POLITECNICO DI TORINO

Dipartimento di Matematica
Mathematical Engineering

Master Thesis

Topological informed optimization of car sharing resources allocation

Supervisors:
Prof. Francesco Vaccarino
Dr. Alessandro De Gregorio
Dr. Luca Vassio
Dr. Alessandro Ciociola

Candidate:
Ilaria Scagno

March 2021

Contents

Abstract

The goal of this thesis is analysing the mobility of the car sharing vehicles over the city of Turin using Topological Data Analysis techniques. The current infrastructure over the city is mainly based on Internal Combustion Engine (ICE) vehicles, for this reason the objective is providing useful information in order to design an electric Free-Floating Car-Sharing system. Specifically, since the main problem this infrastructure needs to face is the optimal placement of charging spots, the desired information consists of zones of the city in which the bookings concentrate the most.

A first exploratory phase consists of analysing the data in order to extract mobility patterns in different days of the week or hours of the day. To begin, five different daily time slots have been individuated by considering the hours of the day that have a homogeneous number of bookings. This has been used to group the bookings per (weekday, time slot). Each group individuates a discrete probability distribution, and hence to obtain a notion of closeness between them the Wasserstein distance can be used. Then a hierarchical clustering algorithm has been applied and it revealed similarities between weekdays in the same time slot. It has also shown that Saturdays and Sundays are the days that differ the most with respect to the others.

Then, before proceeding with the identification of the zones of interest, it has been important understanding if a topological approach identified the same similarities. For this purpose the hierarchical clustering algorithm has been repeated after having built on each of the aforesaid groupings of the data two filtrations of simplicial complexes that are basic representations of a topological space: the Vietoris-Rips and the Alpha complexes. The result confirmed the first insight: the data shows a relevant pattern within the same time slot.

Finally, the last step of the thesis consists of individuating the relevant zones of the city. Such areas are the ones in which is observed a higher number of bookings compared to the rest of the city. Those zones are the ones in which makes more sense the installation of charging poles. From a topological point of view the zones of interest are represented by cycles on the plane. The choice of those cycles is not trivial, but this problem has been solved through the concept of tight cycles that are the ones that individuate the "holes" in a planar simplicial complex in the most accurate way.

To understand the quality of the zones extracted, a validation procedure is done by dividing the data set into train and test sets and checking if the zones individuated in the train phase are densely populated by the booking events of the test set. Results are overall good, in particular for the time slots with more events. They get worse specially in the night time slot due to a much lower amount of data. An interesting aspect noticed is that predictions are more accurate for the departures compared to the arrivals.

Chapter 1

Introduction

Free-Floating Car-Sharing (FFCS) systems have become a popular mobility solution in the past years. In those systems the user is allowed to book a car through his phone and return it anywhere in the designed operational area. In the early years of development of this type of service, the cars provided were almost exclusively based on Internal Combustion Engine (ICE). However, soon it was understood that for sustainability reasons it was needed a partial, or even total, conversion to electric engines. This new type of infrastructure has many challenges, mainly due to the different (lower) autonomy of the cars and hence the need to charge the vehicles. The charging policy is the aspect mostly investigated in many studies ([2], [4], [5], [10]). The possible solutions are mainly of two types: either placing a centralized charging spot in a zone of the city with extremely high utilization rate or constructing a distributed infrastructure with many charging poles around the city. Both the policies have pros and cons that need to be investigated, however the second solution proposes a non trivial problem: the optimal placement of the poles. The objective of this thesis is analysing the mobility of the car sharing vehicles over Turin through Topological Data Analysis in order to identify good candidate locations for such charging spots.

Topological Data Analysis consists of multiple techniques based on algebraic topology whose objective is identifying complex geometric structures, "holes", in the data [3]. Such information is extracted by building on top of the data specific structures called simplicial complexes, basic topological entities. The most important concept of Topological Data Analysis is Homology. It provides a powerful tool that allows to formalize topological features of a simplicial complex in an algebraic way. For any dimension k the "holes" of dimension k are represented by the k-th homology group H_k. Intuitively H_0 counts the number of connected components of the complex, H_1 the loops, H_2 the voids and so on. The zones of interest extracted in this work are loops on the plane, hence in this thesis the attention focuses on the analysis of the first homology group H_1. Another central concept is Persistent Homology. It allows to study the evolution of homological features of families

of nested simplicial complexes, called filtrations. This is done by keeping track of the instants of birth and death of the "holes" in the data.

The goal of the study is to bring back the information given by homology on the starting data, e.g. [19]. To reach this objective, over the past few years several techniques have been proposed. They all have as common aspect the determination of a privileged basis for homology groups, with a good representative in each homology class. For example, minimal homology bases proposed by Dey et al. [7] are used to find specific basis for the first homology group H_1, whereas in [17] are described volume optimal cycles, a generalization to higher order homology groups.

To begin, a first phase consists of analysing the data to find useful information such as mobility patterns between days or hours of the day. This is done, firstly partitioning the day into five time slots in each of which the number of bookings is homogeneous, secondly dividing the data into 35 groups given by (weekday, time slot). In this way each group identifies a discrete probability distribution given by sum of Dirac delta functions and hence a notion of closeness between them is given by the Wasserstein distance, also called Earth mover's distance. Such distance is used to apply a hierarchical clustering algorithm that revealed relevant similarities between weekdays in the same time slot.

In a second phase, the same clustering algorithm is applied in a topological context. Two types of simplicial complexes are built on the data grouped as before: the Vietoris-Rips and the Alpha complexes. The first is the elementary simplicial complex based on the pairwise intersections of closed balls of a given radius centered on the data points given by the longitude and latitude coordinates of the bookings. The second represents a slight improvement by considering as elementary cells the intersection of the closed balls with the Voronoi cell of a given radius. The algorithm applied to the data with those two structures confirmed the first insights: the data follows a hourly pattern. This motivates going deeper in the topological analysis.

Finally, through the homological scaffold and the tight cycles the zones of interest have been extracted. Such zones are identified by the tight representatives of the first homology group H_1. The last step consists of testing the results through a validation approach in which at each step a train set is used to extract the zones of interest and a test set is used to see if the most relevant areas are more "densely populated" by the events of such set.

The thesis is structured as follows: Chapter 2 is devoted to the description of the state-of-the-art works about Topological Data Analysis and the description of the FFCS systems, in Chapter 3 are performed preliminary analysis on the data set, Chapter 4 describes the procedure adopted for the detection of the hourly pattern in the data, Chapter 5 is devoted to the theoretical part of Topological Data Analysis and Chapter 6 presents the results obtained and the validation procedure.

Chapter 2

Related work

Free-Floating Car-Sharing (FFCS) systems have become a popular mobility solution in the past years, however in some cities those services have already reached saturation [10]. The most appealing solution for the service providers, for sustainability reasons, is the conversion from combustion engine fleets to electric ones. This type of FFCS system, however, cannot be applied to every city because of the different car autonomy and because of the different service utilization among the cities. For this reason some preliminary studies need to be performed in order to understand how the service utilization differs from one city to another. To this aim in [10] has been analyzed the service provision in 23 cities across the world. These have been compared under different aspects such as fleet size, operating area, number of bookings, rental distance and duration and car daily usage. Results showed that FFCS system are mainly used for short one-way trips. The city with the highest daily car utilization is Madrid, whose fleet is entirely electric and can be taken as an example of the effectiveness of the new sustainable mobility. On the other hand some cities in the USA such as Columbus and Austin tend to have underused cars and for this reason are not good candidates. Indeed in general the North American cities tend to have a lower utilization rate and this makes them less proper cities for an electric FFCS system.

Finally in case of an electric fleet it is of crucial importance understanding where to locate the charging spots. For this reason a spatial analysis has been performed. To this aim the area of each city has been divided into $500m \times 500m$ squares. The results showed that in some cities rentals are homogeneously spread (New York City and Amsterdam), whereas in other cities (Milan and Vancouver) rentals are concentrated only in few zones. A more detailed spatial analysis has been made dividing the day into time slots in order to understand if a zone is attractive (in this zone start the rentals) or generative (in this zone end the rentals) and at which time. As expected most of the zones that have been identified as attractive in the morning, have become generative in the afternoon. Generative and attractive zones have to be taken in consideration to place charging spots in.

Successively in [2], [4] and [5] the studies focused on the city of Turin. The central point of those articles has been analyzing the mobility in order to individuate the best solution for an electric FFCS system. The data is based on internal combustion engine cars and not on electric ones, but this still captures the actual usage patterns of regular customers. Specifically in [5] are faced the problems of finding the optimal placement of charging stations and of designing the best car return policy. In [4] two different car charging infrastructures are compared evaluating their performance and management costs: a centralized charging hub in a dynamic zone of the city and a distributed set of charging poles around the most used zones of the city. The results show that a distributed infrastructure with some user's help is by far the best solution compared to an optimally placed centralized infrastructure. Finally in [2] the attention focused on the scalability of a Free Floating electric Car-Sharing system with an increase in the intensity of the demand. The three articles have in common the subdivision of the city into squares $500m \times 500m$.

The main problem an electric FFCS system has to face is the reduced autonomy of the cars, hence the system must be equipped with an efficient charging network that minimizes the user's discomfort (i.e. allows the user to park the car close to his destination) and at the same time minimizes the number of infeasible trips (i.e. those trips that are not possible because of the low battery level).

The results obtained in [5] are surprising: equipping just the 5% of the city zones with charging poles is enough to make all trips feasible. This is possible only with the assumption that users cooperate, meaning that when the battery level is low the user returns the car in a charging spot. However, even with this constraint, it would happen rarely and thus the discomfort would be minimum.

The study in [2] about scalability pointed out that the distributed system has a useful economy of scale: the fleet size shall increase sublinearly with respect to the mobility demand intensity.

Thanks to those studies it is now clear that the city of Turin satisfies the requirements needed to develop an electric FFCS system. The challenge is studying the mobility patterns in order to understand which are the best locations for the charging spots.

Up to now the strategy adopted has been dividing the city into $500m \times 500m$ squares obtaining 261 zones. This approach can be limiting because the zone of interest may not be a square, or it can be an area in between of multiple squares. For this reason it is now left aside this subdivision and instead the locations of the trips are considered as points on the plane to which are applied techniques of Topological Data Analysis to extract the zones of greater interest, that can have any shape.

Over the past years Topological Data Analysis has been used in many contexts to give a new perspective based on the topological structure of the data. In the work of Umeda and al. [25], TDA has been in an effective way used to detect internal damage in bridges at an early stage through a Time-Series analysis of signals retrieved by sensors applied on the surface of bridge decks. Many

recent works such as [14], [15] and [24] used TDA in the medical field. Specifically in [24] it was used to detect arrhythmia by analyzing ECG signals. In the article, TDA was used in a classification task, whose aim was detecting and classifying anomalies in the heartbeats. The method adopted proved performing as well as the state-of-the art techniques. In the work of Nicponski et al. [14] a topological approach has been applied for the diagnosis of vascular diseases, primary cause of human mortality worldwide. Specifically persistent homology is applied to a geometric representation of vessels boundaries that allows individuating the stenosis in the vessels. The results showed that persistent homology detected relevant differences in the barcodes between vessels with high and low stenosis. Finally [15] used an alternative representation of the persistent homology information, the persistence landscape, to study a particular protein, the maltose-binding protein. The analysis showed an application of TDA techniques in the three-dimensional space and highlighted that such approach proved to be effective in the individuation of structural changes in the protein bindings.

In [19] and [11] is introduced the homological scaffold, a technique that, applied to a weighted network, allows to summarize the topological information into an attractive network representation. Specifically [19] has applied this tool to compare the characteristics of functional brain networks in 15 subjects after intravenous infusion of placebo and psilocybin,a psychoactive component. The results show that the structure of the homological scaffold in the two cases changes dramatically and this proves the effectiveness of the method. Instead [11] has focused on the main problem of homological scaffold: the arbitrariness in the choice of the representatives of the homology classes. The paper solves this problem using, for the first homology group, the minimal scaffold. It consists of taking as representative for each homology class the cycle of minimal length.

In this thesis a new method for the choice of the representative cycles is investigated: the tight representatives.

Chapter 3

Data exploration

In this chapter is described the first phase of any work that involves analyzing some data coming from any source: exploring the data and its attributes. In section 3.1 is described the Kernel Density Estimation, a nonparametric technique that, given some data estimates the probability distribution it is drawn from. Section 3.2 is devoted to some descriptive plots and analysis useful to understand the structure of the most important attributes of the data set as well as the magnitude of the data, to be taken into consideration for computational reasons.

The data set consists of the trips registered in October 2017 in Turin by Car2Go. Specifically the company detected 103185 trips, each of which has the following attributes:

- *Plate*: plate of the booked car
- *Start_time*, *End_time*: date and time at which the trip begins and ends respectively
- *Start_longitude*, *Start_latitude*: trip's starting point's coordinated
- *End_longitude*, *End_latitude*: trip's ending point's coordinates
- *Euclidean_distance*: trip's distance in meters
- *Duration*: trip's duration in seconds

- *Year*, *Month*: year and month at which the trip occurs
- *Start_hour*, *End_hour*: trip's starting and ending hour
- *Start_weekday*, *End_weekday*: day of the week at which the trips begins and ends respectively
- *Start_daytype*, *End_daytype*: whether the trip's start and end day are weekdays or weekend days.

First some descriptive analysis have been performed. To this aim first it is necessary introducing a nonparametric approach that will be used in this chapter: the kernel density estimation, described in detail in [23].

3.1 Kernel Density Estimation

Given a sample of points $X_1, \dots, X_n$ in $\mathbb{R}^m$ the goal is estimating the distribution from which it is generated. We denote with $\hat{f}_n$ the estimator.

Definition A *kernel* is any smooth function $K : \mathbb{R}^m \longrightarrow \mathbb{R}$ such that:

$$K(x) \geq 0, \quad \forall x \in \mathbb{R}^n$$

$$\int_{\mathbb{R}^m} K(x)dx = 1$$

$$\int_{\mathbb{R}^m} xK(x)dx = 0$$

$$\sigma_K^2 = \int_{\mathbb{R}^m} x^T xK(x)dx > 0.$$

The most famous one is the Gaussian kernel: $K(x) = (2\pi)^{-\frac{m}{2}} e^{x^T x}$.

Definition Given a kernel K and symmetric and positive definite matrix H called *bandwidth*, the *kernel density estimator* is defined as follows:

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{det(H)^{\frac{1}{2}}} K\Big(H^{-\frac{1}{2}}(x - X_i)\Big).$$

The quality of an estimator is evaluated through the integrated mean squared error, or risk $R = \mathbb{E}(L)$, where:

$$L = \int_{\mathbb{R}} (\hat{f}_n(x) - f(x))^2 dx$$

Since the estimators usually depend on a smoothing parameter H that is to be chosen to minimize an estimate of the risk, so let's rewrite L as a function of H:

$$\begin{aligned} L(H) &= \int (\hat{f}_n(x) - f(x))^2 dx \\ &= \int \hat{f}_n^2(x)dx - 2\int \hat{f}_n(x)f(x)dx + \int f^2(x)dx. \end{aligned}$$

Since the last term does not depend on H it is equivalent to minimizing $\mathbb{E}(J(H))$, where:

$$J(H) = \int \hat{f}_n^2(x)dx - 2\int \hat{f}_n(x)f(x)dx.$$

Definition The *cross-validation estimator of risk* is defined as:

$$\hat{J}(H) = \int \left(\hat{f}_n(x)\right)^2 - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{(-i)}(X_i)$$

where $\hat{f}_{(-i)}$ is the density estimator obtained after removing the i^{th} observation.

3.2 Descriptive Analysis

The data set has no missing values, but there are instead many trips with *Euclidean_distance* equal to zero. This can happen for example when a car is booked by an user but the booking later is cancelled. In such case Car2Go registers a trip but in reality the car did not move. Since the analysis focuses on mobility patterns in the city of Turin, those instances have been removed.

The average trip duration is 36 minutes, its average distance is 2.8 km and the car fleet over Turin is composed of 414 cars. In Figure 1 the distributions of these two variables are shown in more detail through violin plots. A violin plot, unlike a box plot in which all of the plot components correspond to actual data points, features a kernel density estimation using a Gaussian kernel of the underlying distribution. Moreover the violin plots show the distribution's extreme values and its mean.

Since Turin is not a very big city, the trips with a high distance run shown in the left side of Figure 1 probably correspond to trips to the Airport and bookings that lasted more than one day. The same consideration can be made about the duration. In fact the corresponding violin plot shows that most of the trips are quite short, but the maximum value is very far from the mean. This can correspond again to trips that lasted multiple days.

