



(a) March (b) April

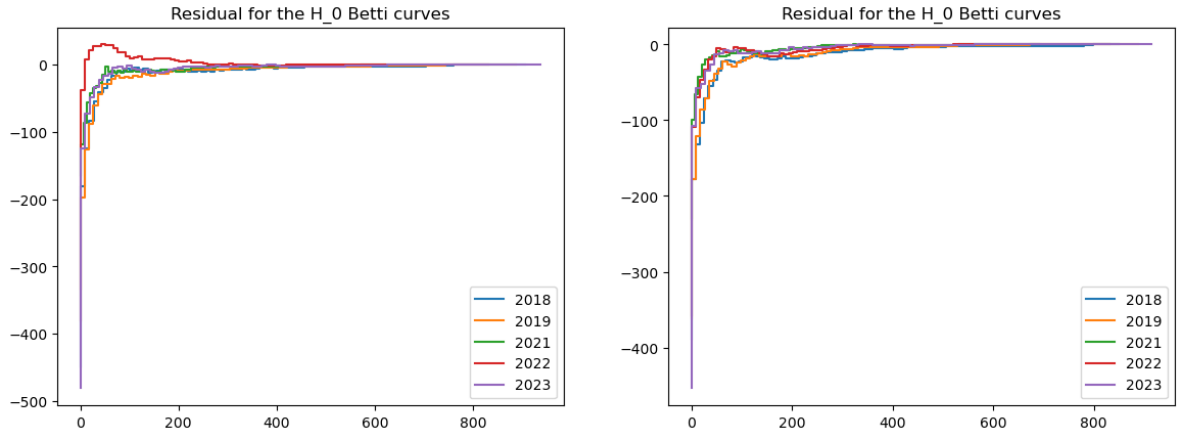
Figure 14 – Residuals of H_0 Betti Curves for March and April in comparison to the 2020 Betti Curve without the Congonhas airport (self-authorship).

In Figure 16 for July and August, the behaviour is practically positive, very different from the behaviour in Figure 10 where the residuals were negative.

In Figure 17a, the behaviour in September is spread around zero, while in Figure 11a the residuals were only negative. In Figure 17b, the residuals' behaviour in October is negative as in Figure 11b.

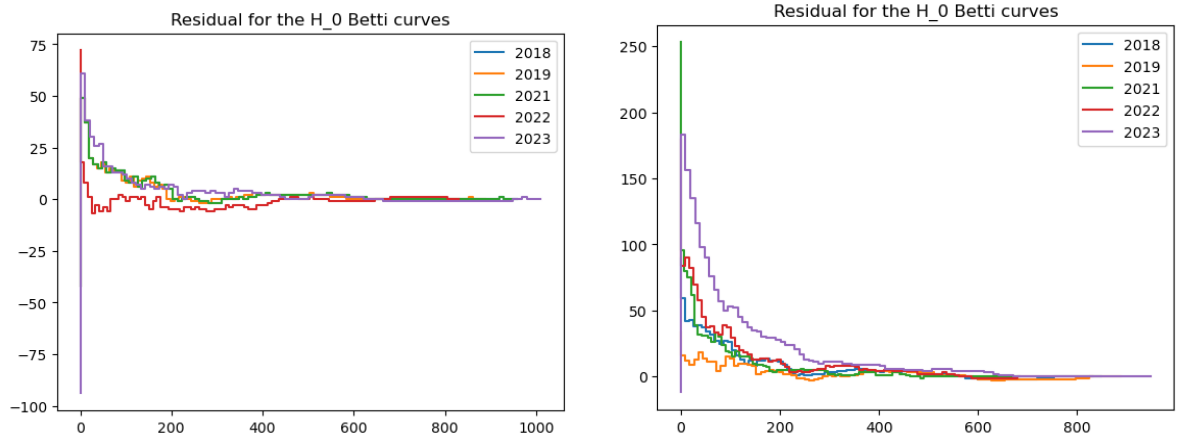
In Figure 18 for November and December, the residuals behaviour is negative as in the case with the Congonhas airport in Figure 11.

