UNIVERSITAT JAUME I

Bachelor's Degree in Computational Mathematics

Bachelor's Degree Final Project

# Introduction to spatio-temporal topological processes analysis

*Autor:*
Diego Galiana Safont

*Mentor:*
Jorge Mateu Mahiques

Reading Date: __ de _____ de 20__
Academic course 2023/2024

## Acknowledgements

Thanks to Anastasios Stefanou, who taught me at the University of Bremen what a simplex was in his 'Introduction to Applied Algebraic Topology' class. Although I failed his subject, I found it really interesting and challenging.

Thanks to Jorge Galindo Pastor, who taught me from scratch basic concepts of topology in his subject 'Differential Geometry and Topology' at the Jaume I University in Castellón. I managed with great effort to pass it but I enjoyed it a lot.

Thanks to Joan Porti and Martin Campos who at the Universidad Autónoma de Barcelona finally made me really understand what I missed in Germany in 2022 in their subject 'Topological Analysis of Data' and made me fall in love with this field of study.

Thanks to Jorge Mateu for being such a good tutor for my final project and helping me whenever I needed it, as well as giving me the opportunity to learn about such an interesting field as spatio-temporal analysis. I would also like to thank Jaime Gómez for providing useful ideas and recommendations.

Finally, I would also like to thank all the teachers I have had in my degree for training me professionally and Vicente Felip who made me like mathematics and studying this degree. I am also very grateful for the support and patience of my family and friends throughout my university adventure.

# Contents

# Chapter 1

# Introduction

We are at a point in history where knowing, collecting, processing and analysing large amounts of data is of vital importance if we want to be better prepared for any kind of situation, whether it is a for-profit company, a country that wants to offer a better service or a person with a specific personal interest.

While we now give data the importance it deserves, there is still a long way to go. While the biggest concern with data is how we can analyse more data in less time, we seem to have taken for granted that our analysis is perfect and that there is nothing more we can get. Like any science, the study of data is constantly evolving and is far from perfect or be finished. We must continue to investigate new ways of obtaining and relating data to obtain information with some kind of value.

Topological data analysis focuses on the geometric structure and spatial relationships that exist between data points. This allows the capture of non-linear relationships and allows the analysis to rely on global properties or connections between points rather than local measures such as mean or standard deviation, which facilitates the interpretation of complex patterns and relationships. One of its features is that its analysis is not limited by the specific nature of the data, making it compatible with spatio-temporal data.

Persistent homology studies the changes in the topological characteristics of a filtering of simplicial complexes and estimates their homology groups from a point cloud. The study of this application of algebraic topology allows us to identify robust patterns and meaningful structures under the presence of variability or noise in the data, in addition to the possibility of detecting persistent features in the data such as holes, which provides us with information about the connectivity and distribution of the data points.

An effective way to gain insights from data is through the visualization of persistence diagrams, as illustrated in Figure 1.1. This figure showcases three diagrams representing different stages of the data filtration process. By constructing a simplicial complex and varying a scale parameter, we capture the birth and death of topological features, such as connected components and holes. The persistence diagrams summarize these features, with the x-axis indicating birth and the y-axis indicating death. This iterative process enhances our understanding of the data's topological structure and helps identify robust features crucial for further analysis.

Spatio-temporal data have two dimensions, a spatial and a temporal component. The spatial component allows us to know the geographical distribution of the data while the temporal component permit us to track and capture changes over time. These properties are well suited to represent most real-world phenomena and allow us through the analysis of these data to understand trends, patterns and fluctuations of different phenomena, geographic variations and spatial interactions between different locations.

A point process is a mathematical model used to describe and analyze random events that occur in time, space, or both. It represents a collection of points, each corresponding to an event, distributed across a given domain, such as a timeline, a spatial region, or a combination of space and time. Key characteristics of point processes include their inherent randomness, where events happen unpredictably, and their intensity, which denotes the rate at which events occur. This intensity can be constant across the domain (homogeneous process), or vary depending on location or time (inhomogeneous process). Point processes can model independent events or capture dependencies between them, making them versatile tools for studying phenomena in various fields, such as ecology, economics or criminology.

The study of the shape of data is not new and has been questioned many times for its real usefulness. During this thesis we will introduce two branches of analysis, one more traditional and focused on global properties as the analysis of spatio-temporal data using point processes and one more innovative as the persistent homology analysis. Later, by means of a study of emergency call data in the city of Valencia, we will put into practice the concepts introduced and we will relate these two ways of analysing data to demonstrate the usefulness and importance of topological analysis.

In recent years, Topological Data Analysis (TDA) has gained significant traction across various fields, demonstrating its versatility in uncovering hidden patterns in complex datasets. Chandola and Kumar [1] applied TDA to anomaly detection in network traffic, revealing insights that traditional methods often miss, particularly in cybersecurity contexts. Turner, Mukherjee, and Boyer [2] provided a comprehensive survey of TDA's integration into machine learning, highlighting its role in enhancing predictive models. Meanwhile, Bubenik [3] advanced the field with his work on Persistence Landscapes, a tool that summarizes topological features across scales, proving essential in areas ranging from biology to medical imaging.

Exploring the relationship between Topological Data Analysis (TDA) and point processes in spatial-temporal data is a burgeoning area of research that promises significant insights. Recent studies have made notable advancements in this field. For example, Ghrist [4] delves into how topological methods can be applied to spatial data analysis, offering new perspectives on data connectivity and structure. Nelson et al. [5] extend this exploration to temporal data, demonstrating how TDA can capture dynamic and temporal features. Additionally, Pratola et al. [6] investigate the application of persistent homology to spatial point processes, revealing how topological techniques can enhance the understanding of point distributions in both spatial and temporal contexts. These contributions underscore the growing importance of integrating TDA with spatial-temporal analysis, highlighting its potential to uncover complex patterns and relationships in such data.

In the second chapter, we focus on spatio-temporal point processes, while in the third, we explore the field of persistent homology. Finally, in chapter three, we introduce a practical application of the theory and present a few conclusions. We hope you enjoy this journey through data analysis and that it inspires many of you to reconsider the way we analyze data.

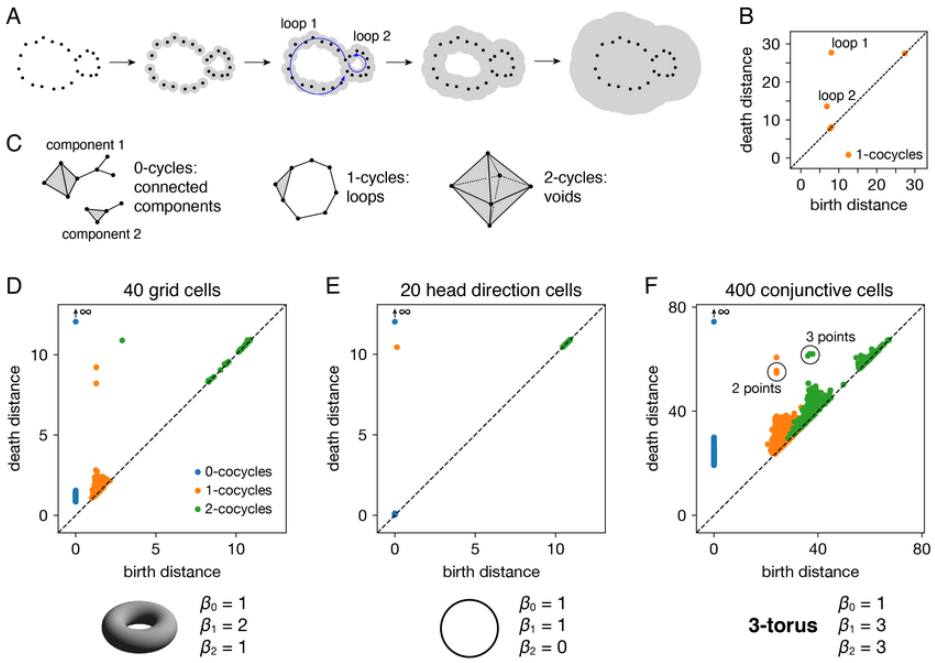


Figure 1.1: Persistence diagrams.

# Chapter 2

# Spatio-Temporal Point Processes

The purpose of this chapter is to cover all the basic definitions of spatio-temporal point processes. We have divided the content into three sections. This chapter is primarily based on the definitions and concepts presented in the comprehensive review by González et al. [7]. For further details and a thorough review of the concepts, we refer the reader to this seminal work.

## 2.1 Basic characteristics

In this chapter, we focus on defining the fundamental theoretical concepts of spatio-temporal point processes, which are essential for modeling and analyzing data that varies across space and time. We will cover concepts such as the intensity function, stationarity, and isotropy, which form the theoretical foundation for models and analysis in this field. This theoretical knowledge is crucial for understanding the structure of spatio-temporal data and sets the stage for the estimation methods that we will explore in the next chapter.

**Definition 2.1.1** Let W $\subseteq \mathbb{R}^2$ be a subset and T $\subseteq \mathbb{R}$ be an interval. We consider as a **spatio-temporal point pattern data** a collection of points $\{X_i\}_{i=1}^n = \{(u_i, v_i)\}_i^n \subseteq$ W x T. We also can define it as a random countable subset of $\mathbb{R}^2 \times \mathbb{R}$ such that $|X \cap (\text{A} \times B)| < \infty$.

We denote $X$ as the point process, and the observations of this process are represented as $X_i$.

**Definition 2.1.2** We define B[(u,v), r, t] centred at (u,v) ∈ W x T with spatial radius $r > 0$ and temporal radius $t > 0$ as the **cylindrical neighbourhood**:

$$B[(u,v),r,t] = B[u,r] \times [v-t,\ v+t] = \{(a,b) \in W \times T : \|u-a\| \leq r,\ |v-b| \leq t\},$$

where $B[u,r] = \{a \in W : \|u-a\| \leq r\}$ is the euclidean ball centred at u ∈ W with radius r.

**Definition 2.1.3** Let A ⊆ W and B ⊆ T. We denote the **number of points of a set** (A × B) ∩ X as N(A × B). If N(W × T) < ∞ with probability one, we call X a finite spatio-temporal point process.

**Definition 2.1.4** Let X be a spatio-temporal point process on $W \times T \subseteq \mathbb{R}^2 \times \mathbb{R}$. We say that X is **stationary** if (u, v) + X has the same distribution as the original process X ∀ (u,v) ∈ $W \times T$.

**Definition 2.1.5** Let X be a spatio-temporal point process on $W \times T \subseteq \mathbb{R}^2 \times \mathbb{R}$. We call X **isotropic** if rX = $\{(ru, v) : (u.v) \in X\}$ has the same distribution as the original process X ∀ rotation r around the origin.

Note that we can define explicit spatial or temporal stationarity assuming only translations such as (u,0)+X,u ∈ $\mathbb{R}^2$, or (0,v)+X,v ∈ $\mathbb{R}$.

**Definition 2.1.6** If X is a finite spatio-temporal point process, we can project X onto W and T, dealing with the **space and time components of X separately**:

$$X_{space} = \{u : (u,v) \in X, v \in T\}, \quad X_{time} = \{v : (u,v) \in X, v \in W\}$$

**Definition 2.1.7** Let $k \geq 1$, we can define the **product densities** $\lambda^k$ through the Campbell theorem. Let X be a spatio-temporal point process and h any non-negative function on $(\mathbb{R}^2 \times \mathbb{R})^k$ we have that:

$$\mathbb{E}[\sum_{A_1,...,A_k \in X}^{\neq} h(A_1,...,A_k)] = \int_{\mathbb{R}^2 \times \mathbb{R}} \int_{\mathbb{R}^2 \times \mathbb{R}} h(A_1,...,A_k)\lambda^k(A_1,...,A_k) \prod_{i=1}^{k} dA_i$$

where $\sum^{\neq}$ express that the summation is taken over distinct k-tuples of spatio-temporal events. We also recall that the left side is infinite if and only if the right side is.

**Definition 2.1.8** Let A × B ⊆ W × T, we define the **intensity measure** as:

$$\mu(A \times B) = \mathrm{E}[\mathrm{N}(A \times B)]$$

**Definition 2.1.9** If $\lambda = \lambda^{(1)}$ exists, we can assume that:

$$\mu(A \times B) = \int_A \int_B \lambda(u,v)dudv$$

and we refer to $\lambda(u, v)$ as the **intensity function** of X.

**Definition 2.1.10** When X is stationary, then $\lambda(u, v)$ is constant, positive and referred to as the **intensity** of X.

**Definition 2.1.11** If first-order intensity function of a spatio-temporal point process can be factorised (almost everywhere) as $\lambda(u, v) = \lambda_1(u)\lambda_2(v)$ through which:

$$\mu(A \times B) = \int_A \lambda_1(u)du \int_B \lambda_2(v)dv$$

where $\lambda_1(\cdot)$ and $\lambda_2(\cdot)$ are non-negative functions, then we say that the process is **first-order spatio-temporal separability**.

Note that a spatio-temporal point process X is first-order separable since its intensity is constant. If X is space-stationary such that $\lambda(u, v)$ depends only on v, it is also first-order separable with $\lambda_1$ being a non-negative constant. If X is time-stationarity such that $\lambda(u, v)$ depends only on u, it is too first-order separable with $\lambda_2$ being a non-negative constant.

**Definition 2.1.12** Let $X_{space}$ and $X_{time}$ , we can define $\lambda_{space}$ and $\lambda_{time}$ as the **marginal spatial and temporal intensity functions**:

$$\lambda_{space}(u) = \lambda_1(u) \int_T \lambda_2(v)dv \text{ and } \lambda_{time}(v) = \lambda_2(v) \int_W \lambda_1(u)du$$

by which $\lambda(u, v)$ proportional to $\lambda_{space}(u)\lambda_{time}(v)$ with $\lambda, \lambda_{space}, \lambda_{time}$ all being constant when X is stationary.

**Definition 2.1.13** We define the **history** $\mathcal{H}_v$ as the family of $\sigma$-algebras generated by the events occurring at times up to, but not including v.

**Definition 2.1.14** We define the **conditional intensity function** $\lambda^*$(u, v| $\mathcal{H}_v$) of a spatio-temporal point process as the expected rate that points occur around the spatio-temporal location (u, v), conditionally on the history $\mathcal{H}_v$, v $\in$ T, consisting of the set of locations and times of all events of the process that occur prior to time v:

$$\lambda^*(u, v \mid \mathcal{H}_v)dudv = \mathbb{E}[N(du \times dv) \mid \mathcal{H}_v], \quad (u, v) \in du \times dv \subseteq W \times T.$$

assuming that the process X is orderly, which means that the probability of observing more than one point in a time interval is decreasing with the order of the size of the interval. Denoting by l+V and U the time and location, respectively, of the first event that occurs after time l, it follows that

$$\mathbb{P}(V > v) = exp\left\{-\int_l^{l+v} \int_W \lambda^*(u, t \mid \mathcal{H}_l)dudt\right\}$$

the conditional probability density of U given V = v is proportional to the conditional intensity, $\lambda^*(u, l+v \mid \mathcal{H}_{l+v}), u \in W$.

We denote an edge-correction factor by a weight $w_{ij}$, where i and j represent two different points of the pattern.

**Definition 2.1.15** Turning to measures of second-order spatio-temporal interaction, we can define the **pair correlation function** in presence of inhomogeneity as:

$$g(\xi_1, \xi_2) = \frac{\lambda^{(2)}(\xi_1, \xi_2)}{\lambda(\xi_1)\lambda(\xi_2)}, \quad \xi_1, \xi_2 \in W \times T.$$

if the process is completely random (Poisson) the pair correlation is 1, larger or smaller values will indicate how much more or less likely it is that a pair of events will occur at the specified locations than in a Poisson process with the same intensity function.

**Definition 2.1.16** The pair correlation function is **separable** if $g((u, v), (s, l)) = g_1(u, s)g_2(v, l)$ where $g_1$ and $g_2$ are non-negative functions.

**Definition 2.1.17** Let X be a spatio-temporal point process and $\bar{g}, g_0$ be some non-negative functions. We say that X is **second-order intensity-reweighted stationary** (SOIRS) if g((u,v), (s,l)) = $\bar{g}$(u-s, v-l) $\forall$ (u,v), (s,l) $\in$ W $\times$ T.

If X is also isotropic, then $\bar{g}$(u-s, v-l) = $g_0$(r,t) such that g($\cdot$,$\cdot$) depends only on the distances r = $\|$u-s$\|$ and t = |v-l|.