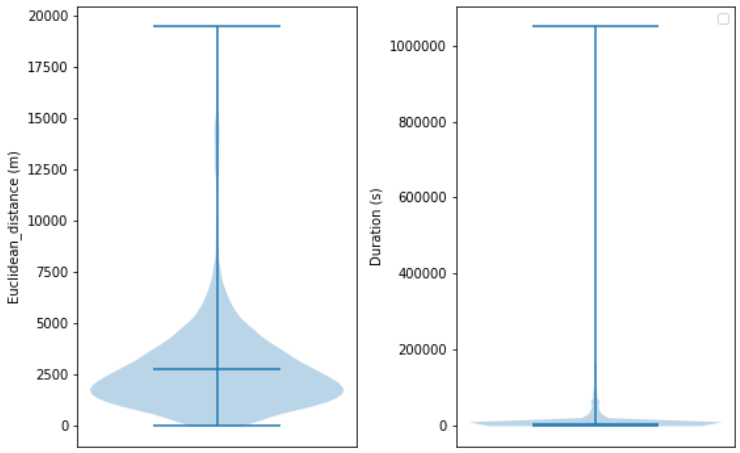


Figure 1: Violin plots for *Euclidean_distance* and *Duration*.

The data has then been divided per weekday in order to count the number of trips registered in each day. In Figure 2 is clear that the number of trips is quite homogeneous within the same weekday: each day has approximately 2500 trips. The only exception are Sundays in which there are at most 2000 trips per day.

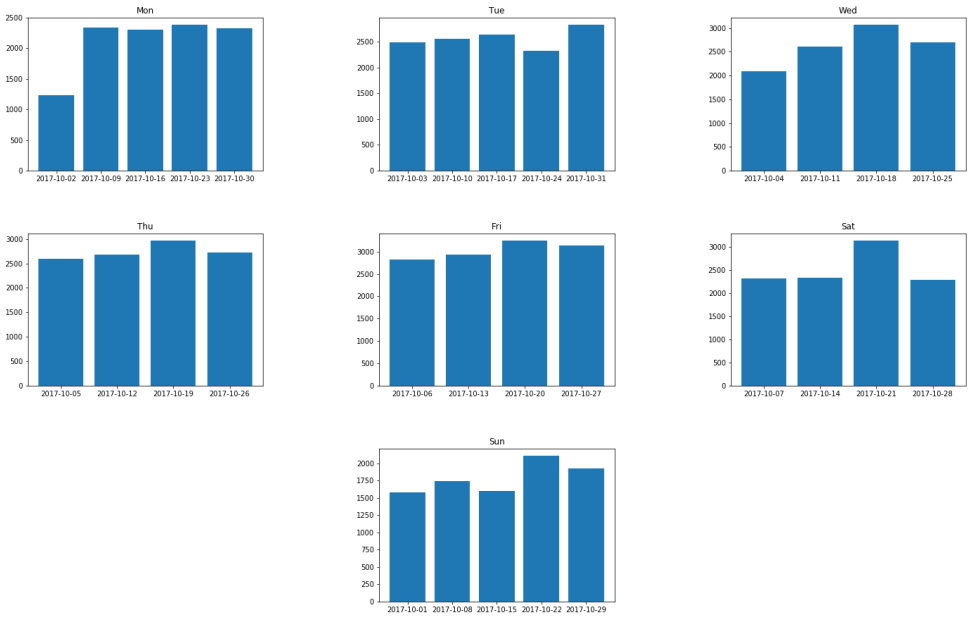


Figure 2: Number of trips per day, divided by weekday.

A similar analysis, shown in Figure 3, has been performed counting the total number of trips per pair (weekday, hour). As expected during working days there are two peaks: the first between 7 a.m. and 8 a.m., time at which people usually goes to work, and the second between 17 p.m. and 19 p.m., time at which people come back home. Moreover during weekdays between midnight and 5 a.m. there are very few trips. Weekends look different, for example between midnight and 2a.m. there are far more trips due to the nightlife.

Figure 2 and Figure 3 have been obtained considering just the starting attributes (i.e. *Start_weekday* and *Start_hour*). Analogous plots can be obtained with the ending attributes but they would not change much because, since the median of trip duration is 20 minutes, most of the trips that start in an hour end in the same hour.

Finally for each pair (weekday, hour) the trip's starting and ending coordinates have been considered in order to visualize how the trips are distributed over the city. In fact the goal of the thesis is to detect the zones in which the trips concentrate the most so as to eventually place in such zones charging spots as well as replacing an amount of cars that satisfies the demand. The starting and ending coordinates have been used to obtain kernel density estimation plots by using a Gaussian Kernel with bandwidth $\sigma = 0.2I$, where I is the identity matrix. In Figure 4 are shown such plots for the pair (Saturday, 6 p.m.).

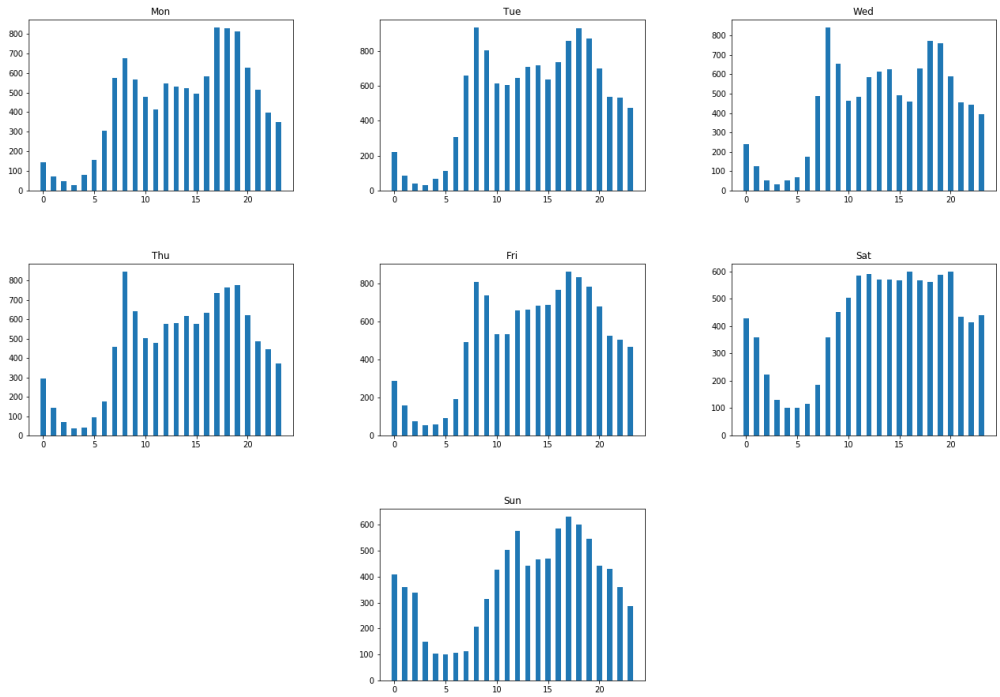


Figure 3: Number of trips per pair (weekday, hour).

Even though is difficult to compare the two plots, one aspect is particularly evident: in this day and at this hour there are flights arriving at the airport (the red isolated point on the right plot of Figure 4), in fact some cars are departing from there. On the other hand there are probably no flights, or very few, departing from the airport since there are no cars arriving to it.

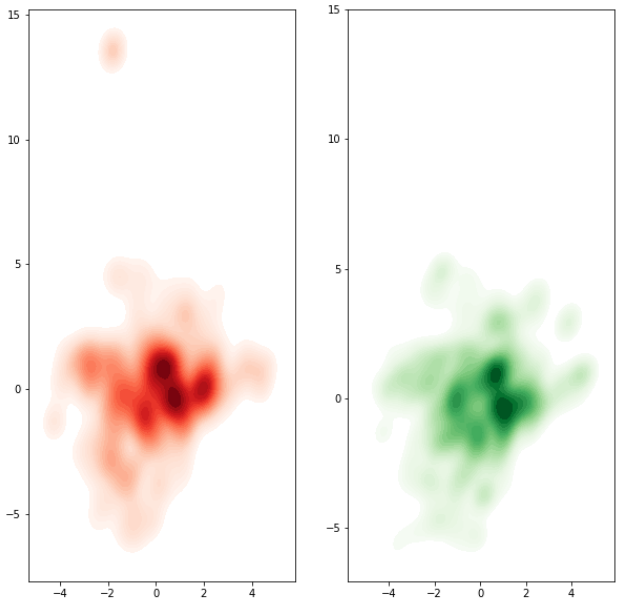


Figure 4: Start (left) and End (right) KDE plots for the pair (Saturday, 6 p.m.).

Chapter 4

Mobility patterns

In this chapter is described the procedure adopted to look for useful patterns in the data. Specifically the goal is to apply the hierarchical clustering technique to see which days of the week or which hours of the day are identified as most similar.

Section 4.1 is devoted to the description of the Equirectangular projection, that for computational reasons will be needed to extract useful results. In section 4.2 is defined the Wasserstein distance, that will be used to compute the similarity between weekdays and hours of the day. In section 4.3 is described the Hierarchical clustering and finally in section 4.4 is showed how the concepts introduced in the previous sections have been applied to obtained the desired results.

4.1 Equirectangular projection

The equirectangular projection, also called equidistant cylindrical projection, is one of the simplest map projections. It is defined in [22] and maps both meridians and circles of latitude to straight lines of constant spacing on the plane to form a grid (Figure 5). The forward projection transforms spherical coordinates to planar ones, while the reverse projection maps planar coordinates into spherical ones.

This projection is usually adopted only with maps covering small areas because if there is no distortion at the specified standard parallel, it increases the more we move north or south from this parallel.

Define the following variables:

- (λ, φ): longitude and latitude coordinates of the point to project
- φ_1: latitude at which there is no distortion
- (λ_0, φ_0): longitude and latitude coordinates of the central point of the map
- R: Earth radius
- (x, y): planar coordinates of the projected point

where longitudes and latitudes are defined in terms of radians. The equirectangular projection in the forward case is the following:

$$x = R(\lambda - \lambda_0)\cos\varphi_1$$
$$y = R(\varphi - \varphi_0)$$

the reverse projection instead is:

$$\lambda = \frac{x}{R\cos\varphi_1} + \lambda_0$$
$$\varphi = \frac{y}{R} + \varphi_0$$

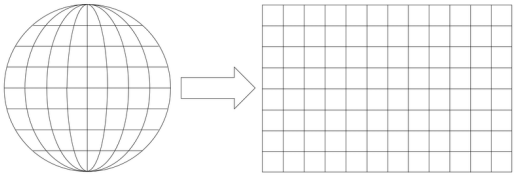


Figure 5: Equirectangular projection.

4.2 Wasserstein distance

In this section the Wasserstein distance is described, or Earth Mover's distance, that is used to compute the distance between two probability distributions. Its definition is based on an optimization problem called Optimal Transport (OT) problem. The OT problem aims to find the most efficient way to move mass between distributions at the minimum cost.

To introduce those concepts some theoretical results are needed. For the concepts of Borel sets and Borel probability measures the source used is [18]. For what concerns tightness and transport measures Chapter 5 of the book by Ambrosio [1] has been taken as reference. Finally for the definitions of the OT problem and Wasserstein distance Chapters 6 and 7 again of [1] have been used.

Consider a Banach space X. Recall that an *algebra* on a set X is a a collection $\mathcal{A}$ of subsets of X such that:

- $X \in \mathcal{A}$
- if $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$
- if A, $B \in \mathcal{A}$ then $A \cup B \in \mathcal{A}$.

Definition A *σ-algebra* on X is a nonempty collection $\mathcal{A}$ of subsets of X such that:

- if $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$
- let $\{A_n\}_{n\geq 1}$ be a numerable family of elements of $\mathcal{A}$ then $\bigcup_{n\geq 1} A_n \in \mathcal{A}$.

Definition The *Borel σ-algebra* $\mathcal{B} = \mathcal{B}(X)$ is the smallest σ-algebra that contains all open subsets of X. Its elements are called *Borel sets*.

Definition A metric space (X, d) is called *separable* if it has a countable dense subset, i.e. there are $x_1, x_2, \dots$ in X such that $\overline{\{x_1, x_2, \dots\}} = X$.

Definition Let (X, d) be a metric space. A *finite Borel measure* on X is a map $\mu : \mathcal{B}(X) \longrightarrow [0, +\infty)$ such that:

- $\mu(\emptyset) = 0$
- if $B_1, B_2, \cdots \in \mathcal{B}$ are mutually disjoint, then $\mu\big(\bigcup_{n\geq 1} B_i\big) = \sum_{n\geq 1} \mu(B_i)$.

μ is called *Borel probability measure* if additionally $\mu(X) = 1$. We denote by $\mathcal{P}(X)$ the family of all Borel probability measures on X.

Definition Consider the Borel σ-algebra $\mathcal{B}(X)$ and a Borel set $A \subseteq X$. Then for a given $x \in X$ the *Dirac probability measure* on A is defined as:

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

Definition A finite Borel measure μ on X is called *tight* if for every $\epsilon > 0$ there exists a compact set $K \subset X$ such that $\mu(X \backslash K) < \epsilon$.

Definition A separable metric space X is a *Radon space* if every Borel probability measure $\mu \in \mathcal{P}(X)$ satisfies:

$$\forall B \in \mathcal{B}(X), \;\; \epsilon > 0 \;\; \exists K_\epsilon \Subset B \;\; s.t. \;\; \mu(B \backslash K_\epsilon) \leq \epsilon, \tag{4.2}$$

where $K_\epsilon \Subset B$ means that the closure of K_ϵ is a compact of B.

Defintion Let X_1 and X_2 be two separable metric spaces, $\mu \in \mathcal{P}(X_1)$ a Borel probability measure on X and $r : X_1 \longrightarrow X_2$ a Borel map. We define the *push-forward of μ through r*, denoted $r_\# \mu \in \mathcal{P}(X_2)$:

$$r_\# \mu(B) = \mu(r^{-1}(B)) \;\; \forall B \in \mathcal{B}(X_2).$$

Definition For an integer $N \geq 2$ and $i, j = 1, \ldots, N$ we define the *projection operators* on the product space $X = X_1 \times \cdots \times X_N$:

$$\pi^i : (x_1, \ldots, x_N) \mapsto x_i \in X_i,$$
$$\pi^{i,j} : (x_1, \ldots, x_N) \mapsto (x_i, x_j) \in X_i \times X_j.$$

Definition If $\mu \in \mathcal{P}(X)$, the *marginals of μ* are the probability measures:

$$\mu^i = \pi^i_\# \mu \in \mathcal{P}(X_i),$$
$$\mu^{i,j} = \pi^{i,j}_\# \mu \in \mathcal{P}(X_i \times X_j).$$

Definition Let $\mu^i \in \mathcal{P}(X_i)$ be the i^{th} marginal of μ, $i = 1, \ldots, N$. The *class of multiple plans* with marginals μ^i is defined by:

$$\Gamma(\mu^1, \ldots, \mu^N) = \Big\{ \mu \in \mathcal{P}(X_1 \times \cdots \times X_N) \;\; : \;\; \pi^i_\# \mu = \mu^i, \; i = 1, \ldots, N \Big\}.$$

If $N = 2$ a measure $\mu \in \Gamma(\mu^1, \mu^2)$ is called *transport plan between μ^1 and μ^2*. To each pair of measures $\mu^1 \in \mathcal{P}(X_1)$ and $\mu^2 = r_\# \mu^1 \in \mathcal{P}(X_2)$ connected by a Borel transport map $r : X_1 \longrightarrow X_2$ we can associate the transport plan

$$\mu = (i \times r)_\# \mu^1 \in \Gamma(\mu^1, \mu^2), \quad i \text{ identity map on } X_1. \tag{4.3}$$

If μ is representable as above we say that μ *is induced* by r.

After giving all the necessary information it is finally possible introducing the Optimal Transportation Problem.

Definition Let X, Y be two Radon spaces and let $c : X \times Y \longrightarrow [0, +\infty]$ be a Borel cost function. Given two Borel probability measures $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(X)$ the *optimal transport problem* is given by:

$$\min \Big\{ \int_{X\times Y} c(x,y) d\gamma(x,y) \ : \ \gamma \in \Gamma(\mu,\nu) \Big\}. \tag{4.4}$$

Definition A transport plan $\Gamma \subset X \times Y$ is said *c-monotone* if:

$$\sum_{i=1}^{n} c(x_i, y_{\sigma(i)}) \geq \sum_{i=1}^{n} c(x_i, y_i)$$

whenever $(x_1, y_1), \ldots, (x_n, y_n) \in \Gamma$ and σ is a permutation of $\{1, \ldots, N\}$.

Definition Let λ be a probability measure. We say that λ is *concentrated* on A if for some set $A \in \mathcal{B}$ we have that $\lambda(E) = \lambda(A \cap E), \quad \forall E \in \mathcal{B}$, where $\mathcal{B}$ is some σ-algebra.

Thanks to this definition is now possible to introduce the following theorem giving necessary and sufficient optimality condition for the optimal transport plan.