The behaviour of the residuals for the connected components, H_0 Betti Curves in Figures 13, 14, 15, 16, 17, and 18 are divided into months for each year. The residuals without the routes connected to the Congonhas airport can be summarized as:



(a) May (b) June

Figure 15 – Residuals of H_0 Betti Curves for May and June in comparison to the 2020 Betti Curve without the Congonhas airport (self-authorship).



(a) July (b) August

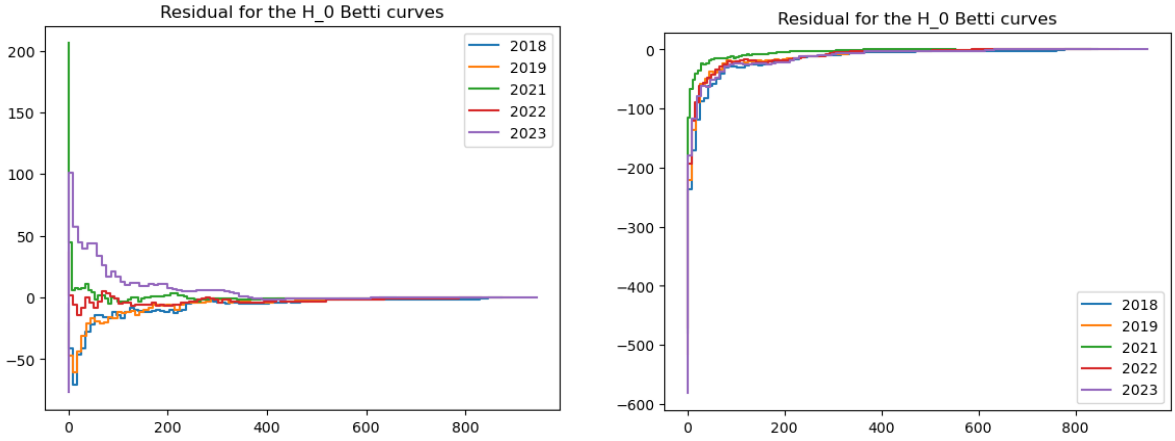
Figure 16 – Residuals of H_0's Betti Curves for July and August in comparison with the 2020 curve without the Congonhas airport (self-authorship).

- the behavior of the curves in January and March (in Figures 13a and 14a) is not so immediately distinguishable compared with the ones in Figures 7a and 8a. Likewise all the curves are grouped together.

- in February the 2020 curve in Figure 13b is above all the others, as the residuals are all positives. While in Figure 7b the residuals are spread around zero.

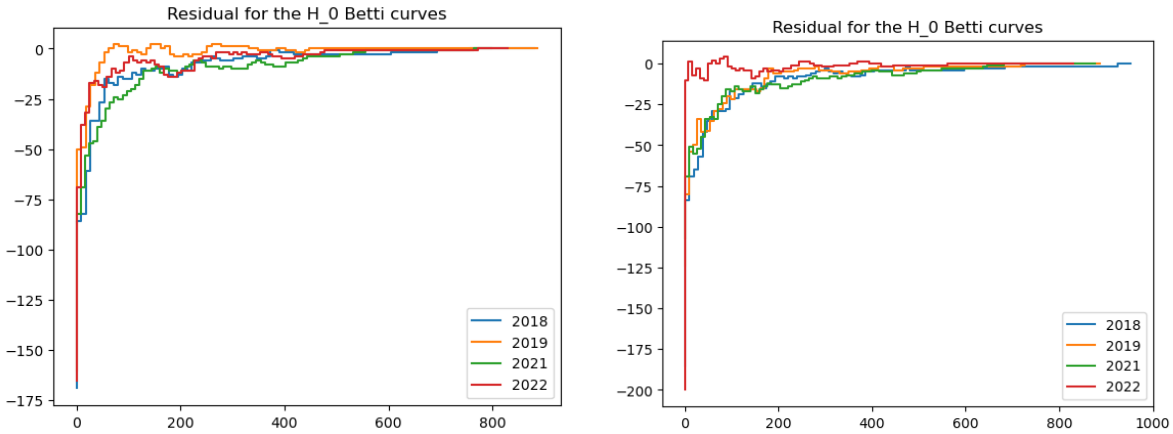
- from April until June (14b, 15a, and 15b) the curves of 2020 and 2021 are below the other curves, as the residuals are all negatives.