**Definition 2.1.18** Let X be a SOIRS spatio-temporal point process, u= u-s, v= v-l and r and t be positive constants, we can define the **inhomogeneous spatio-temporal K-function** as

$$K_{inhom}(r, t) = \int_{\mathbb{R}^2} \int_{\mathbb{R}} 1\{(u, v) \in B_{rt}\} \bar{g}(u, v) du dv.$$

If also X is isotropic, we can define the spatio-temporal inhomogeneous K-function as:

$$K_{inhom}(r, t) = 2\pi \int_0^r \int_{-t}^t s g_0(s, l) ds dl$$

for a spatio-temporal Poisson process we have that $K(r, t) = \pi r^2 t$ such that $K(r, t) - \pi r^2 t$

can be used as a meassure of spatio-temporal aggregation of regularity (variations in the destiny of points that cannot be explained by inhomogeneity alone).

**Definition 2.1.19** Let

$$p_1(u) = \frac{\lambda_1(u)}{\int_W \lambda_1(u)du} \quad and \quad p_2(v) = \frac{\lambda_2(v)}{\int_T \lambda_2(v)du}$$

such that

$$g_{space}(u,s) = g_{space}(u-s) = \int_T \int_T p_2(v)p_2(l)g(u-s, v-l)dvdl,$$
$$g_{time}(v,l) = g_{time}(v-l) = \int_W \int Wp_1(u)p_1(s)g(u-s, v-l)duds$$

if $X_{space}$ and $X_{time}$ are defined, under the assumption of separability we can write the **spatial and temporal components of the K-function** as:

$$K_{space}(r) = \int_{\|u\| \leq r} g_{space}(u)du \; and \; K_{time}(t) = \int_{-t}^{t} g_{time}(v)dv, \; r, t > 0.$$

Now we have established the key theoretical concepts necessary for analyzing spatio-temporal point processes. These theoretical foundations are indispensable for tackling the practical problem of estimating unknown parameters from observed data, which is the focus of the next chapter. With these concepts in mind, we will be able to develop and apply estimation methods that align theoretical models with real-world data, enabling a practical and robust application of the theory in real situations (see Figure 2.1).

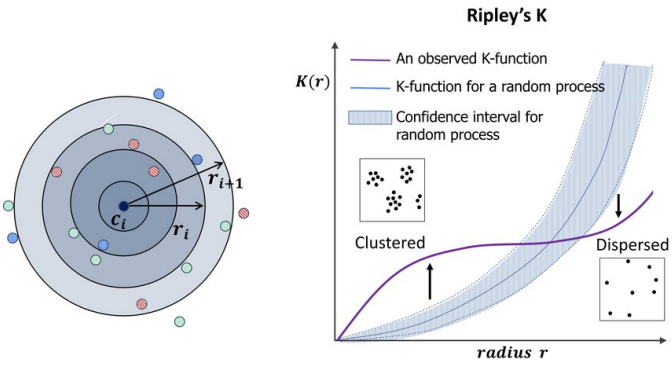


Figure 2.1: Ripley's Spatial K-Function: Comparison Between Observed and Random Process.

## 2.2 Estimation Methods

In this section we will focus on estimation methods that allow us to apply the previously defined theoretical concepts to real data. In practice, we do not have the exact theoretical parameters, but rather observed data from which we need to estimate these parameters. This chapter explores various techniques, such as non-parametric estimators and kernel-based methods, that allow us to approximate the unknown parameters and thus fit the theoretical models to the observed reality.

**Definition 2.2.1** Let $\hat{\lambda}_{space}(\cdot)$ and $\hat{\lambda}_{time}(\cdot)$ be unbiased estimators. If we assume separability, we define the **first-order spatio-temporal intensity function estimation** as:

$$\hat{\lambda}(u, v) = \frac{1}{n} \left( \hat{\lambda}_{space}(u) \hat{\lambda}_{time}(v) \right)$$

This is also an unbiased estimator of the expected number of points.

**Definition 2.2.2** Let $\varepsilon$ be a positive smoothing parameter called bandwidth. Let $\kappa(\cdot)$ be a bivariate kernel such that

$$\kappa_\varepsilon(u) = \frac{1}{\varepsilon^2} \kappa(\frac{u}{\varepsilon}),$$

and let $C_{W\varepsilon}(u_i)$ be an edge-correction factor to guarantee $\int_W \hat{\lambda}_{space}(u) du = n$ such that

$$C_{W\varepsilon}(u_i) = \int_W \kappa_\varepsilon(u - u_i) du.$$

We define the **non-parametric estimation of the spatial intensity function** using a kernel estimation:

$$\hat{\lambda}_{space}(u) = \sum_{i=1}^{n} \frac{\kappa_\varepsilon(u - u_i)}{C_{W\varepsilon}(u_i)}, \quad u \in W.$$

In the same way, we can also estimate $\lambda_{time}(v)$ non-parametrically by kernel estimators. Note that the estimation depends on the bandwidth and there is no magic constant value to get always the best approximation. Small values usually leads to noisy and unrealistic estimations, while large values produce smooth estimations.

We assume that X is a spatio-temporal point process on W $\times$ T $\subseteq \mathbb{R}^2 \times [0, \infty]$ such that $X_{time}$ is well defined. X can be treated as a temporal point process with corresponding marks $X_{space}$ and we can define the cumulative process as $X_{time}(t) := | X_{time} \cap [0, t] |$, t $\in$ T. This approach for spatio-temporal point processes is the classical one.

**Definition 2.2.3** Let $w_{ij}^{(u)}$ be the proportion of the circumference of a circle centred at $u_i$ with radius $\parallel u_i - u_j \parallel$ that lies within W, we define the **isotropic correction** as:

$$w_{ij} =\mid W \times T \mid w_{ij}^{(u)} w_{ij}^{v}$$

here the weight is proportional to the product between its one-dimensional analogue and the ripley edge-correction factor for the spatial region. The temporal edge-correction factor $w_{ij}^{(v)}$ takes the value of 1 if both ends of the interval of length $2\mid v_i - v_j$ that is centred at $v_i$ lie within T, or otherwise 1/2.

**Definition 2.2.4** Let $W_{\ominus r} = \{u \in W : B[u,r] \subseteq W\}$ and $T_{\ominus t} = \{v \in T : B[v,t] \subseteq T\}$ be eroded spatial and temporal regions, obtained by trimming off a margin of width $r \geq 0$ and $t \geq 0$ from the borders of W and T, respectively. We can define the **border method** as

$$w_{ij} = \frac{\sum_{j=1}^{n} 1\{(u_j, v_j) \in W_{\ominus r} \times T_{\ominus t}\} \diagup \lambda(u_j, v_j)}{1\{(u_i, v_i) \in W_{\ominus r} \times T_{\ominus t}\}}, \quad r, t \geq 0$$

Note that $W_{\ominus r} \times T_{\ominus t}$ may be visualised by taking the flat trimmed region $W_{\ominus r} \times \{0\}$ and stretching it in the t-dimension until its height reaches $\mid T \mid -2t$. This method restricts attention to those events lying more than r units away from the boundary of W and more than t units away from the boundary of T.

**Definition 2.2.5** Let $X = \{(u_i, v_i)\}_{i=1}^{n}$. Let $\kappa_{1\varepsilon}$ and $\kappa_{2\delta}$ be one-dimensional kernel functions with spatial and temporal bandwidths $\varepsilon$ and $\delta$ respectively. Let $w_{ij}$ be edge-correction factors, which correct for the loss of information regarding the interaction occurring between points close to the border of W $\times$ T and those (unobserved) ones outside. We can define the **non-parametric estimation of the pair correlation functions** based on kernel methods as:

$$\hat{g}(r,t) = \frac{1}{4\pi r} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{\kappa_{1\varepsilon}(\parallel u_i - u_j \parallel - r)\kappa_{2\delta(|v_i - v_j| - t)})}{\hat{\lambda}(u_i, v_i)\hat{\lambda}(u_j, v_j) w_{ij}}, \; r > \varepsilon, t > \delta.$$

**Definition 2.2.6** Let $X = \{(u_i, v_i)\}_{i=1}^{n}$ we can define the **general** $K_{inhom}(r,t)$ **estimator** as

$$\hat{K}(r,t) = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{1[\parallel u_i - u_j \parallel \leq r] 1[\mid v_i - v_j \mid \leq t]}{\hat{\lambda}(u_i, v_i)\hat{\lambda}(u_j, v_j) w_{ij}}$$

if $w_{ij}$ is Ripley's spatial edge-correction factor, we obtain an approximately unbiased non-parametric estimator of $K_{inhom}(r,t)$.

Let $n_t$ be the number of events $v_i \leq b - t$ whenever T=[a,b] $\subseteq \mathbb{R}_+$, we can define alternatively the past estimation without taking into account the past of the process:

$$\hat{K}^*(r,t) = \frac{n}{n_t} \sum_{i=1}^{n_t} \sum_{j>i} \frac{1[\| u_i - u_j \| \leq r] 1[v_i - v_j \leq t]}{\hat{\lambda}(u_i, v_i)\hat{\lambda}(u_j, v_j) w_{ij}}.$$

We begin with an example of a spatial pair correlation function (see Figure 2.2), utilizing a dataset of emergency calls from the city of Valencia in 2015, which serves as part of the practical case study in Chapter 3. This analysis is spatial in nature, as we exclusively use spatial data, specifically the longitude and latitude coordinates from the dataset. The resulting graph includes the Poisson pair correlation function as a reference, revealing that the calculated pair correlation function exhibits higher values. This observation indicates, as explained in Figure 2.1, that the K function is another type of graph that illustrates a similar degree of data aggregation but in a different way.

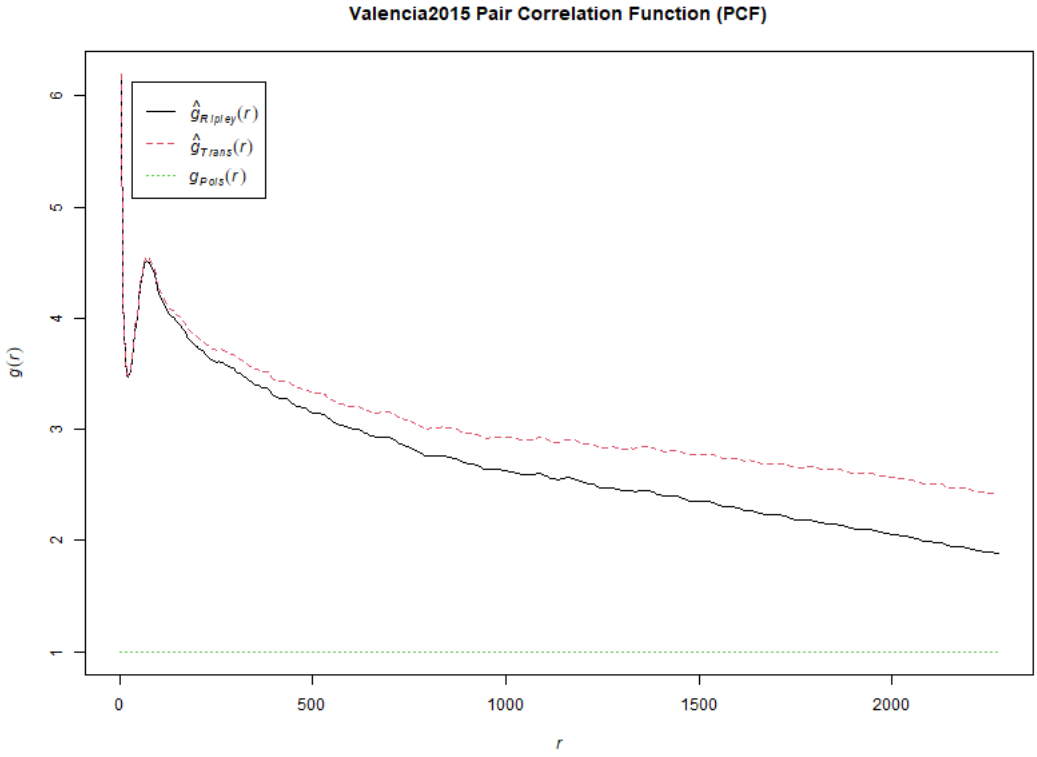


Figure 2.2: Pair Correlation Function Example.

We can extend the calculation of the pair correlation function to spatio-temporal data by incorporating a temporal component. In our previous example using the emergency call dataset from Valencia in 2015, we now include the variable representing days, which ranges from 1 to 365 for the entire year. This allows us to generate a spatial pair correlation function for each temporal value. The resulting pair correlation function illustrates the spatial distribution of the data across different days, similar to the example shown in Figure 2.2, which displays the spatial pair correlation function. A comprehensive example of the spatio-temporal pair correlation function can be observed in Figure 2.3.

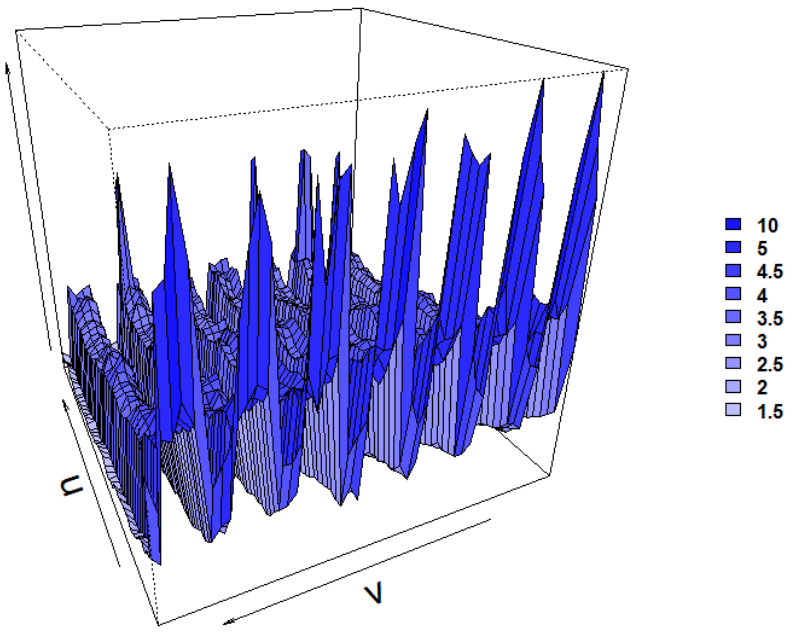


Figure 2.3: Spatial-Temporal Pair Correlation Function Example.

With these estimation tools, we are now ready to move on to the concrete modelling of spatio-temporal processes in the next chapter.

## 2.3 Spatio-Temporal models

In the previous chapters, we have established the fundamental theoretical concepts and estimation methods needed to understand spatio-temporal point processes. With this foundation, we are ready to explore specific models that capture the complexity of spatio-temporal phenomena in practice. In this chapter, we will focus on spatiotemporal Poisson processes, both homogeneous and inhomogeneous, and move towards more sophisticated models such as Cox and Hawkes processes.

**Definition 2.3.1** Let $\lambda$ be a positive constant that represents intensity, we define a **spatio-temporal homogenous Poisson process** as a spatio-temporal point process X satisfying that:

i. Given $A_1 \times B_1, ..., A_m \times B_m \subseteq W \times T$, the corresponding random variables $N(A_1 \times B_1), ..., N(A_m \times B_m)$ follow independent Poisson distributions with the respective means $\mu(A_i \times B_i) = \lambda \mid A_i \times B_i \mid, i = 1, ..., m$.

ii. Points included in $A \times B$ form an independent random sample from the uniform distribution on $A \times B$ (conditioned on $N(A \times B)$).

iii. All product densities exist and $\lambda^{(k)}(\xi_1, ..., \xi_k) \equiv \lambda^k,\ k \geq 1$.

Poisson processes are the starting point for spatio-temporal point pattern data models, they are rarely realistic but provide an approximation of complete spatio-temporal randomness (CSTR).

**Definition 2.3.2** We define a **spatio-temporal inhomogeneous Poisson process** as:

i. Given any disjoint $A_1 \times B_1$,...,$A_m \times B_m \subseteq$ w $\times$ T, the corresponding random variables $N(A_1 \times B_1), ..., N(A_m \times B_m)$ follow independent Poisson distributions with the respective means

$$\int_{A_i} \int_{B_i} \lambda(u, v) du dv,\ i = 1, ..., m.$$

ii. Given $N(W \times T) = n$, the n events in $W \times T$ form an independent random sample from the distribution on $W \times T$ which has density function

$$f(u, v) = \frac{\lambda(u, v)}{\int_W \int_T \lambda(u, v) du dv}.$$

We obtain the homogeneous Poisson process by setting $\lambda \equiv \lambda > 0$. Similarly to the homogeneous case, it follow that the product densities exist and are given by

$$\lambda^{(k)}((u_1, v_1), ..., (u_k, v_k)) = \prod_{i=1}^{k} \lambda(u_i, v_i),\ k \geq 1.$$

Spatio-temporal inhomogeneous Poisson process is the simplest non-stationary spatio-temporal point process, note that it is obtained by replacing the constant intensity of a homogeneous Poisson process by a spatially and/or temporally varying intensity function $\lambda(u, s), (u, s) \in W \times T$.