Theorem(Necessary and sufficient optimality conditions) (Necessity) If $\gamma \in \Gamma(\mu,\nu)$ is optimal and $\int_{X\times Y} c d\gamma < +\infty$, then γ is concentrated on a c-monotone Borel subset of $X \times Y$. Moreover, if c is continuous, then suppγ is c-monotone.
(Sufficiency) Assume that c is real-valued,$\gamma \in \Gamma(\mu,\nu)$ is concentrated on a c-monotone Borel subset of $X \times Y$, and

$$\mu\left(\left\{ x \in X \ : \ \int_Y c(x,y) d\nu(y) < +\infty \right\} \right) > 0 \tag{4.5}$$

$$\nu\left(\left\{ y \in Y \ : \ \int_Y c(x,y) d\mu(x) < +\infty \right\} \right) > 0. \tag{4.6}$$

Then γ is optimal, $\int_{X\times Y} c d\gamma < +\infty$.

Definition A measure $\mu \in \mathcal{P}(X)$ is *regular* if $\mu(B) = 0$ for any null set B. We denote by $\mathcal{P}^r(X)$ the class of regular measures.

Theorem (Optimal transport maps in $\mathbb{R}^d$) Assume that $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, $c(x,y) = h(x-y)$ with $h : \mathbb{R}^d \longrightarrow [0, +\infty)$ strictly convex, and the minimum of the optimal transport problem is finite. Then if μ, ν satisfy (4.5), (4.6) and $\mu \in \mathcal{P}^r(\mathbb{R}^d)$, then the optimal transport problem (4.4) has a unique solution μ and this solution is induced by an optimal transport i.e. there exists a Borel map $r : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ such that (4.3) holds.

Definition We denote with $\mathcal{P}_p(X)$ the subset of measures in $\mathcal{P}(X)$ with finite p-moment:

$$\mathcal{P}_p(X) = \left\{ \mu \in \mathcal{P}(X) \ : \ \int_X d(x,\bar{x})^p d\mu(x) < +\infty \quad \text{for some} \quad \bar{x} \in X \right\}.$$

Finally the Wasserstein distance can be defined.

Definition Let X be a separable metric space satisfying the Radon property (4.2) and let $p \geq 1$. The p^{th} *Wasserstein distance* between two probability measures $\mu^1,\ \mu^2 \in \mathcal{P}_p(X)$ is defined as follows:

$$W_p^p(\mu^1, \mu^2) = min\Big\{ \int_{X^2} d(x_1, x_2)^p d_\mu(x_1, x_2) \ : \ \ \mu \in \Gamma(\mu^1, \mu^2) \Big\}$$

In general this problem is very hard to solve, but is easier for discrete distributions. Let us consider for example two discrete probability distributions μ^1 and μ^2 defined respectively over the metric spaces X and Y as follows:

$$\mu^1 = \sum_{i=1}^{m} a_i \delta_{x_i},$$

$$\mu^2 = \sum_{j=1}^{n} b_j \delta_{y_j}$$

where the coefficients are such that $\sum_i a_i = \sum_j b_j = 1$.

The objective is finding the joint distribution $\gamma : X \times Y \longrightarrow \mathbb{R}$ with marginal distributions μ^1 and μ^2 defined as

$$\gamma = \sum_{i,j} \gamma_{ij} \delta_{x_i} \delta_{y_j}.$$

In such case, with $X = \mathbb{R}^m$ and $Y = \mathbb{R}^n$, the problem can be formulated through the following optimization problem:

$$\begin{aligned} \min_{\gamma \in \mathbb{R}_+^{m \times n}} \quad & \sum_{i,j} \gamma_{ij} M_{ij} \\ \text{s.t.} \quad & \gamma \mathbb{1}_n = a, \\ & \gamma^T \mathbb{1}_m = b, \\ & \gamma \geq 0 \end{aligned}$$

where $M \in \mathbb{R}_+^{m \times n}$ is the cost matrix defined as $m_{ij} = c(x_i, y_j) = d^p(x_i, y_j)$.

4.3 Hierarchical clustering

Clustering refers to a very broad set of unsupervised techniques for finding subgroups or clusters in a data set such that the elements in the same group are more similar to each other than to the other elements in different groups. In this section a specific clustering technique is described: the hierarchical clustering.

Hierarchical clustering, as explained in [12], is a method of cluster analysis which seeks to build a hierarchy of clusters. With respect to the well known K-means, this approach offers two advantages: first it does not require to specify the number k of clusters we want to obtain, second it provides

an attractive tree-based representation called dendrogram. The most common type of hierarchical clustering is the *bottom-up* or *agglomerative* clustering in which the dendrogram is built starting from the leaves and combining clusters up to it roots.

In a general setting, the hierarchical clustering algorithm works as follows. Given n points and defined a distance d between them, a $n \times n$ symmetric matrix D keeps track of the distances $d(i, j)$ between the points. Denote with $d((g), (k))$ the distance between the clusters (g) and (k). Finally denote with $L(f)$ the level of the f^{th} cluster. The algorithm is the following:

1. At the beginning all points are disjoint, hence $L(0) = 0$ and initialize a counter $k = 0$

2. For each pair of clusters $(i), (j)$ find the most similar ones $(a), (b)$ according to: $d((a), (b)) = \text{FORMULA}_{i,j} d((i), (j))$

3. $k = k + 1$, merge (a) and (b) into a new cluster and set his level $L(k) = d((a), (b))$

4. Update D by removing rows and columns corresponding to (a) and (b) and substitute them with new ones representing the new cluster

5. Continue until all points are inside a single cluster.

The FORMULA in the algorithm depends on the strategy adopted to merge the clusters. Hierarchical clustering uses the concept of linkage. The most common types of linkage are the following:

- complete: maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in two clusters and record the largest

- average: mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in two clusters and record the average

- single: minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in two clusters and record the smallest

- centroid: dissimilarity between the centroids of two clusters.

Average and complete linkage are usually preferred over the other ones since they produce more balanced dendrograms.

Finally all types of linkage are based on the choice of the distance d. The most common one is the Euclidean distance, but other choices can be made. Specifically, in this thesis will be used the Wasserstein distance introduced in Section 4.2.

Let's analyze an example of a dendrogram to understand how it is built. In Figure 6 we can see the dendrogram obtained from a data set made of 14 observations. Each leaf of the dendrogram represents one observation and as we move higher up the tree we notice that some leaves begin to fuse into branches. Fusions occur between similar groups. The height of the fusion indicates how different the two observations are; in fact observations that fuse at the bottom of the tree (e.g {1,2}, {5,6}) are very similar, whereas observations that fuse close to the top will tend to be quite different. In order to identify the clusters we make an horizontal cut at a given height across the dendrogram. The number of branches crossed by the cut gives the number of clusters obtained (5 in the example). Depending on the number of clusters one wants to obtain, the cut will be at a different height. Hence one single dendrogram can be used to obtain any number of clusters.

The term hierarchical hence means that clusters obtained by cutting the dendrogram at a given height are nested within the clusters obtained by cutting the dendrogram at any greater height. For example {12,13} is among the first clusters to be created; increasing the height of the cut we will obtain the cluster {11,12,13,14} that contains the previous one.

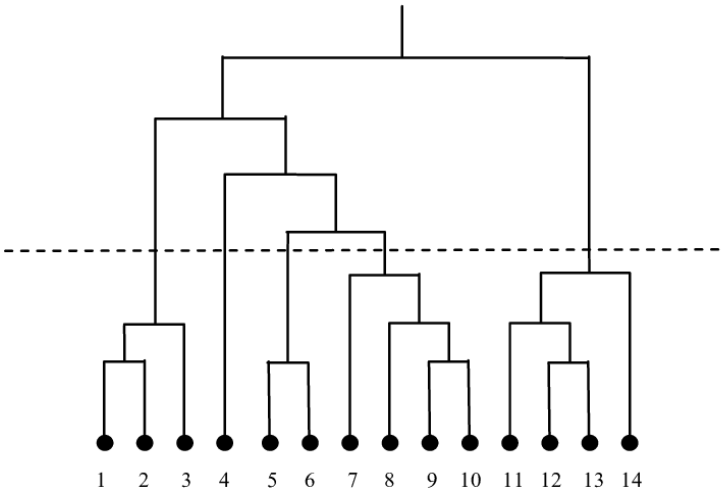


Figure 6: Example of a dendrogram.

4.4 Projection and Hierarchical clustering applied to the data

The analysis performed have been done in three different levels, for each of which the hierarchical clustering has been applied:

- Weekdays: the data has been grouped per weekday
- Days of the month: the data has been grouped per day of the month
- Time slots: each weekday has been divided into five time slots (0-5, 6-10, 11-15, 16-19, 20-23). The data has hence been grouped according to the pair (weekday, time slot)

Those different groupings have the following purposes: the first one is aimed to find out which weekdays are more similar, the second one has as a goal to discover if the clustering technique recognises as similar the days of the month corresponding to the same weekday (for example all Mondays get grouped together), finally the last analysis has as objective finding out if the mobility is similar across different weekdays in the same time slot.

Depending on the results it will be clear which is the distinctive factor in the mobility: either weekdays or time slot.

The starting and ending points in the day or in the time slot of interest identify a discrete probability distribution. For example if the events on Tuesday $x_1, \dots x_n$ are the car bookings that occurred in that day, the distribution can be defined as sum of Dirac delta functions.

Recall the the Dirac delta probability measure defined in equation (4.1) and consider

$$\delta_0(A) = \begin{cases} 1 & \text{if } 0 \in A \\ 0 & \text{otherwise} \end{cases}$$

Then for any point x_0 the Dirac delta measure can be expressed by means of δ_0 as $\delta_{x_0} = \delta_0(x - x_0)$.

Thanks to this observation the discrete probability distribution of the example (and analogously all the others) can be expressed as:

$$D_{Tue}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_0(x - x_i).$$

The idea is to compute, for each one of the different groupings, a matrix of Wasserstein distances between the distributions and then give it as input for the hierarchical clustering. As explained in section 5.2, to compute the Wasserstein distance three elements are needed: the cost matrix M and two unitary vectors a and b.

In this case study the distributions are given by the longitude-latitude coordinates and the cost is simply the distance between all the pairs of points of such distributions. Since the Earth is spherical, to compute the distance between two longitude-latitude points, the Euclidean distance cannot be used and the Haversine formula is used instead. However, since the computation of the Haversine formula is very expensive and the points of the data set are situated in a very restricted area, specifically in the same city, some approximations can be done in order to use the Euclidean distance whose computational cost is by far smaller.

For this reason the Equirectangular projection described in section 5.1 has first been applied to the longitude-latitude points. For simplicity in Figure 7 is shown the result of the projection only on the starting points.

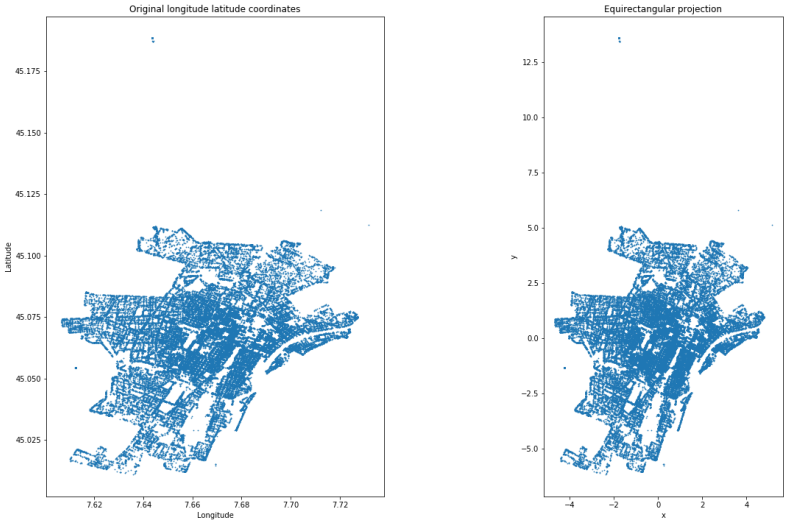


Figure 7: Equirectangular projection over the starting points.

After this step the Euclidean distance has been used to compute the cost matrix M in the three cases. With regards to the unitary vectors a and b, two uniform vectors have been taken because all points have the same importance, no matter where they are situated. Finally the matrix of Wasserstein

distances, separately for starting and ending points, has been computed using the *emd2* function of the Python Optimal Transport (POT) [20] module that implements the W_1 distance:

$$W_1^1(\mu^1, \mu^2) = \min \left\{ \int_{X^2} d(x_1, x_2) d\mu(x_1, x_2) \ : \ \mu \in \Gamma \mu^1, \mu^2) \right\}$$

Notice that the matrices in the three cases are symmetric with zero value on the diagonal and have the following shapes:

- Weekdays: 7×7 matrix (number of weekdays)
- Days of the month: 31×31 matrix (number of days in october)
- Time slots: 35×35 matrix (7 weekdays and 5 time slots per day)

Once those matrices have been obtained, it has eventually been possible to apply to each of them the Hierarchical clustering algorithm with complete linkage.

The result obtained for the first grouping, shown in Figure 8, shows that both in the starting and ending case Sunday is by far the most different weekday. Another aspect common to the two cases is the similarity between Tuesday and Thursday.

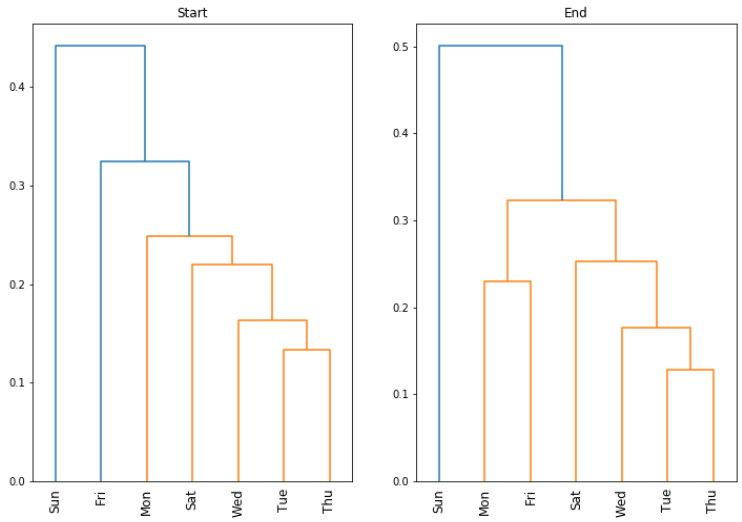


Figure 8: Weekdays Hierarchical clustering.

In Figure 9 are shown the dendrograms obtained for the second grouping separately for starting and ending points. The interesting fact that can be noticed is that in both cases Sundays are almost perfectly separated from the other days, confirming the result obtained in the previous case. With regards to the other days, the Hierarchical clustering did not highlight any relevant pattern. This suggested that the distinctive factor in the mobility was not the weekday.

Hence the last case has been implemented. In Figure 10 is evident that the Hierarchical clustering produced much better results with respect to the previous cases. In fact one can see that in both cases the same time slots have been grouped together most of the times. The most evident separation has been obtained in the start case for the 0-5 time slot. This result is coherent with the expectations, in fact those are the hours of the day in which there is less traffic and there is less people around the city. However this is not the only time slot in which some good results have been produced: in most of the cases all the other time slots are grouped together in the correct way for both the starting and ending points.

Figure 10, compared to the previous ones, clearly indicated that the distinctive factor in the mobility over Turin is the time slot rather than the weekday.

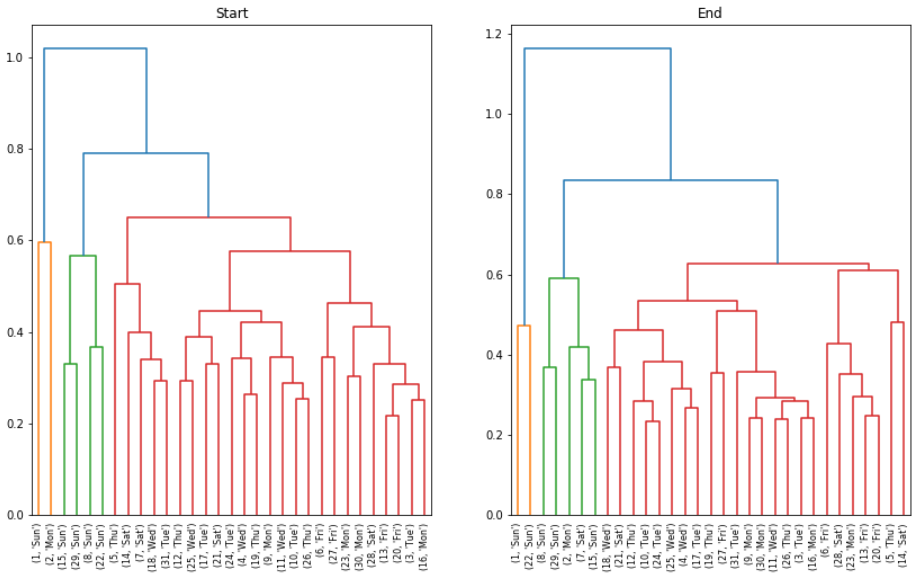


Figure 9: Days of the month Hierarchical clustering.