- from July until September (16a, 16b and 17a) the curves are all mixed together, spread around zero, or positive as in August.



(a) September (b) October

Figure 17 – Residuals of H_0's Betti Curves for September and October in comparison with the 2020 curve without the Congonhas airport (self-authorship).



(a) November (b) December

Figure 18 – Residuals of H_0's Betti Curves for November and December in comparison with the 2020 curve without the Congonhas airport (self-authorship).

- from October until December (17b, 18a, and 18b) the 2020 Betti curve is well below the others, as all the residuals are negative.

We can compare the connected components with the Congonhas airport and without. The graphics illustrate that the Congonhas airport made a significant contribution in February, July, August, and September (13b, 16a, 16b, and 17a). In these figures, the change in behaviour is easily distinguishable without the Congonhas airport. In Figure 13b the trend of the 2020 curve is remarkably above all other curves. Remembering that the Congonhas airport is the most important one for domestic air traffic, the results heavily show its influence in the months most affected by school and work holidays.

To end our chapter, we summarize how we have applied the TDA method:

- we have calculated the Vietoris-Ripser filtration for each month for the years from 2018 until 2023,
- then we have calculated the persistence diagrams for the connected components, H_0, the classes of loops, H_1, and the classes of voids, H_2,
- we have calculated the Betti Curves, for the three homology spaces H_0, H_1, and H_2, divided in months and years,
- we have fixed the 2020 Betti Curve as referential and calculated the residuals of the other Betti Curves.

In the next chapter, we recollect the conclusions and limitations of this project.

CHAPTER

5

CONCLUSION

In Chapter 3, we analyzed the datasets and, after the cleaning process, we applied the TDA method, specifically the persistence homology. In Chapter 4, we presented the results. The persistence homology calculations resulted in vectors, which were simplified in the Betti Curves to better compare the different years. The comparison was taken with the 2020 curve as a reference, as the zero, we calculated the residuals of the other curves.

5.1 Conclusions

We divided the calculations for the Betti Curve for the connected components (H_0), the circles (H_1) and the voids (H_2). The calculations were done for each month of each year. Our problem is concerned with the differences in the topology. For the first of our problems proposed the behavior of the 2020 topology was better in evidence calculating the residuals with this curve as referential. The most evident results were the curves for the connected components.

We can compare the behavior of the Betti Curve for the connected components for the years 2020 with the pre-pandemic years 2018 and 2019 and with the post-pandemic 2021, 2022, and 2023. The 2020 curve is not present in the graphic as it is our zero line. In graphic 9, from April 8b until December 11d all the curves of the residuals are negative, showing that the 2020 curve is the lowest curve with the steepest descent. In the other graphics 7a, 7b, and 8a the curves are more evenly spread around zero, which leads to the conclusion that the 2020 curve has the same behavior as the other curves.

Considering how removing the routes from or to the Congonhas airport changes the topology is shown in Figures 15 and 18. The graphics show the impact of the airport on the results. We can infer that the Congonhas airport was predominant in the air traffic on February 13b, July 16a, August 16b, and September 17a where its extractions had the major impacts.

The results were achieved with the giotto-tda implementation 2.3, here there is a reference

to its use (TAUZIN *et al.*, 2021). It was better suited for rapid calculations and for the simplicity and beauty of the graphics.

In the Introduction 1, we described the topological properties and its applications to Data Science in 2. The translation of the calculations and concepts to the computational applications was developed in the last 10 years, and its applications to various fields have spread in the last 5 years, as shown in the bibliographic review 2. This work is an attempt to spread the TDA method as another tool in the hands of Data scientists. The comparison with other methods is dependent on the problem and the data available, but we have shown the possibilities and the advantages of looking at the "shape" of data.

5.2 Difficulties, Limitations, Improvements

We can summarize the difficulties encountered in this project. The first was about the quality of data available: the data was incomplete, with many missing values (NaN) and zeros. The treatment of the data took time and a trial and error approach. Our final approach was to consider the route of the air traffic, to do so we summed the different entries, which were divided for air companies. This information was not relevant, so we canceled it. After collecting every route divided per month and year, there were still many missing values, the entries with mostly NaN values were discarded. The zeros were maintained, after the consideration that we didn't want to alter the data or didn't have access to better data to incorporate. This approach can be revised to better the performance of the method.