We present a graphical example of an inhomogeneous spatio-temporal Poisson process in Figure 2.4. In this example, we illustrate three independent processes that vary from fewer to more points. The figure provides both a two-dimensional (spatial) perspective and a three-dimensional (spatio-temporal) view, allowing for a comprehensive understanding of how point density and distribution change over time and space. This visualization underscores the characteristics of inhomogeneous processes, where the intensity of points is not uniform, revealing insights into the underlying structure of the data.

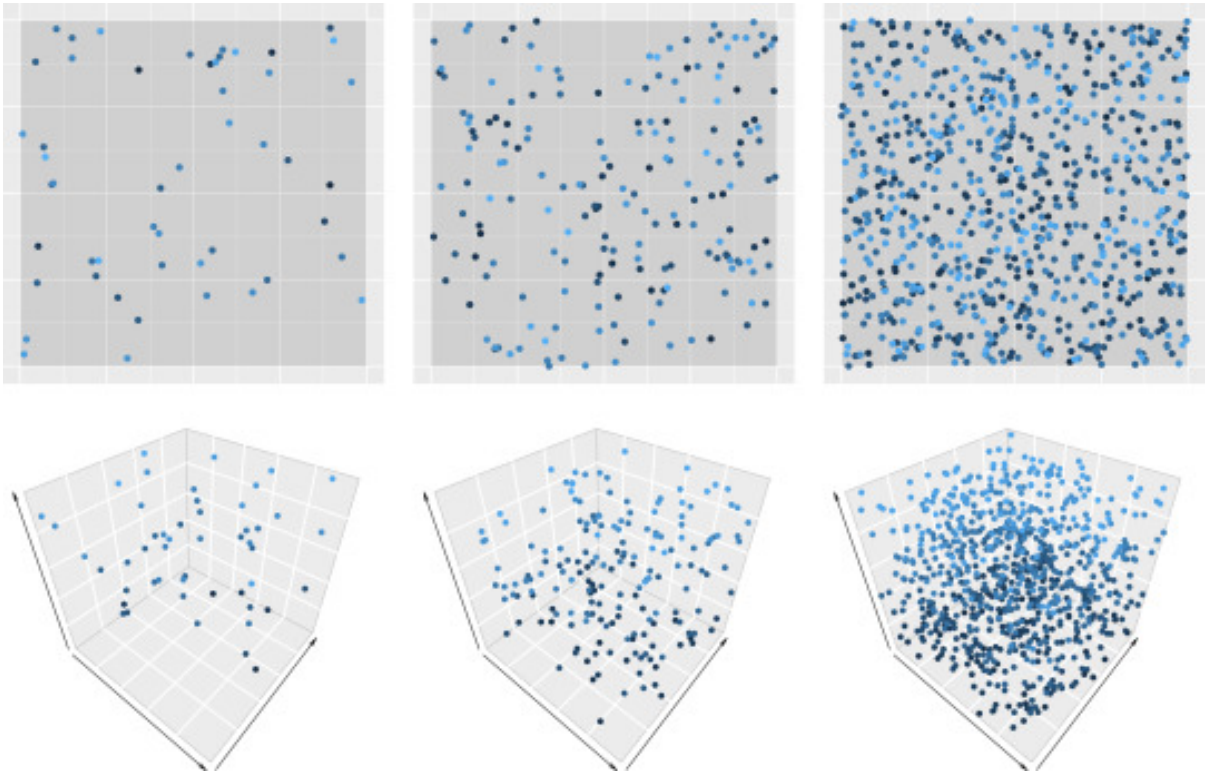

Figure 2.4: Inhomogeneous Poisson Process Example.

**Definition 2.3.3** Let the likelihood be as the probability of obtaining a given number of points in the spatio-temporal observation window, times the joint conditional density for the locations of those points, given their number. Suppose that there are n observations on $W \times T$ at spatio-temporal points $\{(u_i, v_i)\}_{i=1}^n$. Since the distribution of the number of points in Poisson, then the probability of obtaining single points in some differential volume $\triangle$ centred at $(u_i, v_i)$ and no points on the remaining part of $W \times T$ is given by

$$exp\left\{-\int_W \int_T \lambda(u,v)dudv\right\} \prod_{i=1}^n \lambda(u_i, v_i) \triangle .$$

Hence, dividing by $\triangle^n$, letting $\triangle \to 0$ and taking logs, we have that the **log-likelihood** for $\lambda(\cdot, \cdot)$ based on data is given by

$$L(\lambda) = \sum_{i=1}^n log\lambda(u_i, v_i) - \int_W \int_T \lambda(u,v)dudv.$$

In practice, it is particularly useful if $\lambda(u, v)$ can be specified through a regression model, for example let $z_j(u, v)$ be covariates that may vary in space and time, consider

$$log\lambda(u,v) = \sum_{j=1}^p \beta_j z_j(u,v).$$

**Definition 2.3.4** We define a **spatio-temporal Cox process** as a process that satisfies:

i. $\{\Lambda(u, v) : (u, v) \in \mathbb{R}^2 \times \mathbb{R}\}$ is non-negative-valued stochastic process.

ii. Conditionally on $\{\Lambda(u, v) = \lambda(u, v) : (u, v) \in \mathbb{R}^2 \times \mathbb{R}\}$, the events form an inhomogeneous spatio-temporal Poisson process with intensity function $\lambda$(u,v).

Cox processes are natural models for point patterns that are thought to be determined by environmental variability, we can think about it like a "doubly stochastic" process formed as an inhomogeneous Poisson process with an intensity function coming from some stochastic mechanism.

Moment properties of a Cox process are inherited from those of the process $\Lambda(u, v)$ and thus first and second order properties are obtained from those of the inhomogeneous Poisson process by taking expectations with respect to $\Lambda(u, v)$. Assuming that the covariance structure $\gamma(r, t) = Cov\{\Lambda(u_1, v_1), \Lambda(u_2, v_2)\}$, for r = $\parallel u_1 - u_2 \parallel$ and t = $\mid v_1 - v_2 \mid$, is stationary, a convenient reparametrisation is $\Lambda(u, v) = \lambda(u, v)S(u, v)$, where S(u,v) is a stationary process with expectation 1 and covariance function $\gamma(r, t) = \sigma^2$ is the variance of S(u,v) and s$(\cdot,\cdot)$ is a spatio-temporal correlation function. It follows that $\lambda(u, v)$ is the first-order intensity of the point

process, and the stationarity of S(u,v) implies that the point process is intensity-reweighted stationary.

Given the last parametrisation, we can say that $\Lambda(u, v)$ is first-order separable if $\lambda(u, v) = \lambda_1(u)\lambda_2(v)$ holds, and second-order separable if $\gamma(r, t) = \sigma^2 s_1(r)s_2(t)$, where $s_1(r)$ and $s_2(t)$ can be chosen as any pair of valid correlation functions in $\mathbb{R}^2$ and $\mathbb{R}$, respectively. The assumption of second-order separability is more difficult to deal with, but undeniably convenient.

The K-function of an intensity-reweighted stationary Cox process (with the last parametrisation) is given by

$$K(r, t) = \pi r^2 t \ + \ 2\pi\lambda^{-2}\sigma^2 \int_0^t \int_0^r xs(x, y)\ dx\ dy.$$

For a Poisson process, by the independence of the events, we can obtain that

$$\lambda^*(u, v \mid \mathcal{H}_v)dudv = \mathbb{E}[N(du \times dv) \mid \mathcal{H}_v] = \mathbb{E}[N(du \times dv)] = \lambda(u, v)dudv,$$

the conditional intensity function and the first-order intensity function are the same. Note that the intensity function can be estimated either by smoothing the observations or by fitting some parametric model, and here the situation is similar. In the conditional intensity setting, the non-parametric estimation procedure of the first-order intensity is directly followed.

**Definition 2.3.5** A spatio-temporal point process X is called **self-exciting** if $Cov[N(A \times B), N(A \times B + (u, v))] > 0$ for small values of (u,v). Stationary spatio-temporal point processes are sometimes described by the covariance between the number of points in some spatio-temporal regions $A \times B$ and $A \times B + (u, v)$ ($A \times B$ shifted by (u,v)). If such covariance is negative, the process is **self-correcting**. Thus the occurrence of points in a self-exciting point process causes other points to be more likely to occur in space-time, whereas in a self-correcting process, the points have an inhibitory effect, these models are commonly used in seismology.

**Definition 2.3.6** Let $X = \{(u_i, v_i)\}$ be a Poisson cluster process with events $(u_i, v_i) \in \mathbb{R}^2 \times \mathbb{R}$. Let the cluster centres of X given by certain events be called 'immigrants' and the other events known as 'offspring'. A **spatio-temporal Hawkes process** X satisfies that:

i. All immigrants follow a Poisson process with intensity function $\psi(u, v)$.

ii. Each immigrant $(u_i, v_i)$ generates a cluster $C_i$, which consists of events of generations of order $m = 0, 1, \ldots$ with the following branching structure. First we have $(u_i, v_i)$, which is said to be of generation 0. Given the 0,...,m generations in $C_i$, each $(u_i, v_i) \in C_i$ of generation $m$ recursively generates a Poisson process $X_j$ of offspring of generation $m + 1$ with intensity function $\kappa_i(u, v) = \kappa(u - u_i, v - v_i)$. Here, $\kappa$ is a non-negative function defined on $(0, \infty)$.

iii. Given the immigrants, the clusters are independent.

iv. X consists of the union of all clusters.

v. The conditional intensity function of the Hawkes process, given the history $\mathcal{H}_t$ up to time $t$, is defined as:

$$\lambda^*(u, v \mid \mathcal{H}_v) = \psi(u, v) + \sum_{i:v_i<v} \kappa_i(u, v),$$

where $\psi(u, v)$ is the baseline intensity function representing the immigrant process, and $\kappa_i(u, v) = \kappa(u - u_i, v - v_i)$ represents the influence of past events (offspring) on the current rate of occurrence.

These approaches allow us to address situations where the intensity of events varies in space and time, or where events are correlated with each other. Such techniques are fundamental for the analysis of spatio-temporal point patterns in environments where the assumption of homogeneity is unrealistic and where interactions between events play a crucial role. By understanding and applying these methods, we are better equipped to capture and analyze the inherent complexity of spatio-temporal phenomena in real contexts. This knowledge is essential for practical applications that require accurate modeling of observed data in space and time.

# Chapter 3

# Persistent Homology

The aim of this chapter is to provide a general but solid background on persistent homology. We have divided the chapter into three parts.

## 3.1 Simplicial complexes

The first section is intended to relate the topological spaces to simplicial complexes, defining the bases on which persistent homology is built. This relation allows us to analyse a topological space from a simplicial point of view, applying the tools and techniques of simplicial analysis that we will see in the next section.

**Definition 3.1.1** We recall that a finite set of vectors $\{v_0, ..., v_n\}$ from a vector space is said to be **linearly independent**, if there only exist all zero scalars $a_0, ..., a_n$ / $a_0v_0$+...+$a_nv_n$ = 0 (defining 0 as the null vector).

**Definition 3.1.2** A finite subset $\{v_0, ..., v_n\}$ of points in $\mathbb{R}^m$ is called to be **affine independent** if the set of vectors $\{v_1\text{-}v_0,...,v_n\text{-}v_0\}$ is linearly independent.

**Definition 3.1.3** Let n,m $\in \mathbb{R}$ / m$\geq$n. A **n-simplex** is the convex hull of an affine independent set of n+1 points $\{v_0, ..., v_n\} \in \mathbb{R}^m$:

$$\left\{ \sum_{i=0}^{n} a_i v_i \ / \ \sum_{i=0}^{n} a_i = 1, \ a_i \geq 0 \ \forall \ i \ = \ 0, ..., n \right\}$$

Figure 3.1 provides a graphical representation of various types of $n$-simplices, illustrating the concept defined in Definition 3.1.3. This example encompasses simplices ranging from $n = -1$ to $n = 3$. Specifically, it includes vertices (0-simplexes), line segments (1-simplexes), triangles (2-simplexes), and tetrahedra (3-simplexes). Each simplex is formed by the convex hull of an affine independent set of $n + 1$ points in $\mathbb{R}^m$. This visualization aids in understanding the geometric properties and relationships between simplices, emphasizing their role in the broader context of topological analysis.

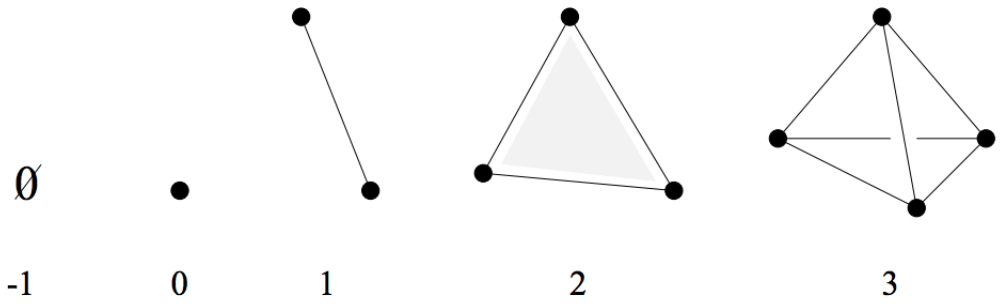


Figure 3.1: Examples of simplexes.

**Definition 3.1.4** Given a simplex $[v_0, ..., v_n]$, we define a **l-face** of the simplex as the l-simplex generated by l+1 points $[v_{i_0},...,v_{i_l}]$ $\forall$ i = 0,...,n / n > l. Every simplex generated by a not empty subset of $\{v_0, ..., v_n\}$ will be a face.

**Definition 3.1.5** A **simplicial complex K** is a set of simplices such that:

1. Let $\alpha \in$ K be any n-simplex and $\beta$ be any face of $\alpha$, then all $\beta \in$ K.

2. Let $\alpha_1$, $\alpha_2 \in$ K be any n-simpleces, then $\alpha_1 \bigcap \alpha_2$ is a face of both $\alpha_1$ and $\alpha_2$ or empty.

**Definition 3.1.6** Let X be a set and P(X) be the family of all subsets of X. We define the topology on X as a collection T $\subset$ P(x) of subsets of X such that:

1. X, $\emptyset \in$ T

2. Let $U_i \in$ T, then:
$$\bigcup_{i \in I} U_i \in \text{T}, \ \forall i \in I$$

3. Let $U_1, U_2 \in$ T, then:
$$U_1 \cap U_2 \in \text{T}$$

We call the pair (X, T) a **topological space** and the sets U $\in$ T open sets.
**Definition 3.1.7** Let K be a simplicial complex and $\alpha$ any n-simplex of K. We define the **simplicial polyhedron** as:
$$|K| := \bigcup_{\alpha \in K} \alpha$$
Note that $|K|$ is a topological space with the usual induced topology.

**Definition 3.1.8** Let $f, g$: X $\to$ Y be two continuous functions. A **homotopy** between f and g its the application H: X x [0,1] $\to$ Y that satisfies:

1. H is continuous

2. H(x,0) = f(x) $\forall$ x $\in$ X

3. H(x,1) = g(x) $\forall$ x $\in$ X

If f and g are homotopic, we write f $\backsimeq$ g.

**Definition 3.1.9** Let $f$: X $\to$ Y be a function between topological spaces. We denote $f$ as an **homeomorphism** if:

1. $f$ is continuous

2. $f$ is bijective

3. Inverse function $f^{-1}$: Y $\to$ X is continous.

We denote $A \cong B$ when $\exists$ a hemeomorphism between topological spaces. Also note that two homeomorphic spaces have the same type of homotopy (if $A \cong B$ then $A \backsimeq B$).

**Definition 3.1.10** Let X be a topological space, K a simplicial complex and $|K|$ its simplicial polyhedron. A topological space **triangulation** is a simplicial complex such that $|K| \cong$ X.

**Theorem 3.1.11** All compact surfaces are **triangulable**.

A sphere, a cylinder, and a torus are all examples of compact surfaces, which means they have a finite extent and are closed. Due to their topological properties, these surfaces are also triangulable, allowing them to be represented using triangular meshes. This triangulation process is essential for various applications in computational geometry and computer graphics. In Figure 3.2, we can see an example of the triangulation of a torus in three dimensions, illustrating how this surface can be divided into a set of triangles that maintain its geometric structure.

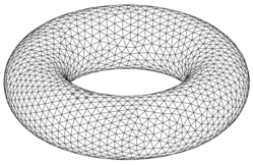

Figure 3.2: A 3d triangulation of a torus.