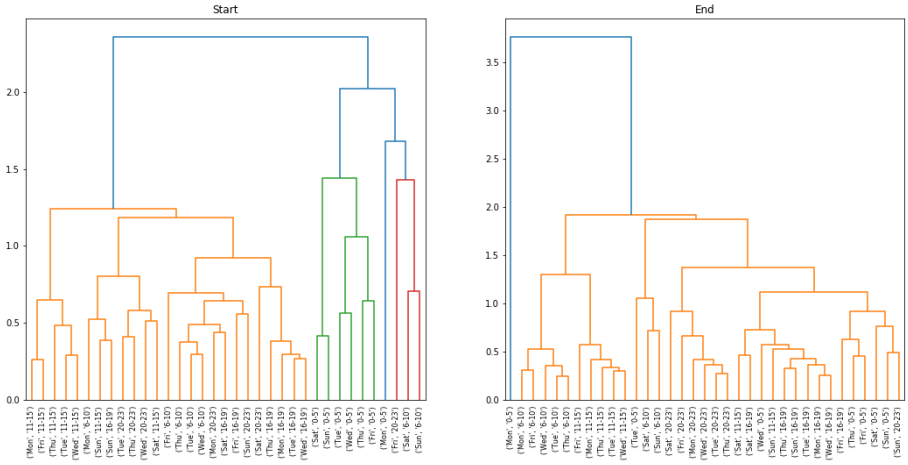


Figure 10: Time slots Hierarchical clustering.

Chapter 5

Topological data analysis

Topological data analysis (TDA) is a collection of techniques whose aim is finding structures in complex data sets using algebraic topology.

For the most part of the chapter the work of Edelesbrunner [8] has been taken as reference. Section 5.1 is devoted to introducing the elementary structures of algebraic topology: simplicial complexes. sections 5.2 and 5.3 describe respectively complexes that originate from convex sets, among which the most famous are the Vietoris-Rips complexes, and the Alpha complexes. Sections 5.5-5.7 introduce the two key concepts of TDA: Homology, its computation in matrix form and Persistent Homology. In section 5.8 are described two concepts that will be needed in the last section: Cohomology and Alexander Duality. Finally section 5.9 is devoted to the introduction of the Homological Scaffold, moreover in this section a new method for the choice of the representatives of the homology classes is proposed: tight cycles.

5.1 Simplicial Complexes

A topological space can be represented in various ways, one of which is as a decomposition into simple pieces. Such decomposition can be called complex if the pieces are topologically simple and their intersections are lower-dimensional pieces of the same type. In particular in this chapter simplicial complexes are used as primitive data structure to represent topological spaces.

Definition Let u_0, u_1, ..., u_k be points in $\mathbb{R}^d$. A point $x = \sum_{i=0}^{k} \lambda_i u_i$ is said *affine combination* of the u_i if the λ_i sum to 1. The *affine hull* is the set of affine combinations. Two affine combinations $x = \sum_{i=0}^{k} \lambda_i u_i$ and $y = \sum_{i=0}^{k} \mu_i u_i$ are said *affinely independent* iff $\lambda_i \neq \mu_i \ \forall i$. An affine combination $x = \sum_{i=0}^{k} \lambda_i u_i$ is a *convex combination* if all λ_i are non-negative. The *convex hull* is the set of convex combinations.

Definition A *k-simplex* is the convex hull of $k+1$ affinely independent points $\sigma = conv\{u_0, u_1, ..., u_k\}$. Its dimension is $dim\sigma = k$

Even though in high dimensions it is difficult to imagine them, the first few simplices are quite intuitive: a 0-simplex is a vertex, a 1-simplex is an edge, a 2-simplex is a triangle and a 3-simplex a tetrahedron. Moreover it is clear that an n-dimensional simplex is composed of $n+1$ points.

Definition A *face* τ of a simplex σ ($\tau \leq \sigma$) is the convex hull of a non-empty subset of the u_i. It is said *proper* if the subset is not the entire set ($\tau < \sigma$). If τ is a proper face of σ, we call σ a *proper coface* of τ.

Definition The *boundary* of a simplex σ ($bd\sigma$) is the union of all the proper faces; its *interior* is everything else $int\sigma = \sigma - bd\sigma$.

Definition A *simplicial complex* is a finite collection of simplices K such that:

i. $\sigma \in K$ and $\tau \leq \sigma$ implies $\tau \in K$

ii. $\sigma, \sigma_0 \in K$ implies $\sigma \cap \sigma_0$ is either empty or a face of both.

The *dimension* of K is the maximum dimension of any of its simplices.

Definition The *underlying space* of a simplicial complex K ($|K|$) is the union of its simplices along with the topology inherited from the ambient Euclidean space in which the simplices live. A *polyhedron* is the underlying space of a simplicial complex.

Definition A *subcomplex* of K is a simplicial complex $L \subseteq K$. It is said *full* if it contains all simplices in K spanned by the vertices in L.

Often, rather than constructing a complex directly in the Euclidean space, it is easier building it abstractly first.

Definition An *abstract simplicial complex* is a finite collection of sets A such that $\alpha \in A$ and $\beta \subseteq \alpha$ implies $\beta \in A$. The sets in A are its *simplices*. The *dimension* of a simplex is $dim\alpha = card\alpha - 1$ and the dimension of the complex is given by the maximum dimension of any of its simplices. A *face* of α is a non-empty subset $\beta \subseteq \alpha$ which is proper if $\beta \neq \alpha$. The *vertex set* of A is the union of all its simplices, i.e. $VertA = \bigcup A$. A *subcomplex* of A is an abstract simplicial complex $B \subseteq A$

Definition Given a simplicial complex K we can construct an abstract simplicial complex A. A is said *vertex scheme* of K, whereas K is said *geometric realization* of A.

The following fundamental theorem ensures that every abstract simplicial complex can be reconstructed in the Euclidean space.

Theorem (Geometric Realization Theorem) Every abstract simplicial complex of dimension d has a geometric realization in $\mathbb{R}^{2d+1}$.

Simplicial maps between simplicial spaces are now introduced, the equivalent of continuous maps between topological spaces.

Definition Let K be a simplicial complex with vertices $u_0, u_1, ..., u_n$. Every point $x \in |K|$ belongs to the interior of exactly one simplex in K. Let $\sigma = conv\{u_0, u_1, ..., u_k\}$ be this simplex. Then we have $x = \sum_{i=0}^{k} \lambda_i u_i$ with $\sum_{i=0}^{k} \lambda_i = 1 \ \forall i$. Setting $b_i(x) = \lambda_i$ for $0 \leq i \leq k$ and $b_i(x) = 0$ for $k+1 \leq i \leq n$ we have that $x = \sum_{i=0}^{n} b_i(x) u_i$ and we call the $b_i(x)$ *barycentric coordinates* of x in K.

Definition A *vertex map* is a function $\varphi : VertK \longrightarrow VertL$ with the property that the vertices of every simplex in K map to vertices of a simplex in L. Then the barycentric coordinates can be used to extend φ to a continuous map $f : |K| \longrightarrow |L|$ called *simplicial map* induced by φ defined as follows:

$$f(x) = \sum_{i=0}^{n} b_i(x) \varphi(u_i)$$

Definition If the vertex map φ is bijective and $\varphi^{-1} : VertL \longrightarrow VertK$ is also a vertex map, then the induced simplicial map f is a homeomorphism. In such case we say that f is a *simplicial homeomorphism* or an *isomorphism* between K and L.

Definition The *diameter* of a set in the Euclidean space is the supremum over the distances between its points.

5.2 Convex Set Systems

Simplicial complexes often originate from the intersections of sets. The most simple case is when such sets are convex sets, in particular balls. Hence this section is devoted to the description of this family of complexes.

In $\mathbb{R}^d$ the following general result holds.

Theorem (Helly's Theorem) Let $X_1, X_2, ..., X_n$ be a finite collection of closed, convex sets in $\mathbb{R}^d$ with $n > d+1$. The intersection of every $d+1$ of these sets is nonempty if and only if the whole collection has a nonempty intersection.

Another important notion that is left to define is the equivalence between topological spaces.

Definition Consider two continuous maps $f, g : X \longrightarrow Y$. A *homotopy* between f and g is a continuous map $H : \mathbb{X} \times [0,1] \longrightarrow Y$ defined as $H(x,0) = f(x)$ and $H(x,1) = g(x)\ \forall x \in X$. A homotopy defines an equivalence relation between f and g ($f \simeq g$). Hence if we interpret $t \in [0,1]$ as time, the homotopy can be taught as a time series that starts at f and ends at g.

Definition Given two topological spaces X, Y such that $Y \subseteq X$ we call Y a *retract* of X if there is a continuous map $r : X \longrightarrow Y$ with $r(y) = y\ \ \forall y \in Y$. The map r is called *retraction*.

Definition We call Y a *deformation retract* and r a *deformation retraction* if there is a homotopy between r and id_X, i.e. if $r \simeq id_X$.

Obviously, a deformation retract is a retract, but the opposite is not true. The concept of deformation retract can be generalized considering maps in both directions as follows.

Definition We say that two topological spaces X, Y not necessarily nested are *homotopy equivalent* ($X \simeq Y$) if there are continuous maps $f : X \longrightarrow Y$ and $g : Y \longrightarrow X$ such that $g \circ f \simeq id_X$ and $f \circ g \simeq id_Y$. In such case we say that X and Y have the same *homotopy type*.

Definition Given a finite collection of sets F, we define the *nerve*:

$$NrvF = \{X \subseteq F | \bigcap X \neq \emptyset\}$$

in other words it consists of all non-empty subcollections whose sets have a non-empty common intersection.

Note that the nerve is always an abstract simplicial complex. In fact if $\bigcap X \neq \emptyset$ and $Y \subseteq X$, then $\bigcap Y \neq \emptyset$.

Theorem (Nerve Theorem) Let F be a finite collection of closed, convex sets in the Euclidean space. Then NrvF and $\bigcup F$ have the same homotopy type.

After having given all those definitions and theoretical results, we can introduce some simplicial complexes formed by geometric balls.

Definition (Čech complex) Let S be a finite set of points in $\mathbb{R}^d$, $B_x(r) = x + r\mathbb{B}^d$ is the closed ball with center x and radius r. The *Čech complex* of S and r is defined as:

$$\check{C}ech(r) = \{\sigma \subseteq S | \bigcap_{x \in \sigma} B_x(r) \neq \emptyset\}$$

i.e. the Čech complex of S and r is isomorphic to the nerve of such collection of balls.

It is easy noticing that for $r_0 \leq r$, $\check{C}ech(r_0) \subseteq \check{C}ech(r)$. Hence increasing the radius the result is a family of nested complexes.

To understand which set of points form simplices in the Čech complex, the miniball strategy can be used. Let $\sigma \subseteq S$ be a subset of points and let the miniball of σ be the smallest closed ball that contains σ and call its radius r_0. Then $\sigma \in \check{C}ech(r)$ if and only if $r_0 \leq r$.

The following complex is a generalization of the Čech complex that, rather than checking all sub-collections, checks the pairs.

Definition (Vietoris-Rips complex) The *Vietoris-Rips complex* of S and r is defined as:

$$Vietoris - Rips(r) = \{\sigma \subseteq S | diam\sigma \leq 2r\}$$

Lemma (Vietoris-Rips) Let S be a finite set of points in the Euclidean space and $r \geq 0$. Then $Vietoris - Rips(r) \subseteq \check{C}ech(\sqrt{2}r)$

In Figure 11 is shown an example of Vietoris-Rips complex. There are eighteen 0-simplices (points), two 0-simplices form a 1-simplex (an edge) if their closed balls of radius r intersect. Three vertices form a 2-simplex (a triangle) if they are pairwise connected by edges. Four vertices form a 3-simplex (a tetrahedron) if they are pairwise connected by edges.

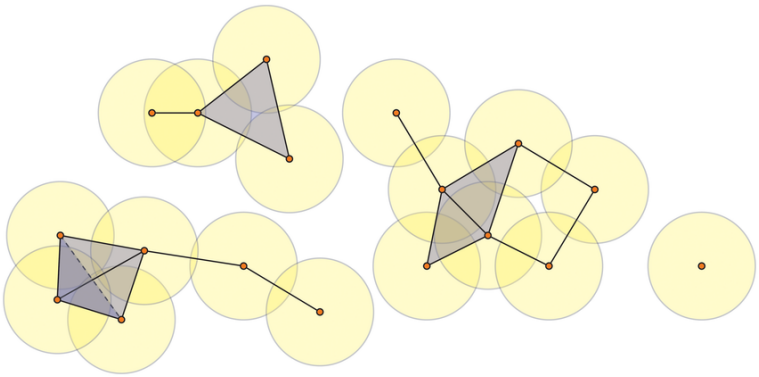


Figure 11: Example of Vietoris-Rips complex.

5.3 Delaunay Complexes

In this section are described new structures that limit the dimension of the simplices obtained from a nerve.

Definition Let $S \subseteq \mathbb{R}^d$ be a finite set. The *Voronoi cell* of a point $u \in S$ is the set of points for which u is the closest: $V_u = \{x \in \mathbb{R}^d | \|x - u\| \leq \|x - v\|, v \in S\}$.

Notice that V_u is a convex polyhedron in $\mathbb{R}^d$ and that two Voronoi cells have at most one segment of their boundary in common (Figure 12).

Definition The *Voronoi diagram* of S is the collection of the Voronoi cells of its points.

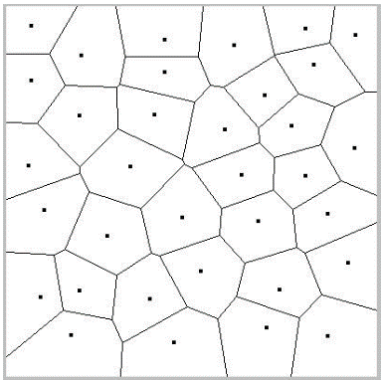


Figure 12: Example of Voronoi diagram of points on a plane.

Definition The *Delaunay complex* of a finite set $S \subseteq \mathbb{R}^d$ si isomorphic to the nerve of the Voronoi diagram:

$$Delaunay = \{\sigma \subseteq S | \bigcap_{u \in \sigma} V_u \neq \emptyset\}$$

Definition We say that the set S is in *general position* if no $d+2$ Voronoi cells have a non-empty common intersection, or equivalently if the dimension of any simplex in the Delaunay complex is at most d.

Assuming general position, taking the convex hulls of abstract simplices is called *Delaunay triangulation* (Figure 13).

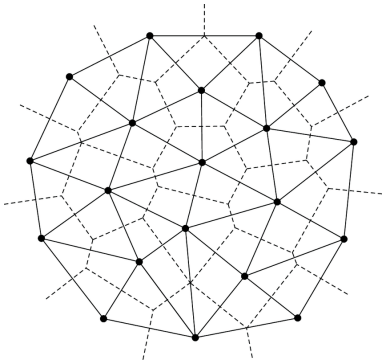


Figure 13: Delaunay triangulation (solid lines) superimposed on the Voronoi diagram.

5.4 Alpha Complexes

In this section a new family of complexes is described, specifically subcomplexes of the Delaunay complex: Alpha complexes.

Let $S \subseteq \mathbb{R}^d$ be a finite set. For each point $u \in S$ let $B_u(r)$ be the closed ball with center u and radius $r > 0$. Now define $R_u(r)$ as the intersection between $B_u(r)$ and the corresponding Voronoi cell $V_u(r)$, $R_u(r) = B_u(r) \cap V_u(r)$. Since both $B_u(r)$ and $V_u(r)$ are convex, so is $R_u(r)$.

Definition The *Alpha complex* of a finite set $S \subseteq \mathbb{R}^d$ is defined as:

$$Alpha(r) = \{\sigma \subseteq S | \bigcap_{u \in \sigma} R_u(r) \neq \emptyset\}$$

i.e. the Alpha complex is isomorphic to the nerve of the cover of the union of the $B_u(r)$s.

Notice that the Alpha complex is a subcomplex of the Delaunay complex because $R_u(r) \subseteq V_u(r)$. Moreover since $R_u(r) \subseteq B_u(r)$ it is also a subcomplex of the Čech complex.

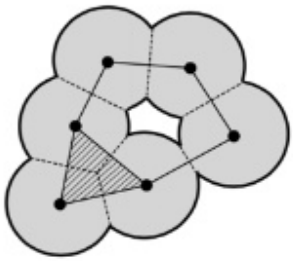


Figure 14: Alpha complex: the union of closed balls is decomposed into convex regions by the Voronoi cells.

Increasing or decreasing the radius leads to respectively larger or smaller alpha complexes. This leads to the following definition.

Definition Let $S \subseteq \mathbb{R}^d$ be a finite set, increasing the radius $r > 0$ it is obtained a 1-parameter family of nested alpha complexes that are all subcomplexes of the same Delaunay complex. Let K_i be the i^{th} alpha complex, hence we get the sequence:

$$\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_m$$

called *filtration* of K_m = Delaunay.

5.5 Homology

Homology is a mathematical concept that allows to associate a sequence of algebraic objects such as groups to topological spaces. Specifically, the reason homology groups have been defined is that they allow distinguishing two shapes by examining their holes.
For the sake of simplicity, **we will use $\mathbb{Z}_2$ as the field of coefficients** that we will use to compute homology. In this case it will be easier to give an interpretation to the results yielded by homology. For a more general description it is possible to see [16].