For what concern the TDA method, our analysis was focused on finding differences in the topology, the evaluation of the differences was assigned through the distances, but graphically presented. To better take advantage of the method the differences would have to be implemented in a machine learning model, to encapsulate the strategic information needed to differentiate the datasets one from the others. The present work was an exploration of the TDA method, its applications and possibilities, therefore we didn't apply it to any model due to the lack of time.

Time is another limitation, the entire analysis and written work were developed more or less in two months of the MBA calendar. The TDA calculations per se are demanding for the calculating process. In our machine, the time spent per year was in the order of several hours (our machine specific: processor Intel Core i5 CPU, 2,20 GHz x4, RAM 8 GiB). Therefore the analysis focused on delivering results for the present work, the connected components were the most evident and impactful results.

The possible further analysis could have delivered more interesting results, as the application of the vector of the Persistence diagram into a model to better take advantage of the lower complexity of the information stored. The calculation returns a vector of 100 entries for every homological space (like H_0), and the simplification of the relation between the data points (more than 1000 entries times 11 columns) is reduced greatly. The connected components can be

utilized for cluster algorithms, for example, to infer the similarities' groups easily.

Another limitation was the graphical presentation of the many curves, the multiplicity of graphics tend to lose the interest and focus of the readers. How to evaluate the results and present them was also a difficulty and object of experimentation.

BIBLIOGRAPHY

ALI, D.; ASAAD, A.; JIMENEZ, M.-J.; NANDA, V.; PALUZO-HIDALGO, E.; SORIANO-TRIGUEROS, M. **A Survey of Vectorization Methods in Topological Data Analysis**. 2022. Citation on page 27.

____. A survey of vectorization methods in topological data analysis. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, p. 1–14, 2023. Citation on page 21.

AYESHA, S.; HANIF, M. K.; TALIB, R. Overview and comparative study of dimensionality reduction techniques for high dimensional data. **Information Fusion**, Elsevier BV, v. 59, p. 44–58, jul 2020. Citation on page 25.

Bazzo Vieira, J. P.; Vieira Braga, C. K.; PEREIRA, R. H. The impact of covid-19 on air passenger demand and co2 emissions in brazil. **Energy Policy**, v. 164, p. 112906, 2022. ISSN 0301-4215. Available: <https://www.sciencedirect.com/science/article/pii/S0301421522001318>. Citation on page 30.

BLEHER, M.; HAHN, L.; PATIñO-GALINDO, J.; CARRIèRE, M.; BAUER, U.; RABADAN, R.; OTT, A. Topological data analysis identifies emerging adaptive mutations in sars-cov-2. 02 2022. Citations on pages 26 and 27.

BUKKURI NOEMI ANDOR, I. K. D. A. Applications of topological data analysis in oncology. **Front Artif Intell.**, v. 4, n. 659037, 2021. Citations on pages 26 and 27.

BYRNE, C.; HORAK, D.; MOILANEN, K.; MABONA, A. Topic modeling with topological data analysis. In: **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. p. 11514–11533. Available: <https://aclanthology.org/2022.emnlp-main.792>. Citations on pages 26 and 27.

CARLSSON, G. Topology and data. **Bulletin of The American Mathematical Society - BULL AMER MATH SOC**, v. 46, p. 255–308, 04 2009. Citations on pages 20 and 25.

____. Topological pattern recognition for point cloud data. **Acta Numerica**, Cambridge University Press (CUP), v. 23, p. 289–368, may 2014. Citations on pages 20 and 26.

____. The shape of biomedical data. **Current Opinion in Systems Biology**, Elsevier BV, v. 1, p. 109–113, feb 2017. Citations on pages 26 and 27.

CARLSSON, G.; VEJDEMO-JOHANSSON, M. **Topological Data Analysis with Applications**. [S.l.]: Cambridge University Press, 2021. Citations on pages 20, 25, 26, and 27.

CARLSSON, G. E. Topological methods for data modelling. **Nature Reviews Physics**, v. 2, p. 697 – 708, 2020. Available: <https://api.semanticscholar.org/CorpusID:227160366>. Citation on page 19.