If we unfold this triangulation into two dimensions, we can visualize it on paper as a square with the corresponding triangularization depicted in Figure 3.3. By folding the paper and joining two opposite sides, we create a cylinder without bases. If we were to further deform and connect the remaining edges of this cylinder, we would ultimately form a torus. This process highlights the relationship between different topological surfaces and how they can be represented through triangulation, illustrating the continuity of shapes as we transition from flat representations to three-dimensional objects.

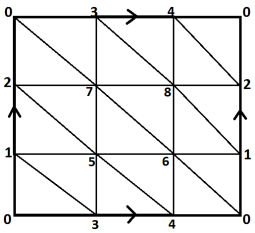


Figure 3.3: A 2d triangulation of a torus

Note that given $A \cong B$, then $A \backsimeq B$, so we can consider continuous deformations of surfaces. Homotopy enables the exploration of how spaces can be continuously deformed while preserving their essential topological properties. For instance, while a sphere and a closed disk are homeomorphic, their triangulations may differ. However, by employing a homotopy that gradually flattens the sphere into a disk, we can continuously triangulate the deformed space while maintaining its topological properties throughout the process. This illustrates how homotopy complements the Triangulation Theorem, facilitating the understanding of surface topology through deformations. Figures 3.4 and 3.5 provide several visual examples of these concepts.

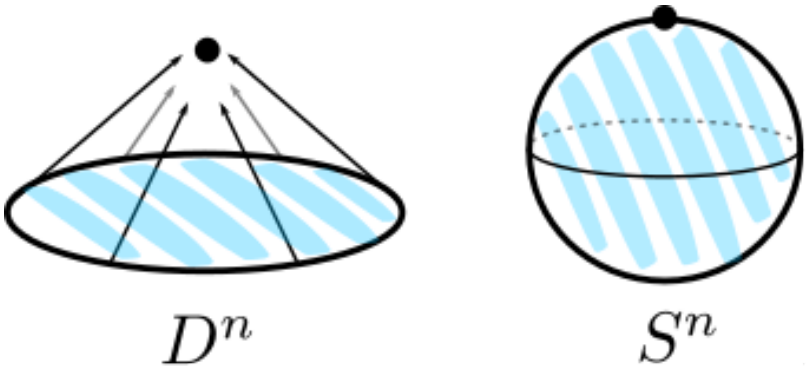


Figure 3.4: Both surfaces can be contracted to one point.

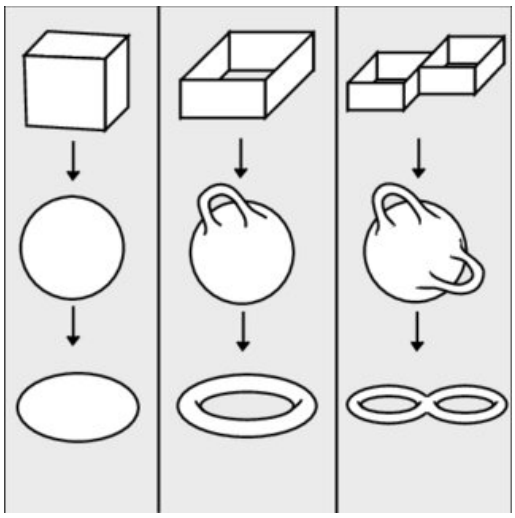

Figure 3.5: Homeomorphism between surfaces.

## 3.2 Simplicial homology

In this section, we delve into the powerful technique of simplicial homology, which provides a systematic approach to quantifying and understanding the 'holes' or non-simply connected regions in a topological space. By associating algebraic structures known as homology groups to simplicial complexes, we gain insight into the topological properties of spaces and their connectivity. We will explore how to calculate the homology groups of simplicial complexes over a given field, laying the groundwork for a deeper understanding of topological spaces and their structures.

**Definition 3.2.1** Let $\Delta = [v_0, ..., v_k]$ be a simplex, an **orientation** of a simplex is an ordering of the vertices $v_0, ..., v_k$, with the convention that two orderings define the same orientation if they differ by an even permutation.

Recall that a transposition or inversion is a permutation of exactly two elements and that a permutation is even when it is the product of an even number of transpositions (inversions).
**Definition 3.2.2** Let F be a field, let K be a simplicial complex and d $\geq 0$ a dimension. We define a **d-chain** as the sum of oriented d-simplices with coefficients in F:

$$c = \sum_{\sigma} a_\sigma \sigma$$

where $\sigma$ is a oriented d-simplex and $a_\sigma \in$ F.

**Definition 3.2.3** The **set of d-chains** forms a vectorial space that we will denote:

$$\mathrm{C}_d(\mathrm{K}, \mathrm{F})$$

Note that its a F-vectorial space with dimension equal to the number of d-simplices of K.

**Definition 3.2.4** The **edge of a d-simplex** is defined as:

$$\partial_d([v_0, v_1, ..., v_d]) = \sum_{i=0}^{d} (-1)^i [v_0, v_1, ..., \widehat{v_i}, ..., v_d]$$

where $\widehat{v_i}$ states that we omitted $v_i$. Note that the edge of one d-simplex is one (d-1)-chain because we omit $v_i$.

**Definition 3.2.5** Let c be a d-chain / $c = \sum a_\sigma \sigma$. We define the **linear edge of a chain** as:

$$\partial c = \partial_d c = \sum_{\sigma} a_\sigma \partial \sigma$$

Given that its a linear definition:

$$\partial_d : C_d(K, F) \to C_{d-1}(K, F)$$

we can say that $\partial_d$ is a morphism of vector spaces:

$$\partial_d(c + c') = \partial_d c + \partial_d c' \; and \; \partial_d(\lambda c) = \lambda \partial_d c \; \forall c, c' \in C_d \; and \; \lambda \in F$$

**Definition 3.2.6** Let c $\in$ $C_d$(K, F) be one d-chain, it will be a **cycle** if $\partial_d$(c)=0

**Definition 3.2.7** Let c $\in$ $C_d$(K, F) be one d-chain, c is an **edge** if $\exists$ a chain c' $\in$ $C_{d+1}(K, F)$ / $\partial_{d+1} c' = c$

**Definition 3.2.8** We define the **set of d-cycles** as:

$$Z_d = Z_d(K, F) = \{c \in C_d(K, F) \; / \; \partial c = 0\}$$

**Definition 3.2.9** We define the **set of d-edges** as:

$$B_d = B_d(K, F) = \{c \in C_d(K, F) \; / \; c = \partial c' \; for \; some \; c' \in C_{d+1}(K, F)\}$$

Note that $Z_d$ and $B_d$ are vector subspaces of $C_d$ because:

$$Z_d = ker(\partial_d) \quad and \quad B_d = Im(\partial_{d+1})$$

**Definition 3.2.10** Let K be a simplicial complex and F a field. A **d-Homology group** is defined as:

$$H_d(K, \mathrm{F}) = Z_d(K, \mathrm{F}) / B_d(K, \mathrm{F})$$

We can also define the homology group dimension as:

$$dim(H_i) = dim(Z_i) - dim(B_i)$$

In the study of algebraic topology, homology groups provide valuable insights into the shape and structure of topological spaces. For instance, consider a disk; it has no holes, which means its zeroth homology group, denoted $H_0$, is isomorphic to $\mathbb{Z}$, representing its connected components. Next, a circle possesses a one-dimensional hole, characterized by its first homology group $H_1$, which is also isomorphic to $\mathbb{Z}$. Similarly, an annulus features a one-dimensional hole as well, and interestingly, both a circle and an annulus are homotopy equivalent, reflecting the same topological structure in terms of their homology groups. Lastly, a sphere exhibits a two-dimensional hole, corresponding to its second homology group $H_2$, which is isomorphic to $\mathbb{Z}$. Therefore, we can summarize the homology groups for these examples as follows: $H_0 \cong \mathbb{Z}$ for the disk, $H_1 \cong \mathbb{Z}$ for both the circle and the annulus, and $H_2 \cong \mathbb{Z}$ for the sphere, illustrating the relationship between the dimensions of holes and their corresponding homology groups. All of these concepts can be visualized in Figure 3.6.

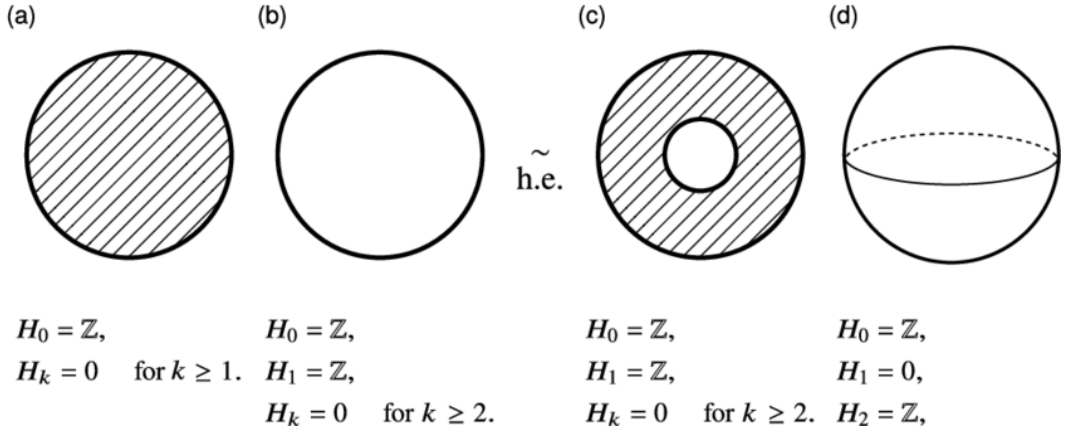


Figure 3.6: Homology Groups Example.

**Definition 3.2.11** Let K be a simplicial complex and d its dimension, its **Betti Number** is defined as

$$\beta_d(K) = \beta_d(K, \mathrm{F}) = dim(H_d(K, \mathrm{F}))$$

**Definition 3.2.12** We define the euler characteristic of a simplicial complex K as:

$$\chi(K) \ = \ \sum_d (-1)^d (number\ of\ d-simplexs\ of\ K)$$

**Theorem 3.2.13** Let K be a simplicial complex:

$$\chi(K) \ = \ \sum_d (-1)^d \beta_d(K) \ = \ \sum_d (-1)^d dim(H_d(K, F))$$

In conclusion, we have explored the fundamental concepts of homology groups and their significance in understanding the topology of simplicial complexes. By defining homology groups and their dimension, we have established a framework for quantifying the 'holes' or non-trivial cycles in a simplicial complex. Furthermore, we have discussed the relationship between homology groups and Betti numbers, as well as their connection to the Euler characteristic for a simplicial complex. Through these insights, we gain a deeper understanding of the topological structure encoded within simplicial complexes, paving the way for further exploration into the rich interplay between algebraic and geometric concepts in topology.

Moreover, armed with the ability to quantify the 'holes' of a simplicial complex through homology groups, we can now undertake comparative analyses between different complexes. By computing the homology groups of a given complex, we can determine its topological characteristics and discern whether it shares the same homology groups as another complex. This enables us to identify topological similarities or differences between complexes and investigate their structural properties. Furthermore, we can extend this analysis to compare the homology groups of a complex to those of a known compact surface.

It's important to note that the equality of homology groups between two simplicial complexes doesn't automatically guarantee their homotopy equivalence. While homology groups provide valuable information about the topology of complexes, homotopy is a broader notion involving the possibility of continuously deforming one complex into another. Therefore, while the equality of homology groups suggests similarities in the topological structure of the complexes, other topological properties must be considered for a complete characterization of their relationship.

For instance, consider the case of a solid sphere and a hollow sphere: both have the same homology groups, yet they are not homotopically equivalent due to their differing topological structures.

## 3.3 Filtrations and persistence

Now that we understand how to compare topological structures among simplicial complexes, compact surfaces, and their continuous deformations, the question arises: how do we compute a simplicial complex that represents a given point cloud of data? It's crucial to find an optimal method for generating the simplicial complex, as it will be analyzed to gain a deeper understanding of the topological structure of the point cloud. This analysis allows us to determine if the structure resembles any known compact surface or exhibits unique topological features.

In this section we aim to explain how to compute a simplicial complex given a point cloud data and the problem of its computational complexity. Moreover we will define persistence given a filtration and a useful way to illustrate it.

**Definition 3.3.1** Let X be a set of finite points of the euclidean space such that:

$$X = \{x_0, x_1, ..., x_N\} \subset \mathrm{R}^n$$

We fix a radius $r > 0$ and d as the euclidean distance to consider the open balls of radius r centred at the points $x_i \subset X$:

$$B(x_i, r) \ = \ \{a \in \mathrm{R}^n \ / \ d(a, x_i) < r\}$$

We define the **Čech complex** Č(X,r) as the simplicial complex with the following simplexs:

1. Vertexs of Č(X,r) are in bijection with the x points: we call $v_i$ to the corresponding vertex to $x_i$ $\forall$i = 0,...,N.

2. A simplex $[v_0, ..., v_k]$ belongs to Č(X,r) if:

$$B(x_0, r) \cap B(x_1, r) \cap ... \cap B(x_k, r) \ \neq \ \emptyset$$

(at least one point of $\mathbf{R}^n$ is at a distance < r to the points $x_0, x_1, ..., x_k$)

Čech vertexs (0-simplexs) are the points $x_0, ..., x_N$ of X, the edges (1-simplexs) correspond to the pair of balls that cut into each other and 2-simplexs correspond to three balls that intersect, and so on.

We define the Čech environment as the union of the set of balls of radius r of the points of X:

$$N_r(X) \ = \ \bigcup_{x \in X} B(x, r)$$

**Theorem 3.3.2** Čech complex Č(X,r) has the homotopy type of $N_r(X)$:

$$|\check{C}(X,r)| \backsimeq N_r(X)$$

Note that one of the consequences is that the homology of the Čech complex is that it is the homology of the tubular environment $N_r(X)$.

Čech complex can be calculated algebraically from the points of X by calculating distances, but it is computationally very complex. Let n be the number of points of X, then we have $2^n$ intersections, which is a very high value.

Čech complexes are constructed by varying the radius between points in a dataset. As the radius changes, different simplices emerge, reflecting the underlying connectivity of the points. For small radii, isolated points or small clusters may form, while as the radius increases, more complex structures appear, leading to the formation of higher-dimensional simplices. This dynamic illustrates how the choice of radius can significantly impact the shape and structure of the resulting Čech complex. An illustrative example of these concepts can be seen in Figure 3.7.

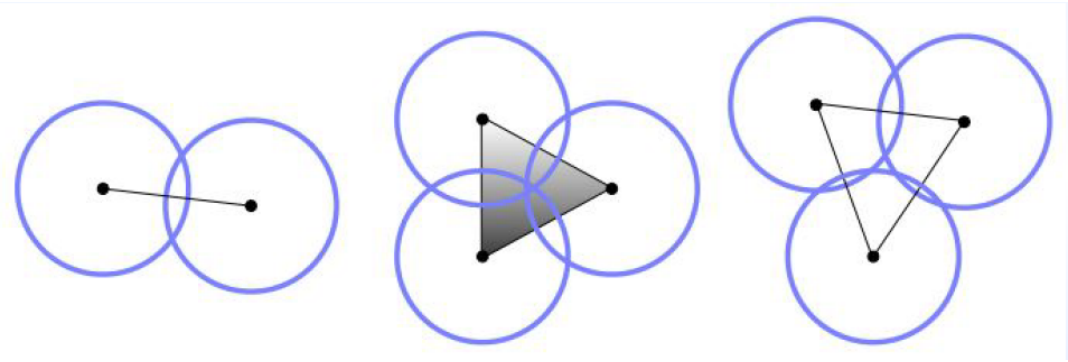

Figure 3.7: Čech complex construction.

**Definition 3.3.3** A **Vietoris-Rips** complex VR(X,r) is a simplicial complex such that:

1. Each point of X its one VR(X,r) vertex.

2. $[v_0, v_1]$ edge belongs to VR(X,r) if $B(x_0, r) \cap B(x_1, r) \neq \emptyset$. Or in other words there is some $\mathbb{R}^n$ point to any distance lower than r to $x_0, x_1$ points.

3. If $k \geq 2$, $[v_0, ..., v_k]$ simplex belongs to VR(X,r) if all $[v_i, v_j]$ edges belong to VR(X,r) for $0 \leq i < j \leq k$.

VR(X, r) calculation complexity is lower than Čech one. Let n be the points of X, we have to calculate $\binom{n}{2} = \frac{n(n-1)}{2}$ distances among points, but it is still a high computational complexity value. Recent research has investigated methods to improve this problem, including GPU-accelerated computation techniques for Vietoris-Rips persistence barcodes, as discussed by Zhang et al. [8].