Definition Let K be a simplicial complex and p a dimension. A *p-chain* is a formal sum of p-simplices in K and is denoted as $c = \sum a_i \sigma_i$ where the σ_i are the p-simplices and the a_i are coefficients chosen in $\mathbb{Z}_2$.

Proposition Let $c = \sum a_i \sigma_i$, $c' = \sum b_i \sigma_i$ be two p-chains. Then $c + c' = \sum (a_i + b_i)\sigma_i$.

Proposition Two p-chains with the addition operation form a group denoted as $C_p = C_p(K)$ and called *group of p-chains*.

Proposition The set $C_p(K)$ is a vector space. The set of elementary chains $\{\sigma_i \mid dim(\sigma_i) = p\}$ forms a basis of $C_p(K)$

For each $i > 0$, the vector spaces C_{i+1}, C_i are related by a boundary operator that associates to each i-simplex σ the formal sum of $i - 1$ simplices that are faces of σ.

Definition We define the *boundary* of a p-simplex as the sum of its $(p-1)$-dimensional faces. In other words let $\sigma = [u_0, u_1, ..., u_p]$ be the simplex spanned by the vertices u_i. The boundary of σ is:

$$\partial_p \sigma = \sum_{j=0}^{p} [u_0, ..., \hat{u_j}, ..., u_p]$$

where $\hat{u_j}$ indicates that u_j is omitted.

Proposition Let $c = \sum a_i \sigma_i$ be a p-chain. Its boundary is the sum of the boundaries of its simplices, i.e. $\partial_p c = \sum a_i \partial_p \sigma_i$.

This means that ∂_p maps a p-chain to a (p-1)-chain and we write $\partial_p : C_p \longrightarrow C_{p-1}$.

Proposition Let c, c' be two p-chains. Then $\partial_p(c + c') = \partial_p c + \partial_p c'$.

As a consequence of the last property ∂_p is a homomorphism called *boundary map* or *boundary homomorphism*.

Definition A *chain complex* is the sequence of chain groups connected by boundary homomorphisms:

$$... \xrightarrow{\partial_{p+2}} C_{p+1} \xrightarrow{\partial_{p+1}} C_p \xrightarrow{\partial_p} C_{p-1} \xrightarrow{\partial_{p-1}} ...$$

Definition A *p-cycle* is a p-chain c with empty boundary, i.e $\partial_p c = 0$.

Definition A *p-boundary* is a p-chain that is the boundary of a $(p+1)$-chain $c = \partial d$ with $d \in C_{p+1}$.

We denote with $Z_p = Z_p(K)$ the *group of p-cycles*, which is a vector subspace of C_p. More formally Z_p is the kernel of the p-th boundary homomorphism: $Z_p = ker\partial_p$.

We denote with $B_p = B_p(K)$ the *group of p-boundaries*, which is a vector subspace of C_p. More formally B_p is the image of the $(p+1)$-th boundary homomorphism: $B_p = im\partial_{p+1}$.

Lemma (Fundamental Lemma of Homology) $\partial_p \partial_{p+1} d = 0$ for every integer p and every $(p+1)$-chain d, i.e. boundaries do not have a boundary.

The importance of the last Lemma lies in the fact that states that all p-boundaries are also p-cycles or, in other words, that B_p is a vector subspace of Z_p and hence it is possible taking quotients.

Definition The *p-th homology group* is the p-th cycle group modulo the p-th boundary group: $H_p = Z_p / B_p$. We call *p-th Betti number* the dimension of this vector subspace, $\beta_p = dim H_p = dim Z_p - dim B_p$.

The elements of H_p are called *homology classes* and are obtained by adding all p-boundaries to a given p-cycle, $c + B_p$, $c \in Z_p$. Two cyles in the same homology class are said *homologous* ($c \sim c'$). For each class one cycle c is taken as representative.

Intuitively H_0 counts the connected components, H_1 counts the cycles, H_2 counts the voids and so on.

Consider a simplicial complex K made of just one vertex. Then one would expect that, since homology counts the holes, $H_p(K) = 0 \quad \forall p$. However this is not true, in fact $H_p(K) = 0 \quad \forall p$ except when $p = 0$ when it has dimension 1. In order to have the 0^{th} homology group behaving like the other ones in this special case, one slight modification needs to be made. Adding the *augmentation map* $\epsilon : C_0 \longrightarrow \mathbb{Z}_2$ defined by $\epsilon(u) = 1$ for each vertex u. Hence we get:

$$\ldots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\epsilon} \mathbb{Z}_2 \xrightarrow{0} 0$$

All is defined as before, the only difference is for Z_0 that now requires that each 0-cycle has an even number of vertices.

The result are the *reduced homology groups* $\tilde{H}_p$ and the *reduced Betti numbers* $\tilde{\beta}_p = \tilde{H}_p$. $\tilde{\beta}_p = \beta_p \quad \forall\, p \geq 1$ and $\tilde{\beta}_0 = \beta_0 - 1$.

5.6 Homology computation

Homology can be easily computed using the matrix form.

Let K be a simplicial complex, its p^{th} *boundary matrix* represents the $p-1$-simplices as rows and the p-simplices as columns: $\partial_p = [a_i^j]$, $i = 1, \ldots, n_{p-1}$, $j = 1, \ldots, n_p$. Given a p-chain $c = \sum a_i \sigma_i$ the boundary can be computed as follows:

$$\partial_p c = \begin{pmatrix} a_1^1 & a_1^2 & \ldots & a_1^{n_p} \\ a_2^1 & a_2^2 & \ldots & a_2^{n_p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n_{p-1}}^1 & a_{n_{p-1}}^2 & \ldots & a_{n_{p-1}}^{n_p} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n_p} \end{pmatrix}$$

where $a_i^j = 1$ if the i^{th} $(p-1)$-simplex is a face of the j^{th} p-simplex, otherwise $a_i^j = 0$.

Two types of column operations can be applied to ∂_p, both multiplying ∂_p with a matrix $V = [v_i^j]$ from the right:

- Exchanging two columns k and l: $v_l^k = v_k^l = 1$, $v_i^i = 1$, $\forall i \neq k, l$ and all the other entries equal to zero
- Adding two columns k and l: $v_k^l = 1$, $v_i^i = 1$, $\forall i$ and all the other entries equal to zero (Figure 14).

Analogously the same operations can be defined over the rows by multiplying ∂_p with a matrix $U = [u_i^j]$ from the left:

- Exchanging two rows k and l: $u_l^k = u_k^l = 1$, $u_i^i = 1$, $\forall i \neq k, l$ and all the other entries equal to zero

- Adding two rows k and l: $u_k^l = 1$, $u_i^i = 1$, $\forall i$ and all the other entries equal to zero (Figure 14).

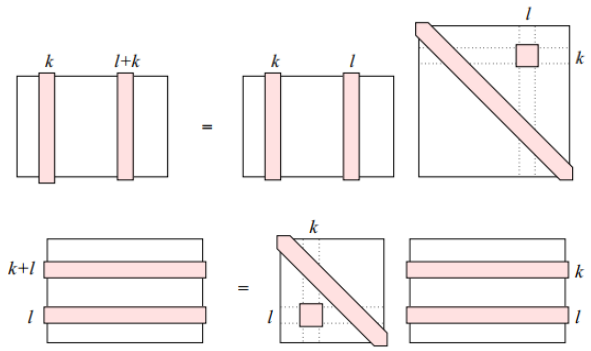


Figure 14: Addition operation between columns (top) and rows (bottom)

Row and column operations can be used to reduce ∂_d to *Smith normal form* that is defined as follows: let A be a non zero $m \times n$ matrix, then there exist $m \times m$ and $n \times n$ invertible matrices S and T such that SAT has the form

$$\begin{pmatrix} \alpha_1 & 0 & 0 & & \dots & & 0 \\ 0 & \alpha_2 & 0 & & \dots & & 0 \\ 0 & 0 & \ddots & & & & 0 \\ \vdots & \vdots & & \alpha_r & & & \vdots \\ & & & & 0 & & \\ & & & & & \ddots & \\ 0 & & & \dots & & & 0 \end{pmatrix}$$

where the diagonal elements satisfy $\alpha_i | \alpha_{i+1} \ \forall 1 \leq i < r$. For arithmetic in $\mathbb{Z}_2$ this means that $\alpha_i = 1 \ \forall i$.

After having reduced the boundary matrices in the Smith normal form, the Betti numbers can be obtained by differences $\beta_p = dim H_p = dim Z_p - dim B_p = dim C_p - rank \partial_p - rank \partial_{p+1}$. The dimensions are directly given by the shape of the reduced matrix, as illustrated in Figure 15.

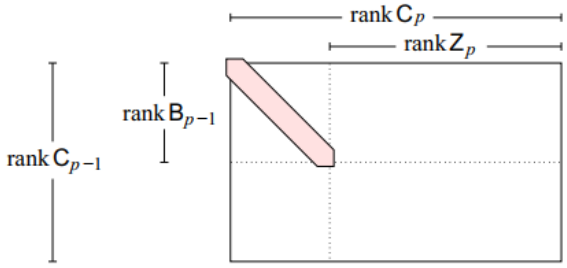
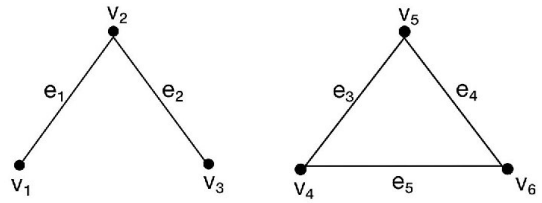


Figure 15: The ranks of the boundary and cycle groups are given by the numbers of non-zero and zero columns

An example is given to clarify the procedure.



Given the simplicial complex above, we want to compute $\beta_0 = dim H_0$. First let's represent in matrix form the boundary map $\partial_1 : C_1 \longrightarrow C_0$ with 1-simplices as columns and 0-simplices as rows:

$$\partial_p = \begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{array} \begin{array}{c} \begin{array}{ccccc} e_1 & e_2 & e_3 & e_4 & e_5 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \end{array}$$

A basis for C_1 is given by the 1-simplices:

$$e_i = \begin{pmatrix} 0 & \dots & 0 & \overset{i}{1} & 0 & \dots & 0 \end{pmatrix}^T, \quad i = 1, \dots, 5$$

whereas a basis for C_0 is given by the 0-simplices:

$$v_i = \begin{pmatrix} 0 & \dots & 0 & \overset{i}{1} & 0 & \dots & 0 \end{pmatrix}^T, \quad i = 1, \dots, 6$$

To compute the boundary of e_1:

$$\partial_p e_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

We now want to reduce ∂_p to the Smith standard form by applying column and row operations. Let's make column operations first:

$$\begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{array} \begin{array}{c} \begin{array}{ccccc} e_1 & e_2 & e_3 & e_4 & e_3+e_5 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \end{array}$$

In practice, what this column operation just did is a change of basis, indeed the basis of C_1 is now $\{e_1, e_2, e_3, e_4, e_3+e_5\}$. We can additionally simplify the matrix by performing the column operation $e_3 + e_4 + e_5$:

$$\begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{array} \begin{array}{c} \begin{array}{ccccc} e_1 & e_2 & e_3 & e_4 & e_3+e_4+e_5 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

Notice that this makes sense because, since $e_3+e_4+e_5$ is a 1-cycle, its boundary is zero. We can not reduce anymore the column space because now e_1, e_2, e_3, e_4 are four linearly independent vectors.

Let's reduce now by performing row operations to obtain the Smith natural form:

$$\begin{array}{c} \\ v_1+v_2 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{array} \begin{array}{c} \begin{array}{ccccc} e_1 & e_2 & e_3 & e_4 & e_3+e_4+e_5 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

$$
\begin{array}{c} \\ v_1+v_2 \\ v_2+v_3 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{array}
\begin{array}{c}
\begin{array}{ccccc} e_1 & e_2 & e_3 & e_4 & e_3+e_4+e_5 \end{array} \\
\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}
\end{array}
$$

$$
\begin{array}{c} \\ v_1+v_2 \\ v_2+v_3 \\ v_3 \\ v_4+v_5 \\ v_5 \\ v_6 \end{array}
\begin{array}{c}
\begin{array}{ccccc} e_1 & e_2 & e_3 & e_4 & e_3+e_4+e_5 \end{array} \\
\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}
\end{array}
$$

$$
\begin{array}{c} \\ v_1+v_2 \\ v_2+v_3 \\ v_3 \\ v_4+v_5 \\ v_5+v_6 \\ v_6 \end{array}
\begin{array}{c}
\begin{array}{ccccc} e_1 & e_2 & e_3 & e_4 & e_3+e_4+e_5 \end{array} \\
\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}
\end{array}
$$

Finally, exchanging the rows the desired reduction is obtained:

$$
\begin{array}{c} \\ v_1+v_2 \\ v_2+v_3 \\ v_4+v_5 \\ v_5+v_6 \\ v_3 \\ v_6 \end{array}
\begin{array}{c}
\begin{array}{ccccc} e_1 & e_2 & e_3 & e_4 & e_3+e_4+e_5 \end{array} \\
\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}
\end{array}
$$

Recall that $\beta_0 = dimH_0 = dimZ_0 - dimB_0$. Keeping in mind Figure 15 it is straightforward to see that $dimB_0 = 4$. To derive $dimZ_0$ it is enough reminding that $Z_0 = ker\partial_0$, but all 0-simplices have empty boundary hence $dimZ_0 = 6$.

Finally $\beta_0 = dimH_0 = dimZ_0 - dimB_0 = 6 - 4 = 2$ and this is correct because we can see that the simplicial complex has two connected components.

5.7 Persistent Homology

Persistent homology is a powerful tool to compute and study topological features of nested families of simplicial complexes and topological spaces. It allows to encode the evolution of the homology groups of the nested complexes as well as to eliminate noise from the data.

Definition Consider a simplicial complex K and a monotonic function $f : K \longrightarrow \mathbb{R}$ i.e. when σ is a face of τ then $f(\sigma) \leq f(\tau)$. Then for every a the sublevel set $K(a) = f^{-1}(-\infty, a]$ is a subcomplex of K. Letting m the number of simplices in K, we get a sequence of $n + 1 \leq m + 1$ different subcomplexes:

$$\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K$$

Such sequence is called *filtration* of f.

Rather than in the sequence of complexes, we are interested in the topological evolution described by sequences of homology groups.

Note that for every $i \leq j$ we have an induced homomorphism $f_p^{i,j} : H_p(K_i) \longrightarrow H_p(K_j)$ for each dimension p. Thus the filtration becomes a sequence of homology groups connected by homomorphisms:

$$0 = H_p(K_0) \longrightarrow H_p(K_1) \longrightarrow \cdots \longrightarrow H_p(K_n) = H_p(K).$$

Going from K_{i-1} to K_i we gain some homology classes and we lose some for example when they merge with each other.

Definition The p^{th} *persistent homology groups* are the images of the homomorphisms induced by inclusion: $H_p^{i,j} = im f_p^{i,j}$, for $0 \leq i \leq j \leq n$. The corresponding p^{th} *persistent Betti numbers* are the ranks of these groups: $\beta_p^{i,j} = dim H_p^{i,j}$.

Notice that, trivially, $H_p^{i,i} = H_p(K_i)$.

The persistent homology groups are composed of those homology classes of K_i that are still alive at K_j. Births and deaths of the homology classes can be described more formally.

Definition Let γ be a class in $H_p(K_i)$, we say that it is *born at* K_i if $\gamma \notin H_p^{i-1,i}$. Moreover if γ is born at K_i we say that it *dies entering* K_j if it merges with an older class as we go from K_{j-1} to K_j i.e. $f_p^{i,j-1}(\gamma) \notin H_p^{i-1,j-1}$ but $f_p^{i,j}(\gamma) \in H_p^{i-1,j}$. In such case we call *persistence of* γ $pers(\gamma) = a_i - a_j$. We can also consider just the indices and define the *index persistence* as $j - i$. If γ is born at K_i but never dies, then both its persistence and its index persistence is infinity.

The most important technique to study persistent homology are persistence diagrams. They are a computable summary of all the persistent homology groups. Persistence diagrams are a very powerful tool to distinguish noise from relevant patterns in the data. Let $\mu_p^{i,j}$ be the number of p-dimensional classes born at K_i and dying entering K_j:

$$\mu_p^{i,j} = (\beta_p^{i,j-1} - \beta_p^{i,j}) - (\beta_p^{i-1,j-1} - \beta_p^{i-1,j})$$

for all $i < j$ and all p. The p^{th} *persistence diagram* $Dgm_p(f)$ of the filtration is obtained drawing each point (a_i, a_j) with multiplicity $\mu_p^{i,j}$ (see Figure 16).

Each point of the persistence diagram represents a class whose persistence is given by its vertical distance from the diagonal. Note that since multiplicities are defined for $i < j$, all points lie above the diagonal. The Betti numbers can be easily obtained from the persistence diagrams, in fact $\beta_p^{k,l}$ is the number of points in the upper left quadrant with corner point (a_k, a_l).