CHAZAL, F.; GLISSE, M.; LABRUèRE, C.; MICHEL, B. Optimal rates of convergence for persistence diagrams in topological data analysis. 2013. Citations on pages 13, 21, and 36.

CHAZAL, F.; MICHEL, B. An introduction to topological data analysis: fundamental and practical aspects for data scientists. In: . [S.l.: s.n.], 2021. Citations on pages 26 and 27.

DEY, T. K.; WANG, Y. **Computational Topology for Data Analysis**. [S.l.]: Cambridge University Press, 2022. Citation on page 25.

EDUARDO OLIVEIRA, A. V. M. de A. An econometric analysis for the determinants of flight speed in the air transport of passengers. v. 13, p. 2045–2322, 2023. Citation on page 30.

GABRIELSSON, R. B.; NELSON, B. J.; DWARAKNATH, A.; SKRABA, P. A topology layer for machine learning. In: CHIAPPA, S.; CALANDRA, R. (Ed.). **Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics**. PMLR, 2020. (Proceedings of Machine Learning Research, v. 108), p. 1553–1563. Available: <https://proceedings.mlr.press/v108/gabrielsson20a.html>. Citation on page 27.

GIDEA, M.; KATZ, Y. Topological data analysis of financial time series: Landscapes of crashes. **Physica A: Statistical Mechanics and its Applications**, v. 491, p. 820–834, 2018. ISSN 0378-4371. Available: <https://www.sciencedirect.com/science/article/pii/S0378437117309202>. Citations on pages 26 and 27.

GIUNTI, B.; LAZOVSKIS, J.; RIECK, B. **DONUT – Creation, Development, and Opportunities of a Database**. 2023. Citation on page 28.

HENSEL, F.; MOOR, M.; RIECK, B. A survey of topological machine learning methods. **Frontiers in Artificial Intelligence**, v. 4, 2021. ISSN 2624-8212. Available: <https://www.frontiersin.org/articles/10.3389/frai.2021.681108>. Citations on pages 21 and 27.

INFRAESTRUTURA, M. da. **Agência Nacional de Aviação Civil (ANAC)**. 2023. https://www.gov.br/anac/pt-br/assuntos/dados-e-estatisticas/dados-estatisticos. Accessed: 11/2023. Citation on page 30.

KARAN, A.; KAYGUN, A. Time series classification via topological data analysis. **Expert Systems with Applications**, v. 183, p. 115326, 2021. ISSN 0957-4174. Available: <https://www.sciencedirect.com/science/article/pii/S0957417421007557>. Citations on pages 26 and 27.

LEYKAM, D.; ANGELAKIS, D. G. Topological data analysis and machine learning. **ADVANCES IN PHYSICS: X**, v. 8, n. 1, 2023. Citations on pages 26 and 27.

LOUGHREY, C. F.; FITZPATRICK, P.; ORR, N.; JUREK-LOUGHREY, A. The topology of data: opportunities for cancer research. **Bioinformatics**, Oxford University Press (OUP), v. 37, n. 19, p. 3091–3098, jul 2021. Citations on pages 26 and 27.

MARIA, C.; BOISSONNAT, J.-D.; GLISSE, M.; YVINEC, M. The gudhi library: Simplicial complexes and persistent homology. In: HONG, H.; YAP, C. (Ed.). **Mathematical Software – ICMS 2014**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. p. 167–174. ISBN 978-3-662-44199-2. Citation on page 27.

OFORI-BOATENG, D.; LEE, H.; GORSKI, K.; GARAY, M.; GEL, Y. Application of topological data analysis to multi-resolution matching of aerosol optical depth maps. **Frontiers in Environmental Science**, v. 9, p. 684716, 06 2021. Citations on pages 26 and 27.

OTTER, N.; TILLMANN, U.; GRINDROD, P. A roadmap for the computation of persistent homology. **EPJ Data Science**, v. 6, n. 17, 9 2017. Citations on pages 26 and 28.

RABADAN, R.; BLUMBERG, A. J. **Topological Data Analysis for Genomics and Evolution: Topology in Biology**. [S.l.]: Cambridge University Press, 2019. Citations on pages 26 and 27.