Similar to Čech complexes, Vietoris-Rips complexes are formed by varying the radii between points in a dataset. However, the key distinction lies in how simplices are constructed: in Vietoris-Rips complexes, a simplex is created whenever all points within a certain radius are pairwise connected, regardless of their actual distances. As the radius increases, this leads to the emergence of more extensive and interconnected structures. This characteristic allows Vietoris-Rips complexes to capture different topological features compared to Čech complexes, where the formation of simplices is contingent on the coverage of balls centered at the points. The differences in these constructions and their implications for topology can be visualized in Figure 3.8.

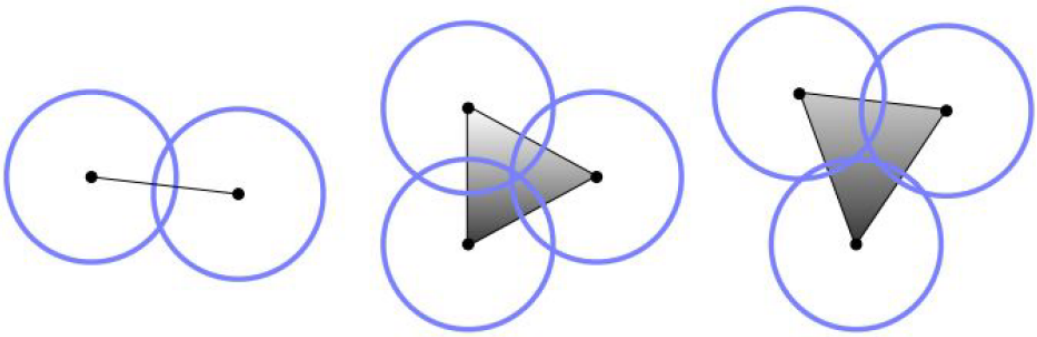

Figure 3.8: Vietoris-Rips complex construction.

**Definition 3.3.4** A simplicial complex **filtration** is a $K_r$ simlicial complex family indexed by $r > 0$ with $r \in \mathbb{R}$ (continuous) or $r \in \mathbb{N}$ (discret) such that $K_r \subset K_{r'}$ for $r' > r$. Let $\sigma$ be a $K_r$ simplex, then $\sigma$ is also a $K_{r'}$ simplex for all $r' > r$.

For all pair of radius such that $r < r'$ we consider the following inclusion:

$$\phi = \phi_{r,r'} : K_r \hookrightarrow K_{r'}$$

which not only is injective but also simplicial. It induces a morphism of chain complexes

$$\phi_{\#} : C_i(K_r) \to C_i(K_{r'})$$

for each i dimension, in other words $\partial_i \phi_{\#} = \phi_{\#} \partial_i$. For each particular $\phi_{\#}$ function carries cycles to cycles and chains to chains and induces an application in homology

$$\phi_* : H_i(K_r) \to H_i(K_{r'}).$$

**Definition 3.3.5** The $\phi_*(H_i(K_r))$ in $H_i(K_{r'})$ is a **persistent homology group** of i dimension and $r < r'$ parameters. We also define a **persistent homology class** as one element of this image, in other words an element of $H_i(K_r)$ seen to $H_i(K_{r'})$.

The morphism of chains $\phi_\# : C_i(K_r) \to C_i(K_{r'})$ is injective.

If $z \in Z_i(K_r)$ is a non-null cycle, then $\phi_\#(z) \in Z_i(K_{r'})$ is a non-null cycle for $r' > r$. If $b \in B_i(K_r)$ is an edge, then $\phi_\#(b) \in B_i(K_{r'})$ is also an edge for $r' > r$.

Not that above properties does not imply that the induced morphism in homology $\phi_* : H_i(K_r) \to H_i(K_{r'})$ is injective but it implies the existence of a birth and a death if we consider r parameter as a time.

**Definition 3.3.6** For each homology class in a t time, we obtain a time of **death**, **birth** and a life time or **persistence**. Formally, for each $\theta \in H_i(K_r)$ we have that:

$$birth(\theta) = inf\{\alpha' \leq \alpha \ / \ \theta \in (\phi_* : H_i(K_{r'}) \to H_i(K_r))\ image\}$$
$$death(\theta) = sup\{\alpha' \geq \alpha \ / \ \theta \notin (\phi_* : H_i(K_r) \to H_i(K_{r'}))/kernel\}$$
$$persistence(\theta) = death(\theta) - birth(\theta)$$

**Theorem 3.3.7** For all $r > 0$, a choice of $H_i(K_r)$ bases is possible, such that all its elements that are still 'alive' for $r' > r$ are part of one $H_i(K_{r'})$ base.

**Definition 3.3.8** A **persistence diagram** is a diagram such that classes are represented in one persistent homology base. Each element of the base is represented as a point in the plane, first coordinate is the birth parameter and second one death parameter.

We can always obtain a diagram for each d degree of a $H_d$ homology or a joint network. In a persistence diagram all points are above the diagonal (points can't die before their birth) and the length of a class is given by the distance from the point to the diagonal.

**Definition 3.3.9** A **barcode** is a diagram such that each element of the persistent homology base is represented as a bar that starts in its birth value and ends in its death value so for each r parameter we have one homology base.

Length of the bars gives the lifetime of the persistent homology classes. Both last two diagrams are equivalent.

Note that in zero dimension we obtain a lot of classes that quickly disappear. Relevant classes are those with the longest duration (long bars in the code, away from the diagonal on the diagram) but there is not a clear or absolute way to decide the relevance of a class. It is possible

to vectorise the diagrams, i.e. transform them into vectors of a finite or infinite vector space for use in statistics or machine learning.

The main problem is obtaining the filtrations, since we have the disadvantages of sensitivity to disturbances and noise and computational complexity.

**Definition 3.3.10** Let $D_1$ and $D_2$ be two persistence diagrams, we consider all $\phi : D_1' \to D_2'$ bijections between the results of joining points on the diagonal to both $D_1$ and $D_2$. We define the **bottleneck distance** as:

$$d_b(D_1, D_2) = \ inf_\phi \ sup_{x \in D_1'} \ \| \ x - \phi(x) \ \|_\infty$$

where $\| \ x - x' \ \|_\infty = \ max\{| \ x_1 - x_1' \ |, | \ x_2 - x_2' \ |\}$ and the infimum is obtained from all possible bijections after choosing the points of the respective diagonals.

**Definition 3.3.11** We define the **p-1 wasserstein distance** as:

$$W(D_1, D_2) = \inf_\phi \sum_{x \in D_1'} \| \ x - \phi(x) \ \|$$

Wasserstein distance and Bottleneck distance differ in how they measure discrepancies between persistence diagrams. Wasserstein distance takes a more global approach by considering the overall cost of transporting points from one diagram to another, summing up all the individual distances between matched points. In contrast, the Bottleneck distance can be understood as a "worst-case" measure, as it focuses on the largest distance between any pair of points across the two datasets. While Wasserstein captures the cumulative difference, Bottleneck highlights the most significant discrepancy between the diagrams.

With all necessary definitions in place, we are now prepared to execute the workflow illustrated in Figure 3.9. This process starts with a given dataset, from which we compute its simplicial complexes and derive the corresponding persistence diagram. This diagram is essential for comparing it with other persistence diagrams using the previously defined distances, allowing us to quantitatively assess differences and better understand the dataset's topological features. In Figure 3.10, we can see a example of one persistence diagram that illustrate these concepts.

The idea is to look at the distance between persistence classes of one diagram with classes of the other diagram or with the diagonal (for cycles that have a short duration). This distance allows us to compare persistence diagrams. In Figure 3.11, we see a graphical example of Bottleneck distance, and in Figure 3.12, we illustrate Wasserstein distance.

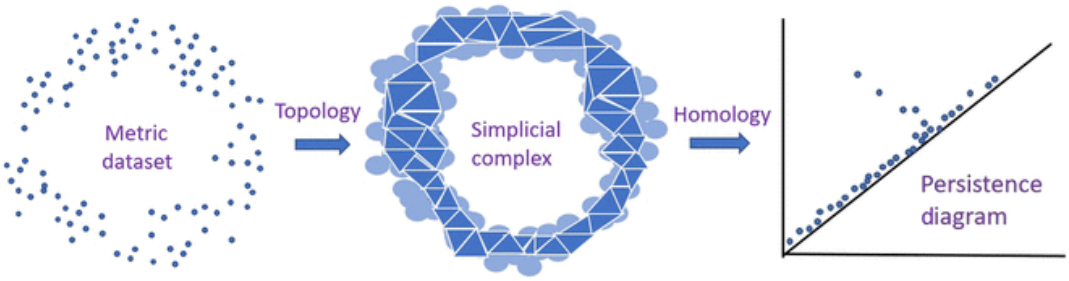


Figure 3.9: Computing Persistent Homology.

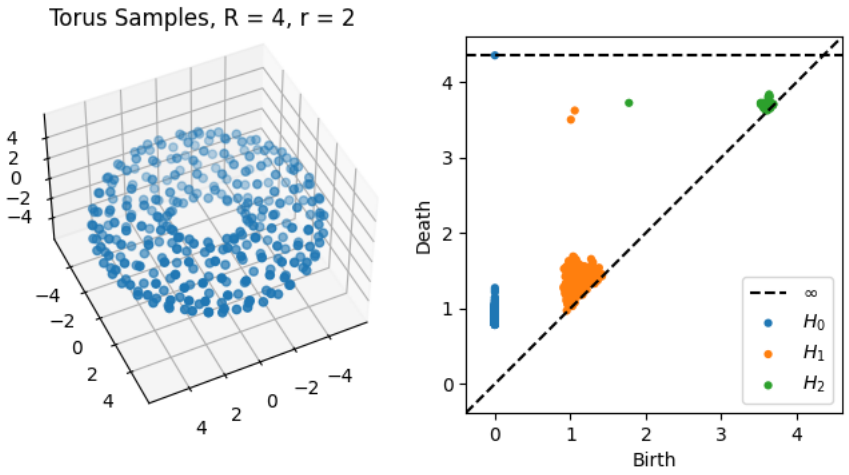


Figure 3.10: Persistence Diagram Example.

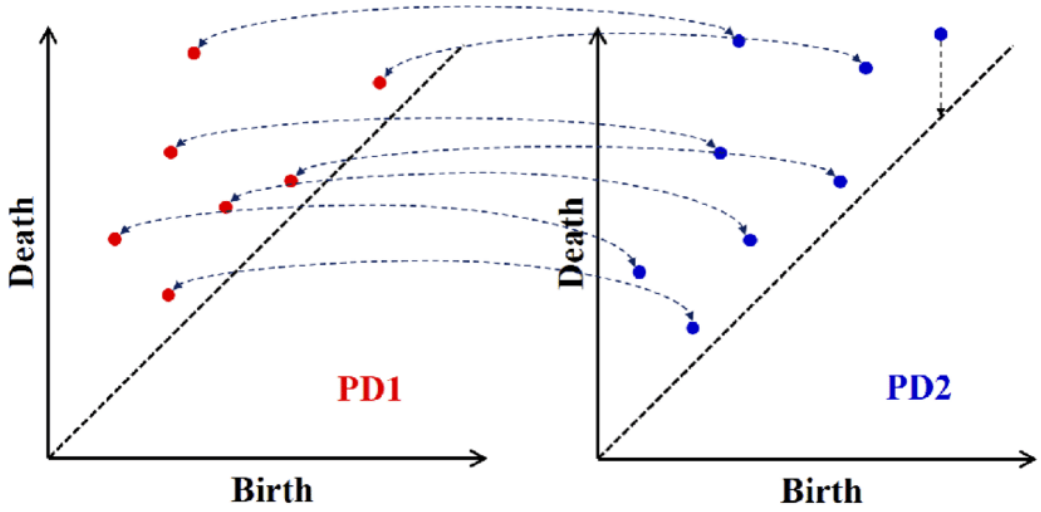


Figure 3.11: Bottleneck Distance Example.

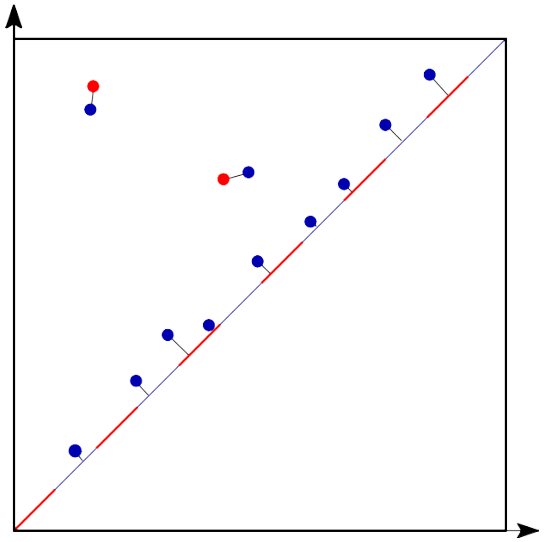

Figure 3.12: Wasserstein Distance Example.

# Chapter 4

# Case of study: Emergency calls in Valencia

In this chapter we will present a practical approach to the above and relate topological analysis to space-time one. The case study presented here is based on the same dataset used by Payares-Garcia et al. [9], whose work provides a dynamic spatio-temporal stochastic modeling approach to emergency calls in an urban context. This study offers valuable insights and serves as the foundation for our analysis.

## 4.1 Presentation and context of dataset

Criminology is a crucial discipline for understanding and addressing security issues within society. Within this field, the analysis of emergency phone calls emerges as a key tool for the prevention and response to crisis situations, such as crimes, accidents, or medical emergencies. The importance of studying and analyzing these calls lies in their ability to provide real-time critical data, including both spatial and temporal information, which enables the rapid location and response to incidents.

Emergency calls not only provide details about the type of incident but also about the location and timing, data that are essential for the efficient allocation of resources and the implementation of mitigation strategies. Additionally, analyzing this data helps identify crime patterns and high-risk areas, which is vital for planning public safety policies.

In this context, the relationship between criminology and emergency calls is evident. Criminology greatly benefits from the analysis of the spatio-temporal data generated by these calls,

allowing for the modeling and prediction of criminal behaviors in urban environments. This is particularly relevant in the digital age, where the increasing availability of information and sensor technology has generated massive volumes of high-resolution data that can be utilized to enhance public safety.

Therefore, studying and analyzing emergency phone calls is not only useful for the prevention and response to emergencies but also contributes to the development of predictive models that can identify trends and prevent future incidents. This, in turn, has a direct impact on people's lives, improving their safety and well-being by enabling a faster and more efficient response to risky situations.

In the context of the city of Valencia, the spatio-temporal analysis of emergency calls has revealed how certain areas of the city experience an increase in criminal events, while others show a decrease. This type of analysis is crucial for adapting public safety strategies to the specific dynamics of each area, allowing for more effective resource allocation and better protection for citizens.

The dataset used in this study comprises approximately 90,000 entries, each representing a crime-related emergency call in Valencia, Spain, collected over several years. This extensive dataset provides a detailed view of various aspects of these incidents, making it a valuable resource for spatio-temporal analysis and criminology research. The dataset's large size and comprehensive scope allow for in-depth analysis, enabling researchers to explore patterns and trends in criminal activity across both time and space.

The dataset includes 28 columns that capture critical information about each incident. These columns provide a wealth of data that can be utilized to understand the context of each crime and its relation to the surrounding environment.

The key columns in the dataset include:

**crime_id**: A unique identifier for each crime entry.

**crime_date**: The date on which the crime occurred, formatted as "MM/DD/YYYY."

**crime_time**: The exact time of the crime in 24-hour format.

**crime_type**: The type of crime reported, such as: 'Agresion', which refers to incidents of assault involving physical violence or threats, 'Sustraccion', which pertains to theft or robbery, where property was unlawfully taken, 'AlarmasMujer', which involves emergency calls related to situations where women were in immediate danger or distress, likely due to domestic violence or similar threats and 'Otros', which encompasses other types of crimes that do not fall into the aforementioned categories, covering a range of various criminal activities. These classifications

enable a nuanced analysis of the different criminal incidents in Valencia, aiding in a deeper understanding of the city's crime dynamics.

**muni**: The municipality where the crime occurred, typically indicating "Valencia".

**year**, **month**, **week**, **day**: These columns break down the date into components, allowing for detailed temporal analysis by year, month, week of the year, and day of the month.

**week_day**: A numerical representation of the day of the week, where 1 corresponds to Monday and 7 to Sunday.

**week_day_name**: The name of the day, which helps in identifying patterns related to specific days.

**crime_hour**: The hour at which the crime occurred, useful for analyzing daily crime patterns.

**crime_lon** and **crime_lat**: The longitude and latitude coordinates of the crime location, essential for spatial analysis.