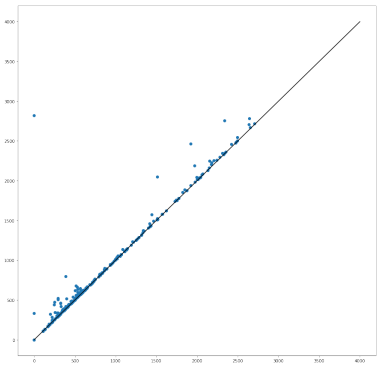


Figure 16: Example of persistence diagram

The points that lie close to the diagonal represent the noise in the data, in fact they represent those classes with very low persistence, that die right after their birth. In contrast, the points far from the diagonal are the ones that give information about the topological structure of the data.

Lemma (Fundamental Lemma of Persistent Homology) Let $\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K$ be a filtration. Then for every pair of indices $0 \leq k \leq l \leq n$ and every dimension p the p^{th} persistent Betti number is:

$$\beta_p^{k,l} = \sum_{i \leq k} \sum_{j > l} \mu_p^{i,j}$$

Hence from the persistence diagrams one can obtain all the information needed.

5.8 Cohomology and Alexander Duality

Another central concept in Topological Data Analysis is cohomology, which is described in this section.

Definition Let K be a simplicial complex. A *p-cochain* is a homomorphism $\varphi : C_p \longrightarrow G$, where $G = \mathbb{Z}_2$. Given a p-chain $c \in C_p$, the cochain evaluates it by mapping it to 0 or 1. The p-dimensional cochains form the *group of p-cochains* C^p.

Let $G = \mathbb{Z}_2$ a group with the addition modulo 2. Let U be a vector space and $\varphi : U \longrightarrow G$ a homomorphism. If φ_0 is another analogous homomorphism, their sum $(\varphi + \varphi_0)(u) = \varphi(u) + \varphi_0(u)$ is again a homomorphism because:

$$\begin{aligned}(\varphi + \varphi_0)(u + v) &= \varphi(u + v) + \varphi_0(u + v) \\ &= \varphi(u) + \varphi(v) + \varphi_0(u) + \varphi_0(v) \\ &= (\varphi + \varphi_0)(u) + (\varphi + \varphi_0)(v).\end{aligned}$$

Taking as neutral element the homomorphism that sends every $u \in U$ to $0 \in G$ and an inverse $-\varphi = \varphi$, a group of homomorphisms from U to G $Hom(U, G)$ is obtained. Given another vector space V and a homomorphism $f : U \longrightarrow V$, there is an induced dual homomorphism $f^* : Hom(V, G) \longrightarrow Hom(U, G)$ that maps $\psi : V \longrightarrow G$ to the composite $f^*(\psi) = \psi \circ f : U \longrightarrow G$.

Recall that the boundary map is a homomorphism between chain groups $\partial_p : C_p \longrightarrow C_{p-1}$. Hence it defines a dual homomorphism between cochains groups: the *coboundary map* $\delta^{p-1} : C^{p-1} \longrightarrow C^p$.

Suppose that φ evaluates a $(p-1)$-simplex to one and all others to zero. Then $\delta\varphi$ evaluates all p-dimensional cofaces of this simplex to one and all to others to zero.

It is now possible to define the *group of cocycles* Z^p and the *group of coboundaries* B^p:

$$\begin{aligned}Z_p &= ker\delta^p : C^p \longrightarrow C^{p+1}, \\ B_p &= im\delta^{p-1} : C^{p-1} \longrightarrow C^p.\end{aligned}$$

Recall that the Fundamental Lemma of Homology states that $\partial_p \circ \partial_{p+1} : C_{p+1} \longrightarrow C_{p-1}$ is the zero homomorphism. Analogously in the cohomology $\delta_{p-1} \circ \delta_p : C_{p-1} \longrightarrow C_{p+1}$ is the zero homomorphism.

Definition The p^{th} *cohomology group* is the quotient of the p^{th} cocycle group modulo the p^{th} coboundary group: $H^p = Z^p / B^p$.

This definition can be again slightly modified to obtain *reduced cohomology groups* $\tilde{H}^p$. The dual homomorphism of the augmentation map introduced in section 5.5 is $\epsilon^* : Hom(\mathbb{Z}_2, G) \longrightarrow C^0$. $Hom(\mathbb{Z}_{\not=}, G)$ has two elements: ϕ_0 that maps 1 to 0 and phi_1 that maps 1 to 1. Hence ϵ^* maps ϕ_0 to ψ_0, which evaluates every vertex to 0, and ϕ_1 to ψ_1, which evaluates every vertex to 1. Finally we get:

$$\ldots \xleftarrow{\delta_1} C_1 \xleftarrow{\delta_0} C_0 \xleftarrow{\epsilon^*} Hom(\mathbb{Z}_2, G) \xleftarrow{0} 0$$

As a consequence of this modification, now the rank of the zeroth cohomology group drops by one.

Cohomology groups are used to define the Alexander duality. Note that the complement of a 2-sphere in the 3 sphere consists of two balls and hence has homology only in dimension 0. Instead the complement of a torus is two solid torii that have homology both in dimension 0 and 1. This observation lead to the intuition that there is a relationship between the homology of a subspace and its complement.

Theorem (Alexander Duality Theorem) Let K be a triangulation of $\mathbb{S}^d$ and let $\mathbb{X} \subseteq \mathbb{S}^d$ be triangulated by a non-empty subcomplex $L \subseteq K$. Then $\tilde{H}_p(\mathbb{X}) \simeq \tilde{H}^{d-p-1}(\mathbb{S}^d - \mathbb{X})$.

Theorem (Universal Coefficient Theorem) Given a simplicial complex K, for every $p \in \mathbb{N}$,there are two isomorphisms:

$$H^p(K) \longrightarrow Hom(H_p(K), \mathbb{Z}_2) \longrightarrow H_p(K),$$

where the first map is a natural isomorphism.

The last theorem is important because states that when the set of coefficients is $\mathbb{Z}_2$, the p^{th} homology group is isomorphic to the p^{th} cohomology group. The same result holds for reduced homology.

Thanks to that, Alexander duality can be formulated in the following way.

Theorem (Alexander Duality Theorem) Let K be a triangulation of $\mathbb{S}^d$ and let $\mathbb{X} \subseteq \mathbb{S}^d$ be triangulated by a non-empty subcomplex $L \subseteq K$. Then $\tilde{H}_p(\mathbb{X}) \simeq \tilde{H}_{d-p-1}(\mathbb{S}^d - \mathbb{X})$.

5.9 Representative cycles and Homological scaffold

In this section is introduced the concept of Homological scaffold, described in [19] and [11].

The homological scaffolds are weighted networks used to summarize the topological information of the homology classes. The weights are used to give a measure of importance to the edges of the graph from the point of view of homology. Given a weighted finite graph $W = (V, E, w)$ and a filtration of simplicial complexes $\mathcal{F}$. Consider the first homology group H_1 for whose homology classes have been taken as representatives the 1-cycles $c_1, \ldots, c_n$. Consider the function $h_W : E \longrightarrow \mathbb{R}^+$

$$h_W = \sum_i \mathbb{1}_{e \in c_i}$$

The homological scaffold of W is the weighted graph whose vertices coincide with the vertices of W, its edge set is a subset of the edge set of W and its weights are given by the function h_W.

The problem to solve now is how to choose the representative cycles for the first homology group. This is in general a non trivial problem since each homology class can be represented by several homologous cycles. As pointed out in [9] this choice can be made by using the Alexander Duality theorem introduced in Section 5.8.

Notice that, given a simplicial complex K, when we consider the homology group $H_0(K)$ a canonical basis is given by the connected components of the complex. The simplicial complexes we consider

in our application are embedded into the plane, and hence can be considered as compact sets on the sphere $\mathbb{S}^2$

For the reasons explained above, given a simplicial complex K a canonical basis for the homology group $H_0(\mathbb{S}^2 \setminus K)$ is given by the connnected components of $\mathbb{S}^2 \setminus K$. In order to obtain a canonical basis also for the reduced homology group $\tilde{H}_0(\mathbb{S}^2 \setminus K)$ one needs to fix one connected component. Let's say such component is represented by a vertex x_0. If the other components are represented by vertices $x_1, \ldots, x_n$, then the 0-chains $[x_1] + [x_0], \ldots, [x_n] + [x_0]$ are a basis of $\tilde{H}_0(\mathbb{S}^2 \setminus K)$, where we indicate with $[x_1]$ the class of x_1.

Hence we want to use Alexander Duality to obtain a basis of $\tilde{H}_1(K)$. In fact for $p = 1$ and $d = 2$ the theorem states that $\tilde{H}_1(K) \simeq \tilde{H}_{2-1-1}(\mathbb{S}^2 \setminus K) = \tilde{H}_0(\mathbb{S}^2 \setminus K)$.

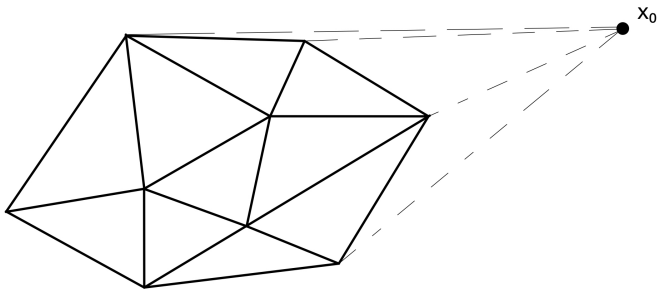


Figure 17: Construction of the structure L

The basis is obtained in the following way. Let K be a triangulation of the plane as shown in Figure 17 and let $T = \sum_i \tau_i$ be the 2-chain formal sum of the 2-simplexes of K. Its boundary $\partial_2 T = \sum_i \sigma_i$ is a 1-cycle composed of the 1-simplexes σ_i. Let x_0 be a point on the plane outside of K. The 1-simplexes of $\partial_2 T$ have vertices v_i^1, v_i^2 for each σ_i. Now adding to $\partial_2 T$ the simplexes $[x_0, v_i^2], [x_0, v_i^2], [x_0, v_i^1, v_i^2]$ the result is a structure L

$$L = K \bigcup_i \left\{ [x0], [x_0, v_i^2], [x_0, v_i^2], [x_0, v_i^1, v_i^2] \right\}$$

that is homeomorphic to S^2. Then, for each vertex x_i representing a connected component of $\mathbb{S}^2 \setminus K$, there exists a maximal set of 2-simplexes of $L \setminus K$ that are in the same connected component of x_i. Such set is obtained by taking those 2-chains that are in L but not in K such that their interior belongs to the connected component of $\mathbb{S}^2 \setminus K$ to which belongs x_i. Finally, taking the boundary of those 2-chains we obtain a 1-cycle that is the i^{th} element of the canonical basis of $\tilde{H}_1(K)$.

Through this process we have obtained a canonical basis $C_1, \ldots, C_r$ of the first homology group of the complex. In the filtration process we would like to express each 1^{st} persistent homology class

born at b and dead at d, in the canonical basis of the homology group $H_1(K_b)$. Then, we can associate to the homology class its unique representative cycle.

In order to find such unique representative cycle we need to introduce the concept of tight cycle.

Definition Given a simplicial complex K, we say that a cycle $z \in Z_1(K)$ is *tight*, if it can be expressed as a linear combination of the elements of the canonical basis: $z = \sum_i \lambda_i C_i$.

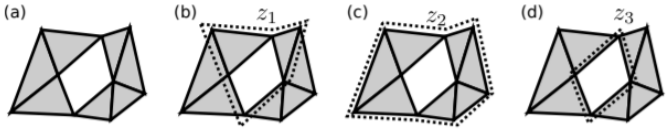


Figure 18: Tight cycle for a simplicial complex with one hole (image taken from [17])

The idea is the following: the simplicial complex in Figure 18 has one homology generator in H_1 and the cycles z_1, z_2 and z_3 shown in Figure 18 b), c) and d) are homologous cycles that represent the same homology class. We consider z_3 as the best cycle to represent this class and call it tight.

Hence, in order to find the desired tight representative for a cycle z one needs to obtain the coefficients λ_i. Notice that since the first homology group is defined as $H_1(K) = Z_1(K)/B_1(K)$, a basis of $Z_1(K)$ can be obtained by extending the canonical basis $< C_1, \ldots, C_r >$ of $H_1(K)$ by adding the 2-boundaries $\partial_2 \sigma_i$. In this way the tight representative for a cycle z can be obtained by solving the linear system:

$$z = \sum_i \lambda_i C_i + \sum_j \mu_j \partial_2 \sigma_j$$

and extracting the coefficients λ_i.

Notice that, for the definition of basis, the coefficients λ_i and μ_j are unique, and hence each cycle has a unique tight representative.

It is now possible to associate to each persistent homology c born in K_b the tight representative in $Z_1(K_b)$. In this way we can define a homological scaffold based on tight cycles.

Chapter 6

TDA applied to the car sharing mobility problem

After having developed the TDA techniques in Chapter 5, in this chapter their application to the mobility pattern identification problem is presented. To this aim in Section 6.1 has been applied the same hierarchical clustering algorithm with complete linkage already used in section 4.4 in a topological context, both with the Vietoris-Rips and the Alpha complexes described in sections 5.2 and 5.4, in order to compare the topological and the classical approaches. In Section 5.9 has been presented a new technique for the choice of good representatives for the homology classes, hence in Section 6.2 is presented its application to the Car-Sharing mobility problem. Finally Section 6.3 validates the results obtained.

6.1 Vietoris-Rips and Alpha clustering

Section 4.4 highlighted that the data shows relevant patterns depending on the hour of the day in which the cars are used, rather than on the day of the week. Hence in this section the objective is further investigating this behaviour by applying again a hierarchical clustering algorithm to the data grouped per pair (weekday, time slot), this time in a topological context.

From now on will be always considered the projected coordinates, obtained by applying the Equirectangular Projection described in Section 4.1.

The approach consists, both for the Vietoris-Rips and the Alpha complexes, in storing the persistence diagrams of the first two homology groups H_0 and H_1 for each pair (weekday, time slot)

separately for starting and ending projected points. Finally two 35×35 matrices have been computed, one for H_0 and one for H_1, with the pairwise sliced Wasserstein distances between persistence diagrams. On those matrices, as in section 4.4, a hierarchical clustering algorithm with complete linkage has been applied and hence it has been possible to compare the results.

Figure 19 shows the results obtained for the Vietoris-Rips complex. There is no big difference between the dendrograms of the starting and ending points. Considering that Vietoris-Rips is the simplest simplicial complex, the results are quite satisfying but, compared to the ones of Figure 10 are worse. In fact the classical approach produced better clusters for the different time slots, for example the 11-15 time slot had been correctly identified whereas now it did not both in H_0 and H_1. On the other hand time slots like 0-5 and 20-23 formed a satisfying cluster in both approaches.

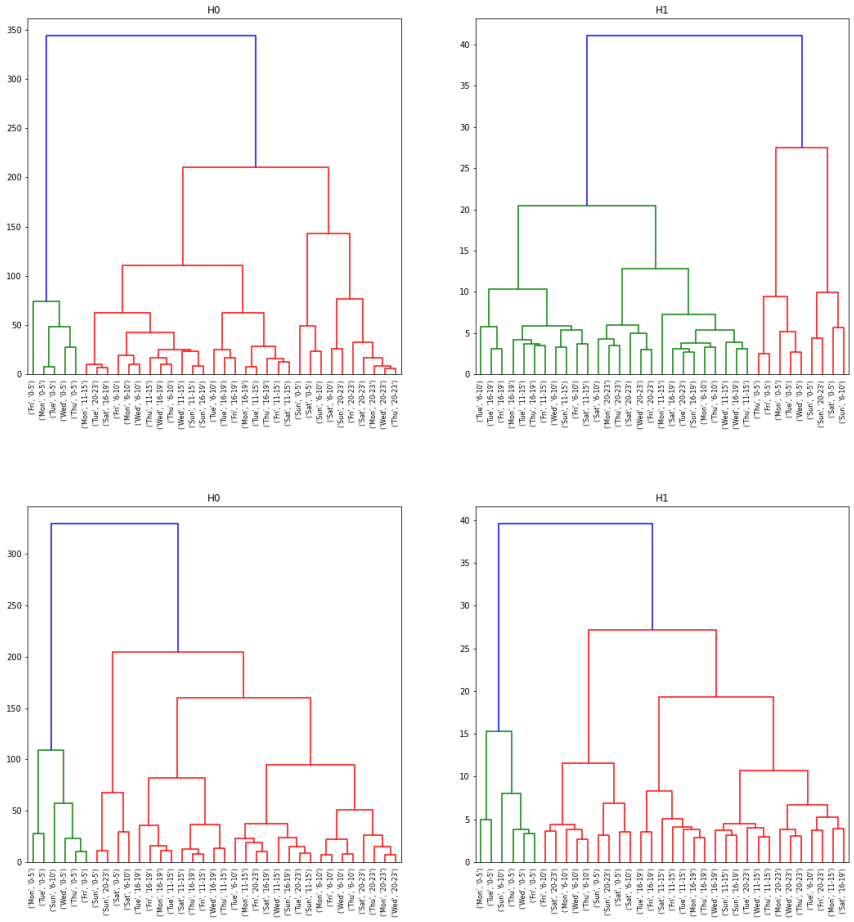


Figure 19: Hierarchical clustering with Vietoris-Rips complexes per pair (weekday, time slot) for starting (top) and ending (bottom) points

Analogously, Figure 20 shows the results obtained with the Alpha complex. First it is interesting noticing that both for the starting and ending points, the dendrograms for H_0 are identical to the ones in Figure 16. Comparing the second Homology groups for the starting points there are slight improvements, the most relevant is that in Figure 19 the 0-5 time slot had been grouped along with 'Sun', 20-23 and 'Sun', 6-10 whereas now the algorithm produced a cluster only with the 0-5 time slot. Also for the ending points there are slight improvements with respect to the previous results, but the most important aspect is that also the TDA approach recognized a relevant pattern in the mobility given by the hourly division.