RAVISHANKER, N.; CHEN, R. Topological data analysis (tda) for time series. In: . [S.l.: s.n.], 2019. Citations on pages 26 and 27.

RAY, J.; TROVATI, M. A survey of topological data analysis (tda) methods implemented in python. In: . [S.l.: s.n.], 2018. p. 594–600. ISBN 978-3-319-65635-9. Citation on page 28.

RIBEIRO, S.; SILVA, A.; DáTTILO, W.; REIS, A.; GóES-NETO, A.; ALCANTARA, L.; GIOVANETTI, M.; COURA-VITAL, W.; FERNANDES, G.; AZEVEDO, V. Severe airport sanitarian control could slow down the spreading of covid-19 pandemics in brazil. **PeerJ**, 06 2020. Citations on pages 13, 22, and 30.

RIECK, B.; SADLO, F.; LEITTE, H. Topological machine learning with persistence indicator functions. In: ____. **Topological Methods in Data Analysis and Visualization V**. Springer International Publishing, 2020. p. 87–101. ISBN 9783030430368. Available: <http://dx.doi.org/10.1007/978-3-030-43036-8_6>. Citations on pages 21 and 36.

SEVERSKY, L. M.; DAVIS, S.; BERGER, M. On time-series topological data analysis: New data and opportunities. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**. [S.l.: s.n.], 2016. p. 1014–1022. Citations on pages 26 and 27.

SHEEHY, D. R. [acm press annual symposium - kyoto, japan (2014.06.08-2014.06.11)] annual symposium on computational geometry - socg'14 - the persistent homology of distance functions under random projection. ACM Press, 2014. Citation on page 21.

SINGH, Y.; FARRELLY, C.; HATHAWAY, Q. e. a. Topological data analysis in medical imaging: current state of the art. **Insights Imaging**, v. 14, n. 58, 2023. Citations on pages 26 and 27.

SIZEMORE, A. E.; PHILLIPS-CREMINS, J. E.; GHRIST, R.; BASSETT, D. S. The importance of the whole: Topological data analysis for the network neuroscientist. **Network Neuroscience**, v. 3, n. 3, p. 656–673, 07 2019. Citations on pages 21, 26, and 27.

SKAF, Y.; LAUBENBACHER, R. Topological data analysis in biomedicine: A review. **Journal of Biomedical Informatics**, v. 130, p. 104082, 2022. ISSN 1532-0464. Available: <https://www.sciencedirect.com/science/article/pii/S1532046422000983>. Citations on pages 26 and 27.

TAUZIN, G.; LUPO, U.; TUNSTALL, L.; PéREZ, J. B.; CAORSI, M.; MEDINA-MARDONES, A. M.; DASSATTI, A.; HESS, K. Giotto-tda: A topological data analysis toolkit for machine learning and data exploration. **J. Mach. Learn. Res.**, JMLR.org, v. 22, n. 1, jan 2021. ISSN 1532-4435. Available: <https://github.com/giotto-ai/giotto-tda.> Citations on pages 27, 36, and 46.

TIERNY, J.; FAVELIER, G.; LEVINE, J. A.; GUEUNET, C.; MICHAUX, M. The topology toolkit. **IEEE Transactions on Visualization and Computer Graphics**, v. 24, n. 1, p. 832–842, 2018. Citation on page 21.

UMEDA, Y. Time series classification via topological data analysis. In: . [S.l.: s.n.], 2017. Citations on pages 26 and 27.

WASSERMAN, L. Topological data analysis. **Annual Review of Statistics and Its Application**, v. 5, n. 1, p. 501–532, 2018. Citations on pages 13, 20, and 21.

YEN, P.; CHEONG, S. A. Using topological data analysis (tda) and persistent homology to analyze the stock markets in singapore and taiwan. **Frontiers in Physics**, v. 9, p. 572216, 03 2021. Citations on pages 26 and 27.

ZOMORODIAN, A. Fast construction of the vietoris-rips complex. **Computers & Graphics**, v. 34, n. 3, p. 263–271, 2010. ISSN 0097-8493. Shape Modelling International (SMI) Conference 2010. Available: <https://www.sciencedirect.com/science/article/pii/S0097849310000464>. Citation on page 21.


ICMC
USP
SÃO CARLOS