Distance to landmarks (**atm_dist**, **bank_dist**, **bar_dist**, etc.): These columns measure the distance from the crime location to various types of landmarks, such as ATMs, banks, bars, cafes, industrial areas, markets, nightclubs, police stations, pubs, restaurants, and taxi stands. These distances are expressed in meters.

**grid_id**: An identifier for the spatial grid where the crime occurred, aiding in spatial aggregation.

**grid_lon** and **grid_lat**: The longitude and latitude of the grid's centroid, useful for mapping and spatial analysis.

The dataset covers the period from 1 January 2010 to 31 October 2020. During these ten years, 83,379 calls were registered by the emergency number in Valencia. These calls are categorized into four main groups based on the nature and reason for the call: 51,533 calls related to assaults, 23,282 calls related to robberies of individuals in the streets, 388 calls related to aggression against women and 8,176 calls related to other unspecified causes.

Data from 1 March 2020 onwards are excluded from validation due to significant disruptions caused by the COVID-19 pandemic, which altered the patterns and structure of emergency calls and criminal behavior.

The detailed breakdown of date and time into various components allows for robust temporal analysis, enabling researchers to identify peak times for different types of crimes and understand how criminal activity fluctuates throughout the day, week, month, and year. The inclusion of precise geographic coordinates and distances to key landmarks facilitates detailed spatial analysis, helping to identify crime hotspots and understand the influence of environmental factors on criminal behavior. By combining these temporal and spatial dimensions, researchers can develop predictive models, assess the effectiveness of law enforcement strategies, and gain deeper insights into the dynamics of crime in urban environments like Valencia. This dataset, therefore, serves as a critical tool for criminologists and public safety officials aiming to enhance public safety and prevent crime.

To handle the large dataset from Valencia effectively, we have opted to use the Google Colab programming environment. We divided the general dataset into multiple smaller subsets using various filters, allowing us to work with reduced data while uncovering relationships of interest. All the code is written in Python and RStudio, and you can find the complete GitHub Repository, where you can find all the dataset partitions, code, files, and resources used in the project.

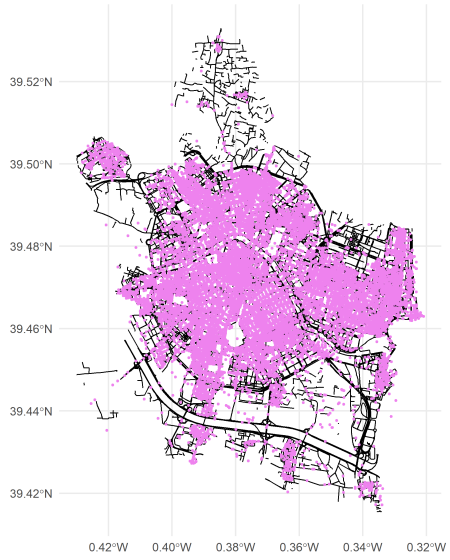


Figure 4.1: Agresion crime type.

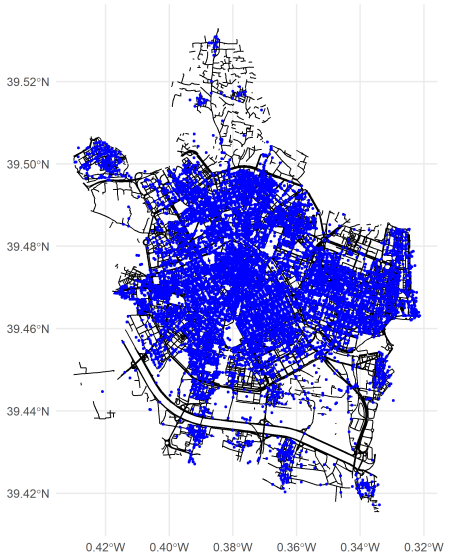


Figure 4.2: Sustraccion crime type.

As an introduction to the dataset, we present a partition into four categories, each representing a different type of crime. Figure 4.1 shows 55,610 entries categorized as *Aggression*, accounting for **61.62%** of the total. Figure 4.2 includes 25,343 entries labeled as *Theft*, representing **28.08%** of the total. Figure 4.3 contains 455 entries corresponding to crimes labeled as *AlarmasMujer*, which makes up **0.50%** of the dataset. Finally, Figure 4.4 features 8,842 entries categorized as *Others*, accounting for **9.80%** of the total.

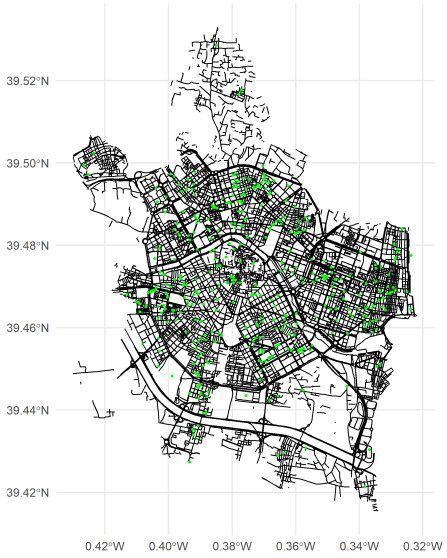


Figure 4.3: AlarmasMujer crime type.

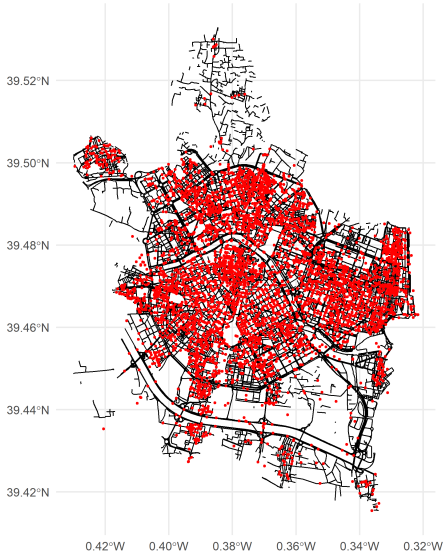


Figure 4.4: Others crime type.

## 4.2 Spatio-temporal analysis

To begin our analysis, we utilized the R libraries sf, ggplot2, and dplyr to visualize datasets effectively. By leveraging the shapefile of the city of Valencia, we can plot various datasets, as illustrated in Section 4.1. In this case, we focus on a filtered partition of the Valencia dataset for the year 2018, specifically highlighting the crime type "AlarmasMujer," which comprises 95 points. The use of these libraries allows us to explore the spatial distribution of the data, providing valuable insights into the geographic patterns of crime occurrences in the region.

As part of our analysis, we generated random datasets to compare with the real data, focusing on studying randomness in crime occurrences. Throughout this project, we considered the range of values within the dataset, specifically the minimum and maximum values of `crime_lat` and `crime_lon`, ensuring that the number of points generated matched the number we aimed to compare—95 points in this case. We generated the random datasets in Google Colab using Python's `numpy` library with the function `np.random.uniform()`, simulating a Poisson random process. Additionally, we utilized the `geopandas` library and `shapely.geometry` to handle `.shp` files and generate random points within these ranges. Figure 4.5 illustrates an example of this Poisson random process for comparison with the data presented in Figure 4.6. It is trivial that some points appear outside the map boundaries, as the maximum and minimum values create a square of possible values. This visual representation suggests that the occurrences of this type of crime are not random.

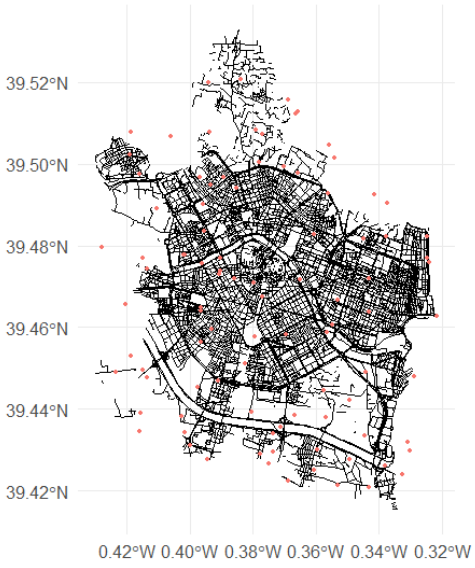


Figure 4.5: Poisson Process.


Figure 4.6: AlarmasMujer 2018.

In our random data generation process, we opted not to use a simple rectangular boundary defined by the maximum and minimum longitude and latitude values. Instead, we employed a convex hull approach to more accurately represent the area where events can realistically occur. This method excludes regions outside the city, such as edges or corners that would otherwise fall within a rectangular bounding box but are irrelevant to our analysis. By focusing on the actual shape of the city, the convex hull maximizes the relevance of the spatial area and minimizes the impact of including non-city zones, ensuring that our calculations are not skewed by the inclusion of these irrelevant areas. This can be observed in Figure 4.7, which shows the convex hull constructed around the entire dataset. Additionally, Figure 4.8 offers another perspective, displaying only the randomly generated points (the same points as shown in Figure 4.5).

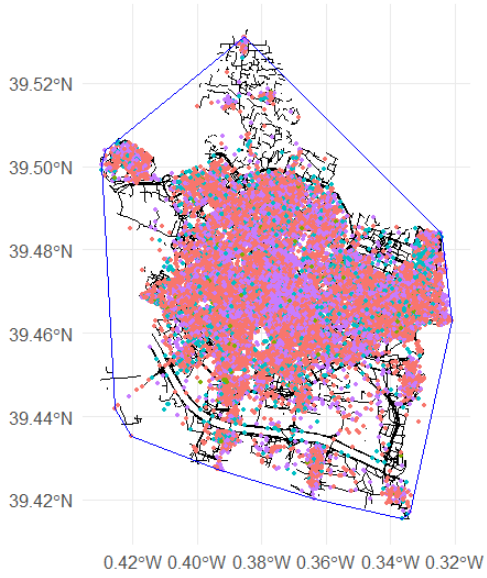


Figure 4.7: Convex Hull of the dataset.

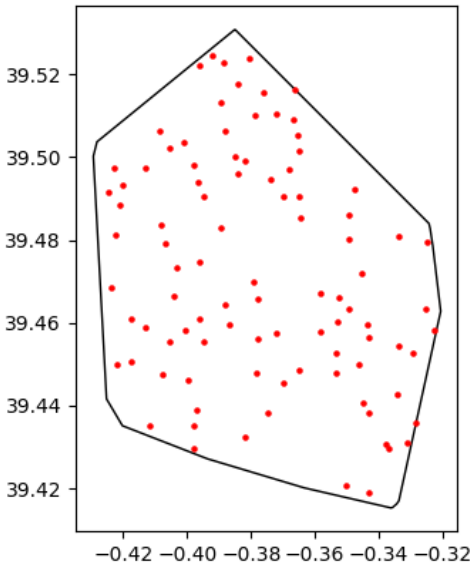


Figure 4.8: Poisson Process

Our space-time analysis is centered on generating Spatial K-function plots, Spatial Pair Correlation function plots, and Space-time Pair Correlation function plots. To perform this analysis, we utilized RStudio along with the `stpp` and `spatstat` libraries. These tools allowed us to assess spatial clustering and correlation, as well as the interaction between spatial and temporal patterns in the dataset.

To utilize spatial data effectively, we transformed the 'crime_lon' and 'crime_lat' coordinates into UTM (Universal Transverse Mercator) values. This conversion allows for more straightforward interpretation of the results, as the distances can be expressed in kilometers.

We will begin our analysis with the 2018 dataset of AlarmasMujer calls. To perform the spatial analysis, we will use the `utm_x` and `utm_y` variables, which represent the spatial coordinates of each event. With these coordinates, we will calculate the pair correlation function and the K-function, both of which will allow us to understand the spatial distribution and clustering patterns of the data.

The K function is a spatial statistical tool used to analyze the degree of clustering or dispersion in a point pattern, in this case, the spatial distribution of gender violence crimes in Valencia during 2018. The function measures the expected number of points within a distance $r$ from an arbitrary point, allowing for the detection of clustering or regularity at different spatial scales (see Fig. 4.9).

In our analysis, we computed several versions of the K function: isotropic ($\hat{K}_{iso}(r)$), translation-corrected ($\hat{K}_{trans}(r)$), border-corrected ($\hat{K}_{bord}(r)$), and Poisson ($K_{pois}(r)$). The isotropic K function assumes that the point pattern is uniform in all directions, which can be useful when the dataset has no clear anisotropy. However, real-world data, particularly in urban environments, often suffer from edge effects. These effects occur because points near the boundary of the study region have fewer neighbors than those near the center, which can lead to underestimation of the K function at large distances.

To address this, both the translation ($\hat{K}_{trans}(r)$) and border corrections ($\hat{K}_{bord}(r)$) were applied. These methods adjust for the edge effects, providing a more accurate estimate of point interactions near the borders. Despite these corrections, the overall trend remained consistent: all observed K functions showed significantly larger values than the Poisson K function. This suggests that the crimes exhibit strong spatial clustering rather than being randomly distributed throughout the city.

The steep increase in the K function values, particularly at larger distances, indicates that spatial clustering persists across a wide range of scales. The differences between the isotropic K function and the edge-corrected versions show that adjusting for edge effects has a significant impact on the results, but the pattern of clustering remains dominant regardless of these corrections.

While the K function provides a global measure of clustering, the Pair Correlation Function (PCF) allows us to explore the local interactions between points at specific distances. The PCF, denoted as $g(r)$, measures the density of points at distance $r$ from an arbitrary point, relative to a random distribution. In our case, the PCF enables us to assess how gender violence crimes are clustered or dispersed at fine spatial scales (see Fig. 4.10).

We computed several versions of the PCF: Ripley's PCF ($\hat{g}_{ripley}(r)$), translation-corrected PCF ($\hat{g}_{trans}(r)$), and the Poisson PCF ($g_{pois}(r)$). The observed PCF values ($\hat{g}_{ripley}(r)$ and $\hat{g}_{trans}(r)$) were significantly higher than the Poisson PCF at small distances (r ¡ 200 meters), indicating

strong clustering of crimes at short distances. This suggests that crimes are not evenly dispersed but tend to occur in close proximity to each other.

As the distance increases beyond 200 meters, the observed PCF values gradually approach the Poisson level, indicating that at larger distances, the distribution of crimes becomes more random. However, the sharp peak at short distances highlights the presence of local crime clusters, possibly indicating specific areas where incidents are more likely to occur close together.

The difference between the translation-corrected PCF and Ripley's PCF is minimal, showing that while edge effects are present, they do not drastically alter the overall conclusions. The strong clustering observed at smaller distances is robust across different methods of PCF calculation.

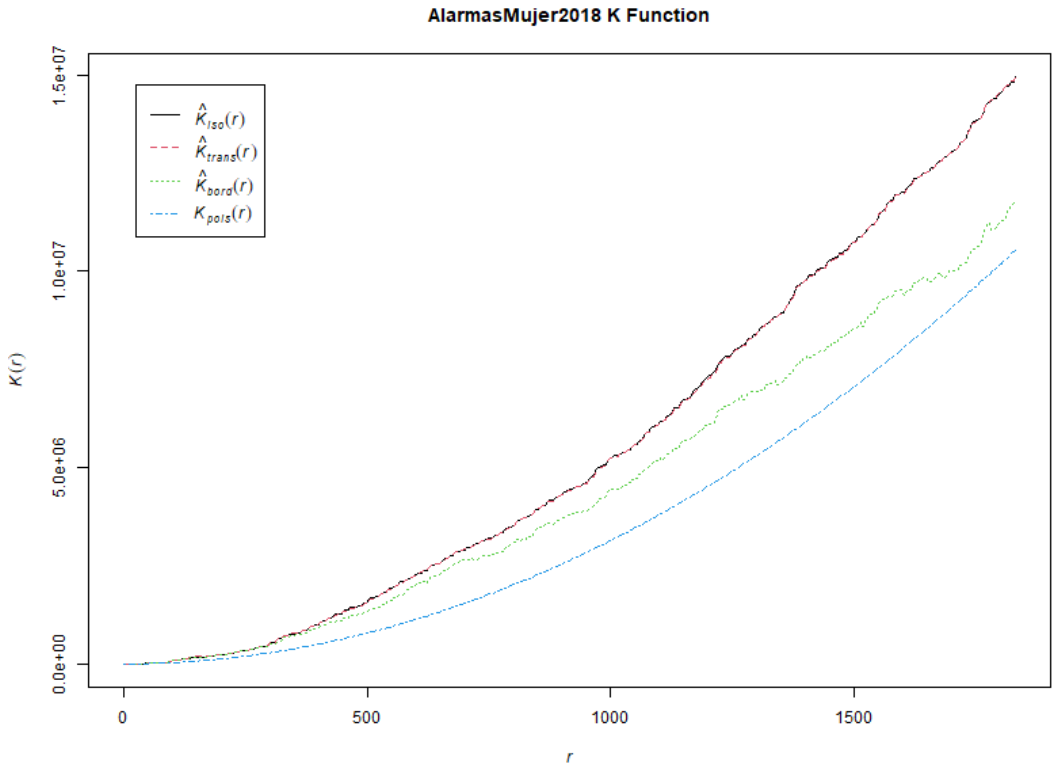


Figure 4.9: Spatial Ripley's K Function.