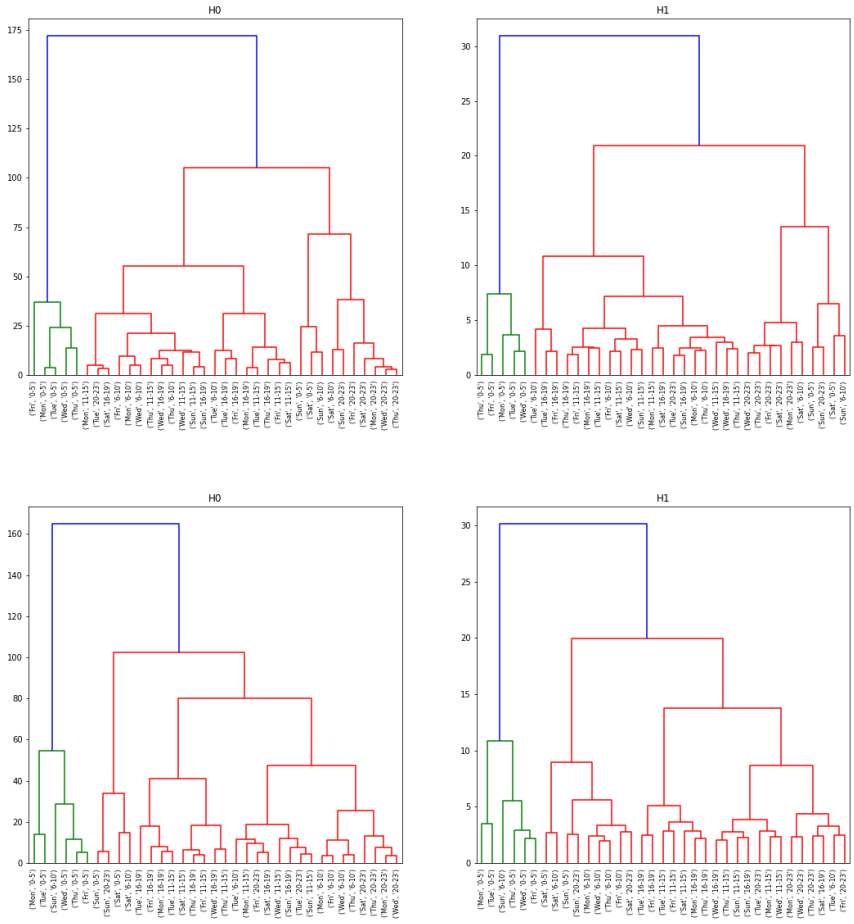


Figure 20: Hierarchical clustering with Alpha complexes per pair (weekday, time slot) for starting (top) and ending (bottom) points

6.2 Extraction of relevant zones

Finally, the zones of interest have been extracted. They are identified by the tight generators of the first homology group obtained through the homological scaffold introduced in Section 5.9.

The procedure adopted has been the following. As explained in Chapter 5, the homological scaffold is built from a filtered simplicial complex. At first a triangulation of the plane is needed. To build such a triangulation a set of points on the plane is taken and the associated Delaunay triangulation, introduced in Section 5.3, is built. This set of points is constituted by a uniform distribution on the plane to which it is added a sample of the booking events registered in the month. This choice is made in order to have a denser triangulation around the Car-Sharing events. A filtering function is defined on the simplicial complex extending monotonically the following density function on the vertices:

$$f_P(x_i) = \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}\Big\{ \|p_j - x_i\|_2 \leq R \Big\}$$

where with x_i are indicated the nodes, with P the set of the events p_j, N is the number of events and R is a radius. In other words for each node, given a radius R, it computes the fraction of events that fall in its neighborhood of radius R. The weight associated to each edge of the Delaunay Triangulation $[x_0, x_1]$ is given by the maximum of the density function values of its vertices: $max\{f(x_0), f(x_1)\}$ and the same procedure is applied to its triangles.

In the right side of Figure 21 is shown the result of the density function described above with a radius of $300m$ applied on the data of Tuesday between 6 a.m and 10 a.m. As expected the centre of the city has a higher density, whereas the peripheral zones have a lower concentration of events.

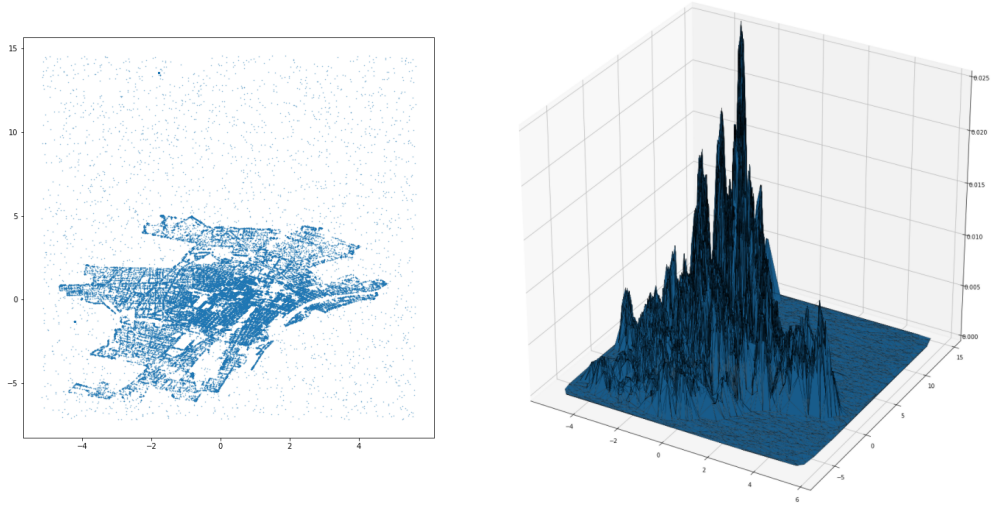


Figure 21: Left: nodes of the Delaunay Tringulation, right: Density function with $R = 300m$ for Tuesday 6 a.m - 10 a.m.

With this construction, for each of the usual groupings (weekday, time slots) we have a filtration of simplicial complexes and hence the persistence homology can be computed. For each homology class of the first homology group, the tight generators are extracted along with their persistence. In order to distinguish between noise and relevant information, only the 10 cycles with the highest persistence have been taken in consideration. To summarize the information given by this set of cycles a subdivision of the plane is defined. It takes into account the number of cycles each point lies into. An integer number is assigned to each point of the plane, equal to the number of tight cycles that contain the point. Therefore we can subdivide the plane in zones:

$$Z_i = \{x \in \mathbb{R}^2 \mid x \text{ is contained in } i \text{ tight cycles}\}.$$

The results, separately for starting and ending points, are shown in Figures 22 and 23. As expected, the cycle with the highest persistence is the biggest one, i.e. the one that indicates the perimeter of the operating area over Turin and it is similar among all the pairs. It has then been decomposed by the algorithm into smaller areas that vary depending on the weekday or time slot. The colors have to be interpreted in the following way. Using a ray casting algorithm [21], for each point of the plane it is possible to count how many times it is included in one of the tight cycles yielded by persistent homology. The darker the color of a point, the higher it is the number of homological cycles that contain that point. Looking at the figures is evident that there is a pattern in the zones of interest detected in the same time slot, in fact for example in Figure 22 between 6 a.m and 10 a.m departures are concentrated in the very centre of the city whereas outside of it there are almost no events. In contrast, between midnight and 5 a.m. departures are spread all over the city. Is interesting noticing that, comparing Figure 22 and Figure 23, the second can be seen as the complementary of the first. In fact looking at the fourth time slot, the arrivals are mainly concentrated in the centre, whereas the departures are spread. Finally, both for the departures and the arrivals, weekends show a different behaviour. This is consistent with what observed at the beginning in Figure 8.

Figure 22: Zones of interest for the starting points for each pair (weekday, time slot)

Figure 23: Zones of interest for the ending points for each pair (weekday, time slot)

To understand the advantages in the detection of the zones of interest given by this topological approach, in Figure 24 is shown the comparison between the approach previously used in [10],[2], [4] and [5] that divided the city in squares $500m \times 500m$ and counts the numbers of events in each square (Figure 24 (a)).

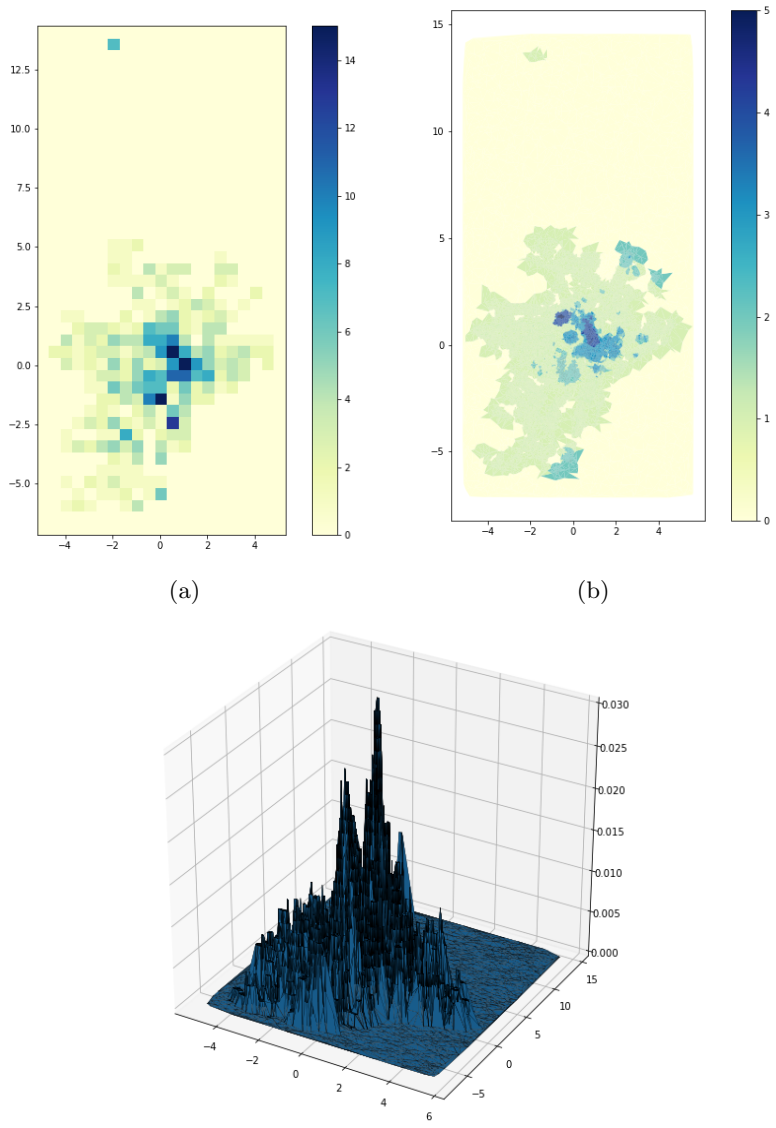


Figure 24: (a) squares subdivision, (b) homological scaffold, (c) density function for October 10^{th} 6 a.m. - 10 a.m.

It is clear that the information given by the topological approach (Figure 24 (b)) is more detailed. This is due to substantial differences in the two methods. The first is that the previous approach fixed on the plane a square grid, whereas the nodes for the Delaunay Triangulation have been chosen

such that the zones in which the bookings concentrate the most are triangulated in a thicker way and hence are the ones in which the analysis focuses the most. Secondly, if in the previous approach the interesting squares were the ones with higher counts, now the filtration process allows to better detect zones of local maxima. This is also the reason why the TDA approach offers complementary results compared to just applying the density function over the triangulation (Figure 24 (c)). One could object that the density function alone could give sufficiently good results, but this is not true. In fact, looking at Figure 24 (c), one could say that in order to detect the most important zones is sufficient taking the zones with a density value greater that 0.01, but in this way all the information relative to the small peaks others than the bigger two would get lost. The biggest advantage given by the TDA approach is that even if a zone has a low density value it could still be detected by homology if, compared to its neighborhood, has higher density values. In fact is evident that, comparing Figure 24 (a) and (b), the boundaries of the operating area of Turin are much more evident in the latter, as well as some other peripheral zones.

6.3 Validation

The last phase of the analysis consists of validating the results obtained. To this aim the approach used reminds the k-fold Cross-Validation procedure: considering the 4 weeks of the month, at each step 3 weeks are taken as training set and the remaining one as test set.

Recall that the city has been divided into zones

$$Z_i = \{x \in \mathbb{R}^2 \mid x \text{ is contained in } i \text{ tight cycles}\}.$$

The objective is proving that the zones with a higher weight are more "densely populated" i.e. are the ones in which occur more events. Obviously in the bigger zones will occur more events, but to get a notion of density the area has to be taken in consideration. For this reason, after extracting the zones of interest from the training set, the goal is counting the number of events in the test set that occur in each zone and dividing it by its area.

The result of this procedure, obtained taking the starting points of the first week as test set, is shown in Figure 25. For each of the (weekday, time slot) groups, is produced a histogram with on the x-axis the weights associated to a certain zone, and on the y-axis the counts of events occurred in the zone, normalized by the area of the zone. The overall behavior observed is that to an increase in the weight corresponds a dramatic increase in the counts. This shows that, as expected, the most important zones are more "densely populated". The worst results are obtained for the time slot 0 a.m - 5 a.m. This is due to the fact that those hours are the ones in which less events occur, and hence the results are less reliable. Another factor to consider is that the data set contains the information of only one month. Collecting more data, and subsequently having bigger training and test sets would probably improve the results.

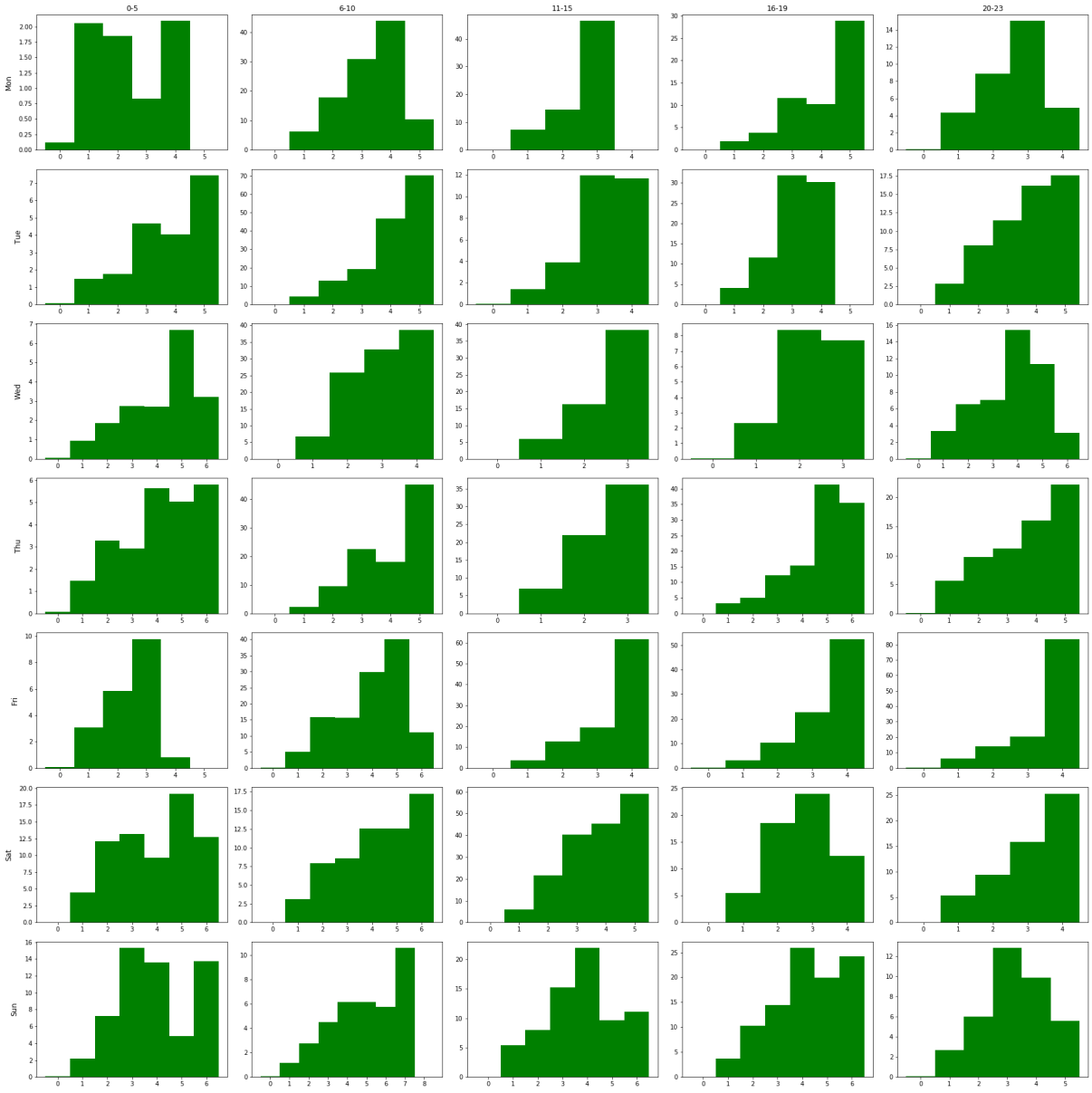


Figure 25: Validation process over the first week

In order to obtain not only a visual comparison but also a numerical one between the validation procedures for the four weeks, two statistics have been computed: the Pearson and Spearman correlation coefficients between the weight of the cycle and the normalized counts. Given two

variables X and Y the Pearson correlation coefficient measures the linear relationship between them and is defined as follows:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

In contrast, the Spearman correlation coefficient measures the monotonic relationship, linear or not. Given a set of items the *ranking* is a relationship such that, for any pair of its elements, either the first is ranked lower or higher than the second or it is ranked equal to the second. Given this definition, the Spearman correlation coefficient between two variables X and Y is defined as follows:

$$\rho_s = \frac{Cov(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}}$$

where rg_X and rg_Y are the ranked variables.

Both coefficients take values between -1 and 1.

Start

	week1	week2	week3	week4
Mean	0.75563	0.819861	0.747244	0.775686
Std	0.277899	0.174013	0.248728	0.267568

End

	week1	week2	week3	week4
Mean	0.656184	0.806155	0.748115	0.778631
Std	0.25475	0.212079	0.289138	0.229871

Table 1: Mean and Standard deviation of Pearson correlation coefficients for starting and ending points

Tables 1 and 2 show respectively the mean Pearson and Spearman correlation coefficients along with their standard deviation for starting and ending points across the four weeks. The mean correlation coefficients are in most cases above 0.7 and it indicates a strong correlation between the number of events occurring in a zone and its importance. The lowest coefficient is obtained for the validation process on the ending points of week 1. As a general trend it can be observed the the Spearman correlation has a higher standard deviation.

Inspecting deeper the behaviour of the coefficients in the validation process, for each week the same statistics have been computed for each time slot and separately for starting and ending points.

Start

	week1	week2	week3	week4
Mean	0.746871	0.839524	0.735646	0.750816
Std	0.335971	0.187614	0.294902	0.34264

End

	week1	week2	week3	week4
Mean	0.63217	0.814218	0.738963	0.792109
Std	0.331486	0.254677	0.342278	0.268362

Table 2: Mean and Standard deviation of Spearman correlation coefficients for starting and ending points

Tables 3 and 4 show the results obtained for the Pearson coefficient. Tables 5 and 6 show, instead, the results for the Spearman coefficient.

In most cases the mean correlation is between 0.7 and 0.9 but across all the weeks the time slot 0-5 is always the one that deviates the most, as already highlighted in Figure 25. The standard deviations are, in general low (between 0.1 and 0.3). Again, as before, the Spearman coefficients have a higher standard deviation. The high mean correlation coefficients along with a low standard deviation are a clear signal of an accurate prediction. Another relevant aspect to notice is that for the ending points, in general, the prediction is less accurate. This indicates that the departures have a more rigid pattern compared to the arrivals.

	Week 1		Week 2		Week 3		Week 4	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
0-5	0.575593	0.447155	0.722915	0.254053	0.618128	0.269713	0.610350	0.366488
6-10	0.788245	0.216842	0.829228	0.159702	0.837354	0.311091	0.841578	0.144495
11-15	0.814875	0.254145	0.914635	0.084107	0.649902	0.216699	0.648940	0.328369
16-19	0.809139	0.212842	0.871892	0.164259	0.841534	0.055163	0.880985	0.219861
20-23	0.790295	0.213874	0.760634	0.154063	0.689301	0.231138	0.896577	0.143687

Table 3: Mean and Standard deviation of Pearson correlation coefficients separately for each time slot for starting points

	Week 1		Week 2		Week 3		Week 4	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
0-5	0.504667	0.300896	0.697052	0.263439	0.742313	0.242551	0.732778	0.252091
6-10	0.536500	0.176845	0.850298	0.158156	0.593567	0.326318	0.845788	0.150480
11-15	0.791657	0.108666	0.872031	0.127257	0.797144	0.297569	0.805265	0.220664
16-19	0.697242	0.323347	0.820103	0.259407	0.802295	0.356339	0.807688	0.248621
20-23	0.750852	0.251129	0.791291	0.250521	0.805254	0.262463	0.701636	0.308374

Table 4: Mean and Standard deviation of Pearson correlation coefficients separately for each time slot for ending points

	Week 1		Week 2		Week 3		Week 4	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
0-5	0.558163	0.492571	0.704762	0.308313	0.525170	0.304104	0.561565	0.435969
6-10	0.786395	0.258677	0.845918	0.169927	0.841837	0.372216	0.838435	0.186751
11-15	0.802041	0.369079	0.941837	0.113870	0.675510	0.327193	0.597959	0.473978
16-19	0.769388	0.318764	0.905102	0.139406	0.920408	0.107177	0.841837	0.321995
20-23	0.818367	0.239999	0.800000	0.093314	0.715306	0.219942	0.914286	0.112788

Table 5: Mean and Standard deviation of Spearman correlation coefficients separately for each time slot for starting points

	Week 1		Week 2		Week 3		Week 4	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
0-5	0.419388	0.402980	0.680272	0.292466	0.759524	0.281889	0.757143	0.249762
6-10	0.516327	0.268332	0.971429	0.054710	0.577551	0.406148	0.910204	0.112702
11-15	0.802381	0.134118	0.917347	117648	0.776106	0.364378	0.840816	0.221993
16-19	0.649286	0.410143	0.773469	0.305469	0.775510	0.403447	0.789796	0.308685
20-23	0.773469	0.307356	0.728571	0.331662	0.806122	0.322681	0.662585	0.400514

Table 6: Mean and Standard deviation of Spearman correlation coefficients separately for each time slot for ending points

Finally, to get the overall behaviour for each pair (weekday, time slot), the mean correlation coefficients have been computed across the four weeks. The results are shown in the heatmaps in Figure 26 (Pearson) and Figure 27 (Spearman) for starting and ending points.

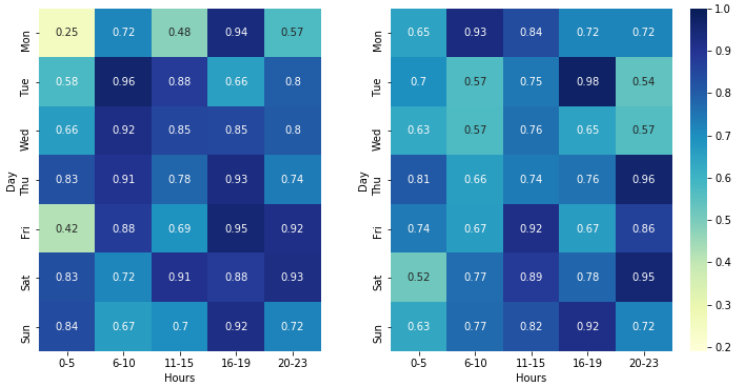


Figure 26: Mean Pearson correlation for starting (left) and ending (right) points

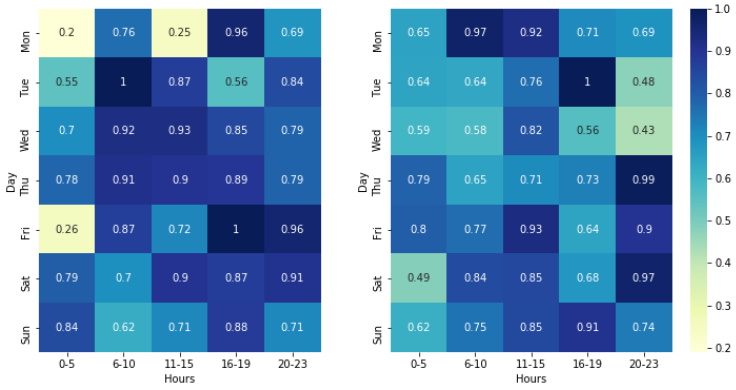


Figure 27: Mean Spearman correlation for starting (left) and ending (right) points

Some interesting aspects can be noticed: the departures' predictions are worse on Mondays and in general in the time slot 0-5, whereas are more accurate in the time slots 6-10 and 16-19. The arrivals, instead, get predicted very well in the time slot 20-23 but only from Thursday to Saturday. Probably this is due to the fact that the nightlife concentrates mostly in those days and in specific zones of the city.

Chapter 7

Conclusions

This thesis had as objective giving a new perspective on the analysis of the Car-Sharing mobility over Turin by using Topological Data Analysis techniques. Given the necessity of introducing electric vehicles based FFCS systems, the desired information was understanding if a topological approach was able to individuate strategic areas of the city for the installation of charging poles, crucial problem for the sustainability of an electric fleet. The approach consisted of a first exploratory phase aimed at identifying patterns in the service utilization. The data set has been divided into 35 groups given by the pairs (weekday, time slot), where the time slots have been chosen by considering the hours of the day that record a homogeneous number of bookings. Specifically in the morning between 6 a.m. and 10 a.m, hours at which people usually go to work, and in the afternoon, between 4 p.m and 7 p.m., hours at which people come back home the utilization rates are higher. In contrast, at night time the utilization rates are much lower.

This subdivision has been used for the implementation of a hierarchical clustering algorithm with complete linkage with dissimilarity measure given by the Wasserstein distance. Such algorithm was applied both in a classical and a topological context. The first through computing the distances between the discrete distributions individuated by the booking events, the latest by building on top of the data of each group the Vietoris-Rips and Alpha complexes. Both the approaches identified a similar behaviour in the mobility between weekdays in the same time slot. Understandably, Saturdays and Sundays proved having slight differences in their behaviour with respect to the other weekdays.

With this insight the analysis proceeded with the individuation of the most relevant zones of the city for each of the aforesaid groups. Such zones, from a topological point of view, are represented by cycles on the plane and hence the attention focused on the first homology group H_1. For

their detection a new technique, based on the tight cycles, has been used. It allowed to solve one of the biggest problems arising from the definition of the homology groups: the choice of good representatives for each homology class. The identified representative cycles allowed to assign to each point of the plane an importance given by the number of tight cycles it lies into. The higher this value, the higher the importance.

Finally the last step consisted of a validation phase aimed at understanding if the detected zones were reliable. By dividing the data into train and test set, the objective was proving that the most important zones of the city were actually the ones in which more booking events occurred. To this aim the Pearson and Spearman correlations have been computed for each group between the importance of a zone and the normalized counts of the events events occurring in it. The results showed that, as expected, the two aspects were highly correlated: in the most important areas occurred more events. Understandably the results were worse in the time slot with less data, between 0 a.m. and 5 a.m., in contrast in the central time slots correlations reached even 0.9 indicating an accurate detection of the areas. One interesting aspect noticed was that prediction were generally better for the starting points if compared to the ending points.

The overall results proved that a topological approach could help in designing an electric FFCS system. Obviously, since the data set used contained the information about the mobility in just one month, more accurate predictions can be obtained by increasing the data set.

In conclusion, the approach used in this work can be very useful for studying the periodicity of the Car-Sharing mobility events. Some aspects that can be investigated in future works are the differences encountered between the classical clustering approach and the topological one. In fact it is not possible to conclude that one is better than the other since they both may capture useful information not detected by the other. Moreover the fact that the departures get better predicted with respect to the arrivals can be further analysed to understand the reasons of this behaviour. The same can be done with Mondays, or all the other weekdays or time slots in which the validation procedure in Section 6.3 did not find satisfying results.

The work done in this thesis focused on the planar case, but it can be extended in higher dimensions considering simplicial complexes to whom the Alexander Duality theorem can be applied.

Obviously the methodology of this thesis is not constrained to the Car-Sharing mobility problem. An attractive area of application are sensor networks and coverage problems related to them, as shown in the work of Robert Ghrist [6]. Another study over sensor networks by Jennifer Gamble [13] showed that persistent homology can be used in a complementary field, to look for zones of lack of coverage.

Bibliography

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. Gradient flows in metric spaces and in the space of probability measures. 2005.

[2] Michelangelo Barulli, Alessandro Ciociola, Michele Cocca, Luca Vassio, Danilo Giordano, and Marco Mellia. On scalability of electric car sharing in smart cities. 2020.

[3] Gunnar Carlsson. Topology and data. *Bull. Amer. Math. Soc. (N.S.)*, 46(2):255–308, 2009. doi:10.1090/S0273-0979-09-01249-X.

[4] Alessandro Ciociola, Dena Markudova, Luca Vassio, Danilo Giordano, Marco Mellia, and Michela Meo. Impact of charging infrastructure and policies on electric car sharing systems. 2020.

[5] Michele Cocca, Danilo Giordano, Marco Mellia, and Luca Vassio. Free floating electric car sharing design: Data driven optimisation. 2020.

[6] Vin Desilva and Robert Ghrist. Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, April 2007. doi:10.2140/agt.2007.7.339.

[7] Tamal Dey, Tianqi Li, and Yusu Wang. Efficient algorithms for computing a minimal homology basis. *Latin American Symposium on Theoretical Informatics*, 10807. URL: https://par.nsf.gov/biblio/10069268, doi:978-3-319-77404-6_28.

[8] H. Edelsbrunner and J. Harer. Computational topology: An introduction. *American Mathematical Society*, 2010.

[9] J. Gamble, H. Chintakunta, and H. Krim. Adaptive tracking of representative cycles in regular and zigzag persistent homology. *ArXiv*, abs/1411.5442, 2014.

[10] Danilo Giordano, Luca Vassio, and Luca Cagliero. A multi-faceted characterization of free-floating car sharing service usage. 2020.

[11] Marco Guerra, Alessandro De Gregorio, Ulderico Fugacci, Giovanni Petri, and Francesco Vaccarino. Homological scaffold via minimal homology bases. *ArXiv*, 2021.

[12] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An introduction to statistical learning: with applications in r. *New York: Springer*, 2014. doi:10.1007/978-1-4614-7138-7.

[13] Hamid Krim Jennifer Gamble, Harish Chintakunta. Adaptive tracking of representative cycles in regularand zigzag persistent homology. *ArXiv*, November 2014.

[14] Jae-Hun Jung John Nicponski. Topological data analysis of vascular disease: A theoretical framework. *bioRxiv*, May 2019. URL: https://doi.org/10.1101/637090.

[15] Violeta Kovacev-Nikolic, Peter Bubenik, Dragan Nikoli´c, and Giseon Heo. Using persistent homology and dynamical distances to analyze protein binding. *ArXiv*, July 2015.

[16] James R. Munkres. *Elements of algebraic topology*. Addison-Wesley Publishing Company, Menlo Park, CA, 1984.

[17] Ippei Obayashi. Volume optimal cycle: tightest representative cycle of a generator on persistent homology. *ArXiv*, 2017.

[18] K.R. Parthasarathy. Probability measure on metric spaces. *Journal of the American Statistical Association*, 1968. doi:10.2307/2283907.

[19] G. Petri, P. Expert, F. Turkheimer, R. Carhart-Harris, D. Nutt, P. J. Hellyer, and F. Vaccarino. Homological scaffolds of brain functional networks. *Journal of The Royal Society Interface*, 11(101):20140873, 2014. doi:10.1098/rsif.2014.0873.

[20] Nicolas Courty Rémi Flamary, 2016. URL: https://pot.readthedocs.io/en/gromov/index.html.

[21] M. Shimrat. Algorithm 112: Position of point relative to polygon. *Commun. ACM*, 5(8):434, August 1962. doi:10.1145/368637.368653.

[22] John P. Snyder. Flattening the earth: Two thousand years of map projections. *Chicago: University of Chicago Press*, 1993.

[23] Larry Wasserman. All of nonparametric statistics. 2006.

[24] Frederic Chazal Yuhei Umeda, Merryll Dindin. Topological data analysis for arrhythmia detection through modular neural networks. *ArXiv*, 2019.

[25] Hideyuki Kikuchi Yuhei Umeda, Junji Kaneko. Topological data analysis and its application to time-series data analysis. *Fujitsu Scientific and Technical Journal*, 55(2), 2019.