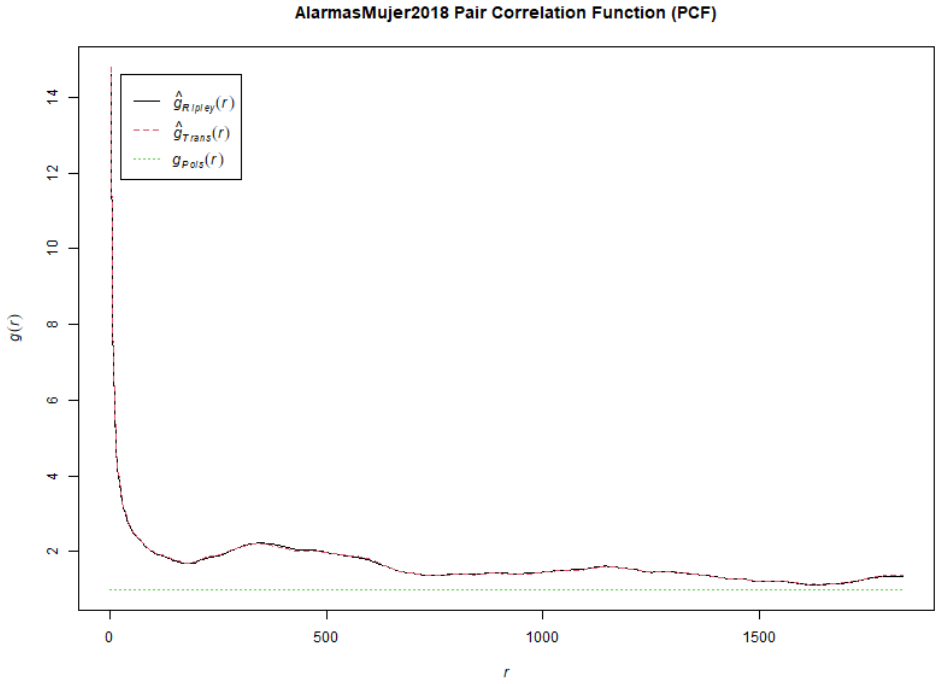


Figure 4.10: Spatial Pair Correlation Function.

In addition to the spatial analysis, we extended our study to include a spatiotemporal analysis using the spatiotemporal pair correlation function. This function $g(u, v)$ takes into account both the spatial coordinates (u, v) and time, allowing us to assess whether crimes are correlated not only in space but also in time (see Figs. 4.11 and 4.12).

The spatiotemporal PCF is represented by a 3D surface plot, where the spatial axes (u, v) represent distance and the vertical axis represents the intensity or strength of correlation over time. The sharp peaks in the surface indicate areas of strong spatiotemporal clustering. These peaks suggest that not only are the crimes clustered in space, but there is also a temporal component to this clustering. Specifically, crimes tend to occur close to each other in both space and time, forming clusters that are temporally and spatially correlated.

The steep rise in the spatiotemporal PCF surface, particularly in specific regions, suggests the presence of temporal hotspots. These are periods when crime activity is highly concentrated in certain spatial areas, indicating that crimes are not only geographically clustered but also occur in bursts over specific periods. This pattern could be linked to underlying factors such as local events, socio-economic conditions, or policing activity.

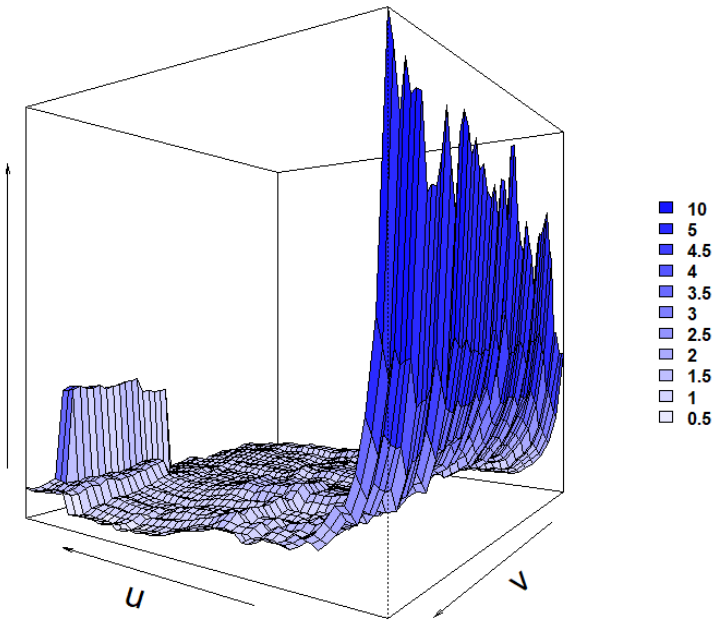


Figure 4.11: Spatio-Temporal Pair Correlation Function (Viewpoint 1).

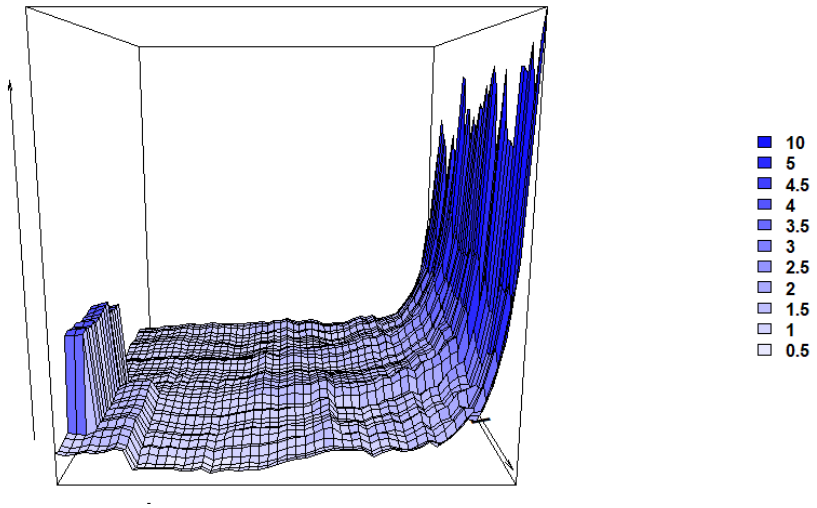


Figure 4.12: Spatio-Temporal Pair Correlation Function (Viewpoint 2).

## 4.3 Topological analysis

Our topological analysis tool is to calculate a persistence diagram, which is derived from the corresponding persistence matrix for each dataset. Similarly to the initial spatial analysis presented in Section 4.2, we generated persistence diagrams for the emergency call dataset classified as `AlarmasMujer` during the year 2018, utilizing the columns `crime_lon` and `crime_lat` (see Figure 4.13). We also generated a diagram for the same dataset based on the Poisson process with the same characteristics, which is shown in Figure 4.14.

Similarly to before, we can visually observe that the two diagrams differ significantly from one another; however, to quantify this difference, we employ the Wasserstein and Bottleneck distances between the diagrams.

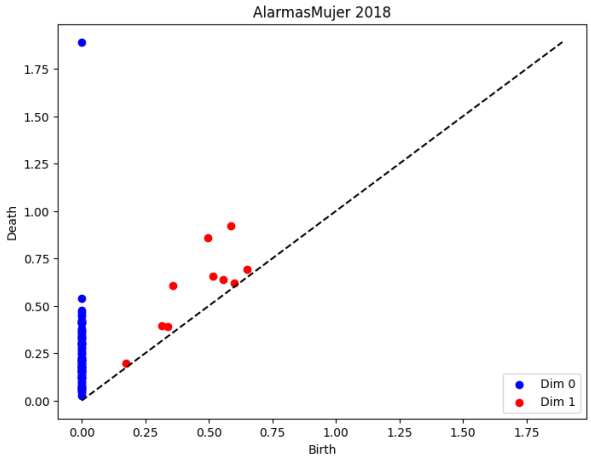


Figure 4.13: Non-random Persistence Diagram.

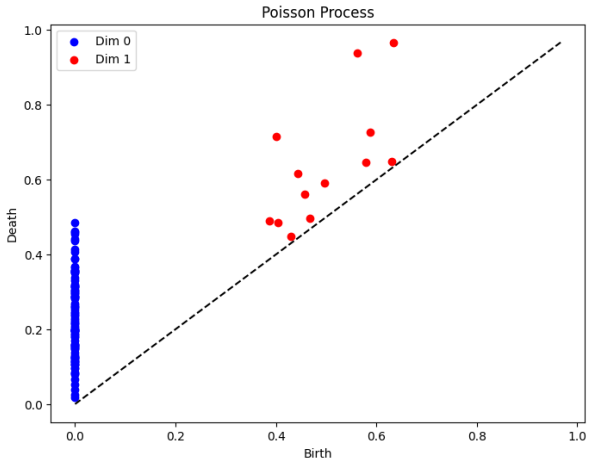


Figure 4.14: Random Persistence Diagram.

As we can observe, we can generate a matrix of Wasserstein distance differences (Figure 4.15) and another for Bottleneck distances (Figure 4.16) between datasets, where each dataset is compared with all others. The diagonal is null, reflecting the symmetric nature of the matrix, as $R_1$ and $R_2$ are two distinct Poisson processes.

We utilized Google Colab and Python, leveraging the `giotto-tda` library to compute persistence diagrams with the `VietorisRipsPersistence` method. For data manipulation, we employed `pandas` and `numpy`. The `gudhi` library was used for the bottleneck distance, while `scipy` calculated the Wasserstein distance. Visualization was accomplished using `seaborn` and `matplotlib` to generate heatmaps and plots.

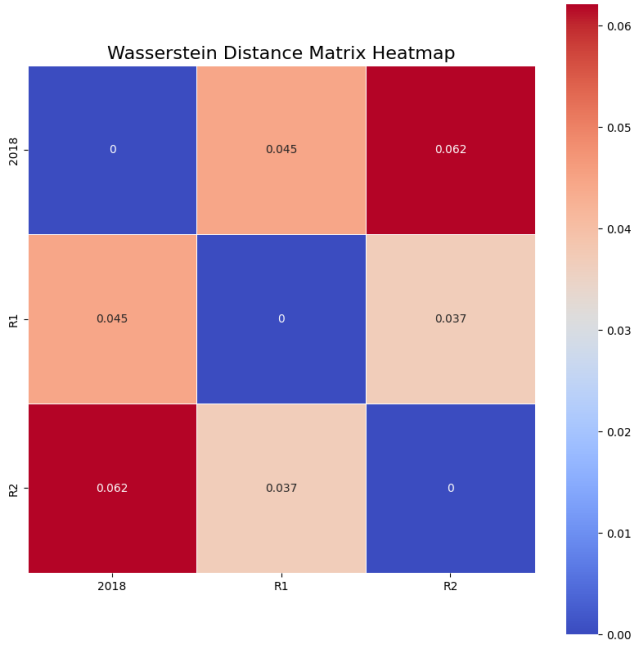


Figure 4.15: Global Differences AlarmasMujer 2018.

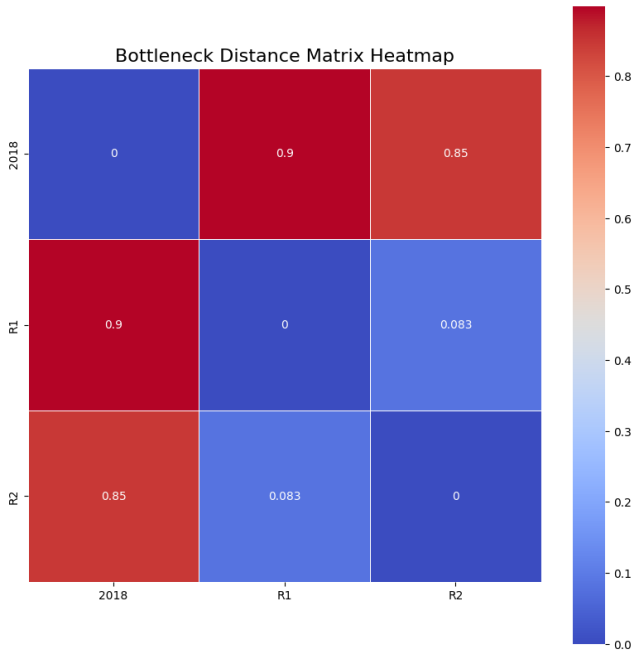


Figure 4.16: Maximum Difference AlarmasMujer 2018.

We obtain interesting values, but how does the 2018 dataset compare to a random one? How can we assess these values across different heatmaps? Additionally, how similar are the randomly generated datasets to each other? Can we find a typical value that reflects this randomness?

To ensure the best calculations and comparisons, we have aimed to always compare datasets that are mutually exclusive (no point is contained in another dataset). All datasets have been sampled to maintain the same number of points; in the case of dataset reduction, the first $n$ values have been taken to preserve the temporal structure.

In the case of the Wasserstein heatmap, in addition to each value obtained from the difference between datasets, we have subtracted a certain 'randomness value' from all values (excluding the diagonal). This randomness value has been calculated separately and prior to each test as the mean of the averages of $j$ experiments of the Wasserstein differences between $m$ datasets of $n$ points. For instance, in the previous case, $n$ would be 95, $m$ would be 3 since we have 3 datasets, and $j$ we would attempt to maximize while considering the necessary computational power (aiming for 100 or 1000 in most cases).

With this adjusted matrix, each value indicates how similar the topological properties are between two datasets, and how close or far these differences are from those expected in random processes. Values close to 0 suggest that the datasets have similar topological properties, with differences comparable to those of random processes. Positive values indicate increasing divergence in topological properties, while negative values suggest a greater similarity than expected from random processes, potentially highlighting shared underlying structures or patterns between the datasets.

Additionally, we have multiplied all these Wasserstein values by 1000 to obtain more manageable numbers for comparison. We have also adjusted the color scale of the heatmap (minimum and maximum values) based on the calculations from all the heatmaps, allowing us to see how significant the differences are between them. With all the 'random values,' we calculated the mean and obtained a value of 9.14 in our system. Therefore, based on our approximations, we can consider that any difference greater than this value indicates that the data exhibit topological dependencies. Moreover, any positive value in our Wasserstein matrices already suggests the presence of topological dependencies in the data. We did not apply this calculation to the bottleneck distances because they are more computationally intensive; however, they still provide complementary information without this procedure.

Finally, before beginning the topological spatial analysis using the `crime_lon` and `crime_lat` data, it is important to explain the possible implications of having distinct topological properties. The Wasserstein distance provides a global measure of the proximity between persistence diagrams. In this case, we analyze dimension 0 classes, which correspond to the connected components of the data, and dimension 1 classes, which represent the loops or holes formed.

In a spatio-temporal context of crime data, similarities in connected components could indicate that different datasets share a comparable clustering behavior, meaning that crimes may be occurring in similarly defined clusters or hotspots across datasets. On the other hand, if the persistence of loops or holes is similar, it may suggest that the spatial layout of crimes shares common structural voids, where areas with few or no crimes exist consistently across datasets. Such patterns could indicate that crime tends to concentrate in certain areas while avoiding others in a recurrent way. These topological features, when consistent between real and synthetic data (such as Poisson processes), could suggest deeper spatial dependencies in the underlying crime distribution.

At the end of this document, we have included in Appendix A, specifically in Appendix I within it, the heatmaps representing the differences between the bottleneck diagrams for each test. This was done to avoid taking up too much space in the main text. Instead, we will focus more on the Wasserstein heatmaps, which provide a more general overview of the topological differences between datasets, while the bottleneck diagrams will be particularly useful in practice to assess how similar two datasets are. If the maximum distance in the bottleneck diagram is very low, the datasets will be quite similar.

In this first case, we calculated a Wasserstein heatmap showing the differences between datasets separated by hours, from 0 to 23. The dataset of Valencia was divided by hour, and persistence diagrams were calculated using the variables 'crime_lat' and 'crime_lon'. In Figure 4.17, we observe that all the values are positive, indicating data aggregation.

Regarding the lower values between wasserstein distances, which represent datasets that, despite the aggregation, are more topologically similar, we can see that the most similar hours are 20:00 and 14:00, typical times in Spain when people tend to have meals or stay at home (less movement). We also find low values between 17:00 and 13:00 or 14:00.

As for the higher values, there is a noticeable difference between 05:00 and 10:00, and we observe that the distances from 02:00 to 07:00 are generally higher compared to other times of the day. This suggests that crimes during these nighttime hours exhibit significant aggregation and may form distinct clusters, following a specific pattern, while during the rest of the day, there is less aggregation and the datasets are more topologically similar.

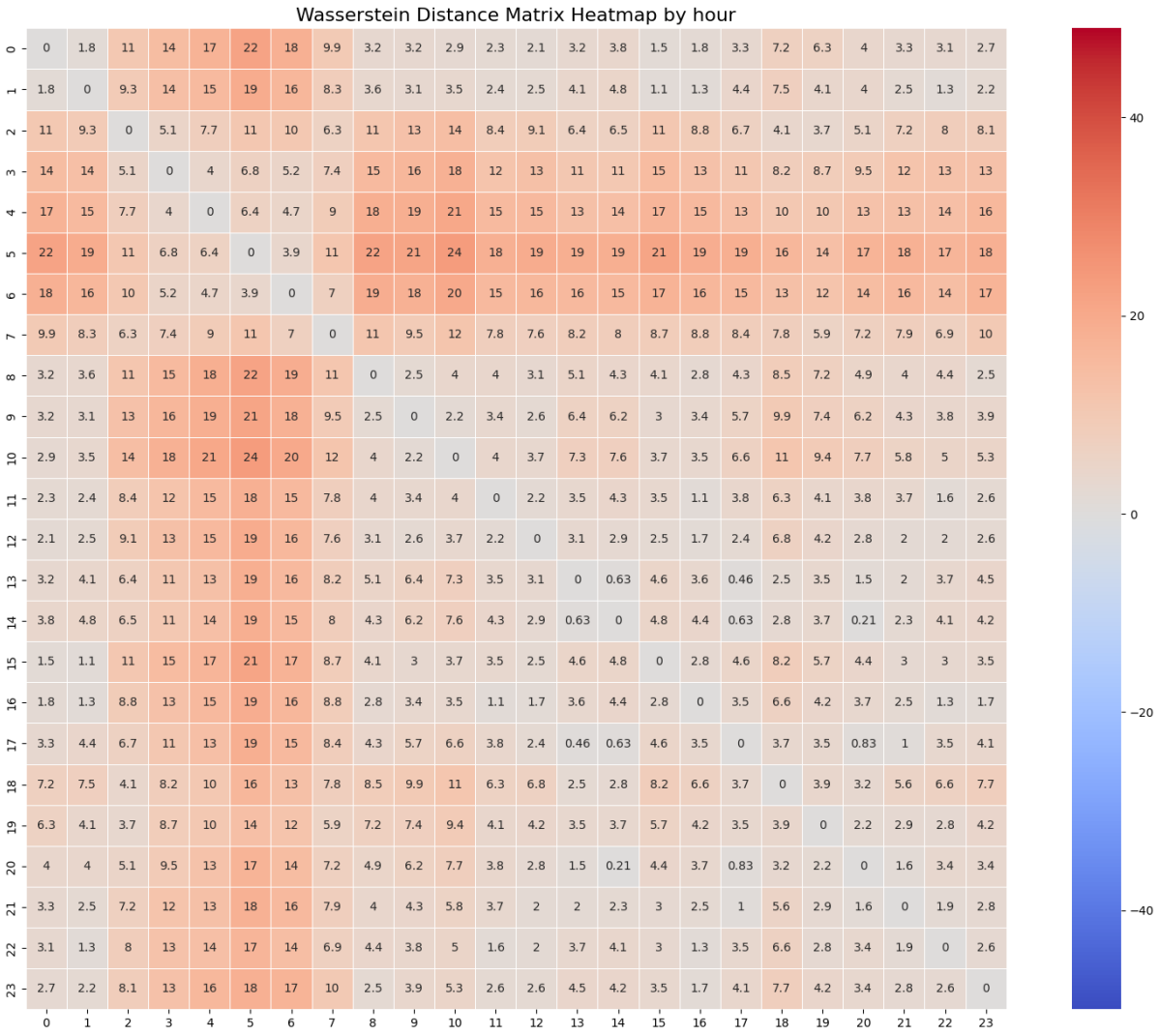


Figure 4.17: Wasserstein Heatmap of Global Distances: Hourly Analysis in Valencia.

In this second case, we calculated a Wasserstein heatmap showing the differences between datasets separated by years, from 2012 to 2019. The entire dataset of Valencia was divided by year, and persistence diagrams were calculated using the variables 'crime_lat' and 'crime_lon'. In Figure 4.18, we observe that all the values are positive, indicating data aggregation. Generally, the topological differences from one year to the next are minimal; however, as we compare 2013 with 2017, 2018, or 2019, the changes become significantly noticeable. This suggests that while year-to-year variations may be subtle, the topology of crime patterns evolves over time, leading to a more distinct transformation in the overall structure.


Figure 4.18: Wasserstein Heatmap of Global Distances: Annual Analysis in Valencia.

In this third case, we present a Wasserstein heatmap illustrating the differences between datasets separated by months, from January to December. The entire dataset of Valencia was segmented by month, and persistence diagrams were computed using the variables 'crime_lat' and 'crime_lon'. In Figure 4.19, we observe that all values are positive, indicating aggregation in the data. Generally, we conclude that the most distinct month is March, which coincidentally is when Valencia celebrates its Fallas festival. This celebration significantly alters the topological structure of crime incidents. Additionally, we can observe notable topological differences between the most representative months of each season: December, August, October, and May, further emphasizing the influence of seasonal variations on the distribution and nature of criminal activity throughout the year.


Figure 4.19: Wasserstein Heatmap of Global Distances: Monthly Analysis in Valencia.

In the fourth case, we analyze a heatmap of Wasserstein differences between datasets separated by days of the week, from Monday to Sunday. In this case, the entire dataset of Valencia has been divided according to the days of the week, and persistence diagrams have been calculated using the variables `'crime_lat'` and `'crime_lon'`. As we can observe in Figure 4.20, all values are positive, indicating aggregation in the data. In general, we can conclude that there is a significant topological difference between the weekend and weekdays.

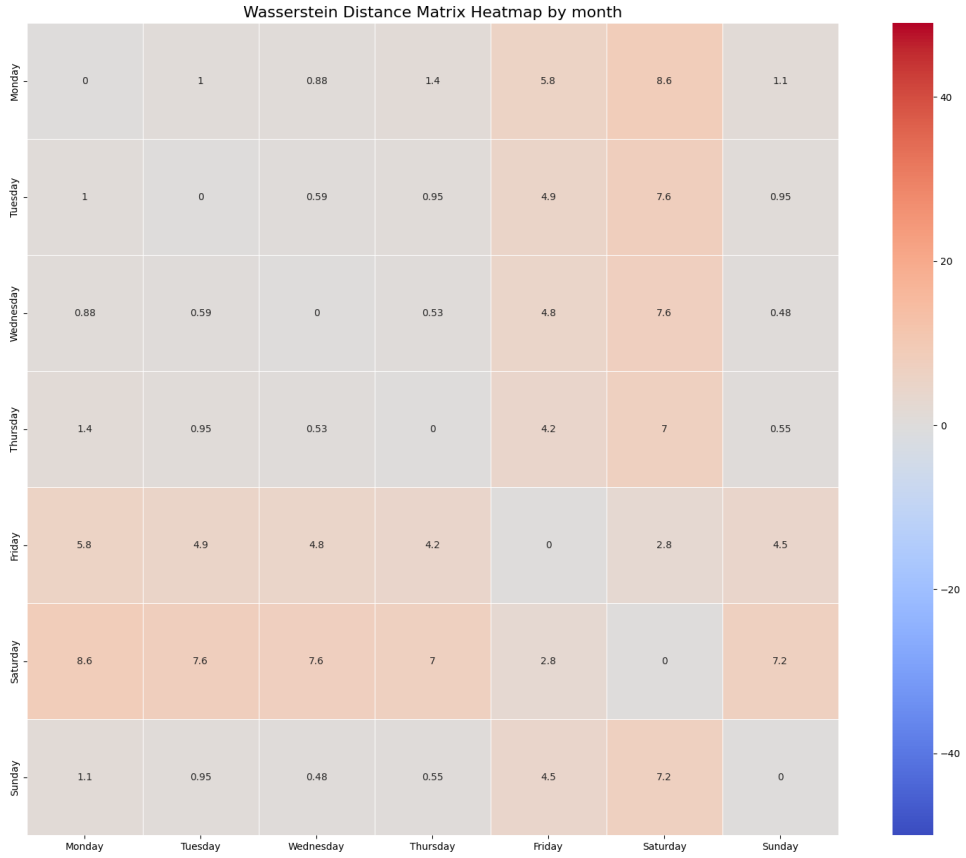


Figure 4.20: Wasserstein Heatmap of Global Distances: Weekly Analysis in Valencia..

We remind the reader that in the final appendix, we can find the respective bottleneck distance heatmaps, which are not bounded by maximum and minimum values across all heatmaps. Therefore, the color coding provides a concrete view of the differences in the diagrams. Additionally, we highlight the maximum difference and leave it to the reader to compare anything they wish, as we have only emphasized the most notable aspects of each case.

In the fifth case, we utilize an innovative approach to calculate the persistence diagrams. We start with the partitioning of the AlarmasMujer crime dataset from the year 2015, which contains 24 entries. It is important to note that the dataset includes more columns than just `crime_lon` and `crime_lat`. For example, we have columns measuring the distance to various landmarks, such as ATMs, banks, bars, cafes, industrial areas, markets, nightclubs, police stations, pubs, restaurants, and taxi stands. These distances are expressed in meters.

The idea behind this fifth case is to generate a separate dataset for each landmark distance, where each dataset consists of three columns: `crime_lon`, `crime_lat`, and the respective landmark distance. We have also taken into account the minimum and maximum values of these

data for generating the Poisson random variable for subtraction. In generating this value, we followed the same scheme as the experiment, using the same dataset but adding a column with random values between the minimum and maximum found. We then compute the differences to obtain a representative average. We calculate the persistence diagram for these three variables and compare it with datasets that have the same spatial variables but vary in the third column. This allows us to observe how a third variable impacts the spatial data.

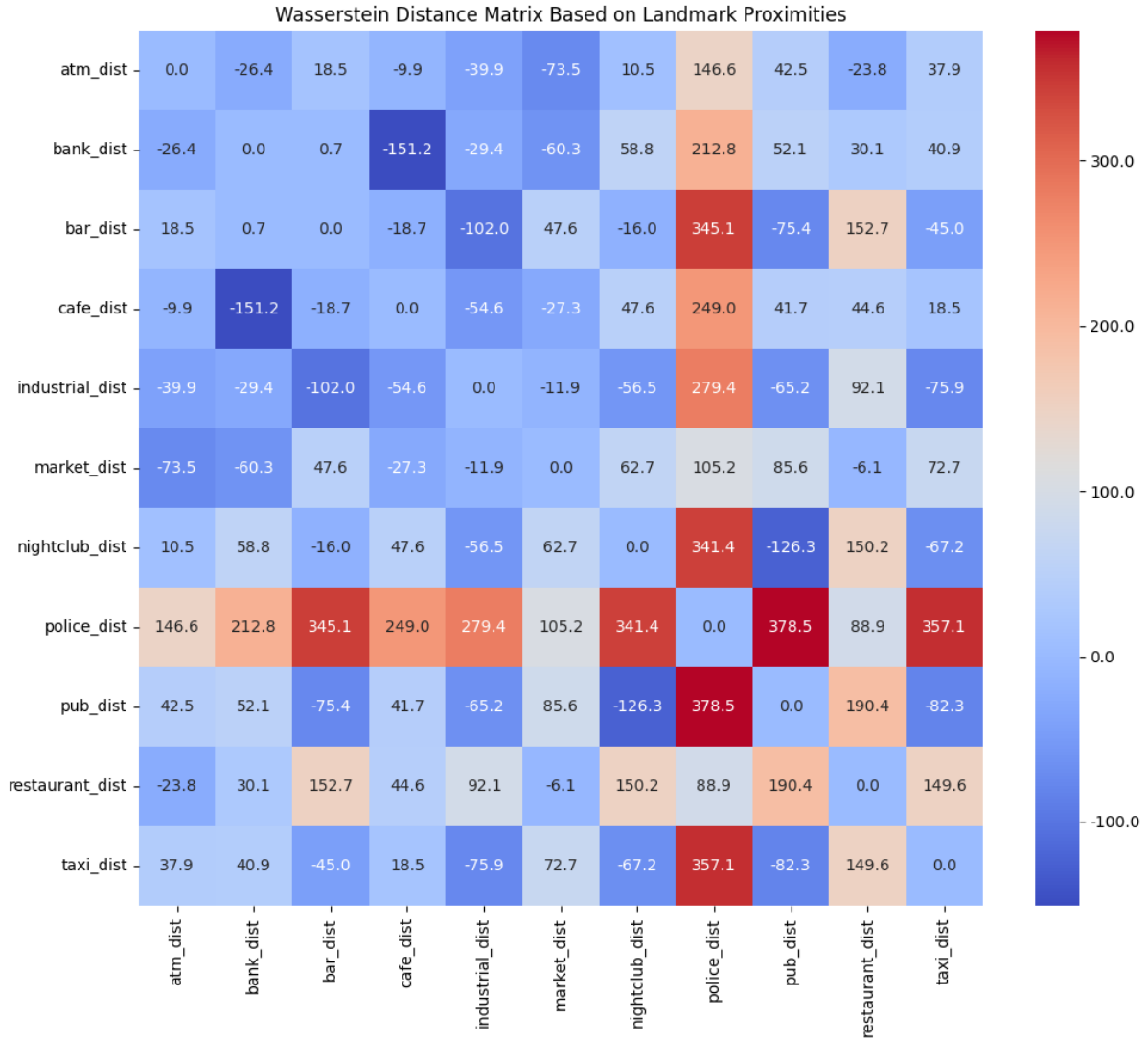


Figure 4.21: Wasserstein Heatmap of Global Distances to Landmarks in Valencia.

In the fifth case, as shown in Figure 4.21, we observe both positive and negative values. Recall that positive values indicate clustering, while negative values suggest dispersion in relation to the difference from Poisson random processes. Among the lower values, we can see that the distance between cafes and banks is the smallest, meaning their topological properties are quite similar. Another notable observation is the low topological distance between nightclubs and pubs, which makes sense, as the topological patterns formed by crimes related to these landmarks may be quite similar. On the positive side, the greatest difference is found between `pub_distance` and `police_distance`, indicating that the topological properties between these two variables differ significantly.

When we add a third temporal variable to our spatial coordinates 'crime_lon' and 'crime_lat', we obtain a persistence diagram with spatio-temporal data. This allows us to detect topological features in dimensions 0, 1, and 2. Dimension 0 represents connected components, indicating clusters of crimes that occur close to each other in both space and time. Dimension 1 corresponds to loops, which suggest recurring patterns or cycles in the crime data. Dimension 2 captures voids or cavities, highlighting gaps in the distribution of crime incidents across space and time. By including time as a variable, the persistence diagram offers deeper insights into the structure of criminal activity, showing how events are distributed not only in space but also temporally.

In Figure 4.22, we present a persistence diagram generated from the so-called 2018 Alarmas-Mujer dataset, where we used the variables 'crime_lon', 'crime_lat', and 'day'. We observe the emergence of some persistent classes in dimension 2, indicating significant topological voids or gaps in the spatio-temporal distribution of the crime incidents.

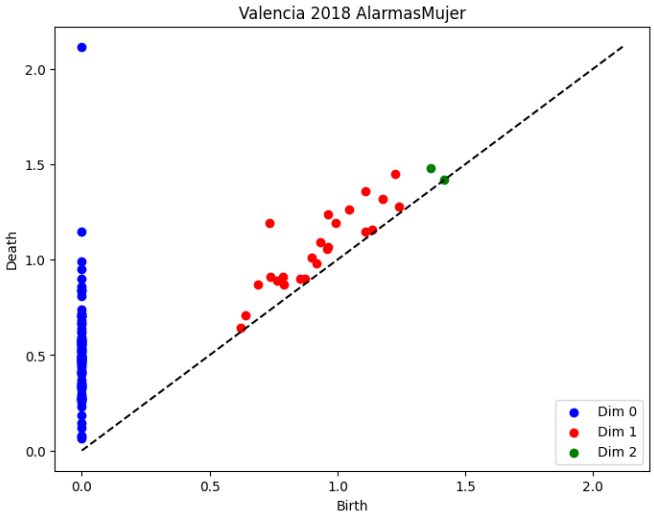


Figure 4.22: Spatio-temporal Persistence Diagram.

We have generated two heatmaps representing the Wasserstein distances between the Valencia AlarmasMujer datasets from the years 2018 and 2019. The first heatmap, shown in figure 4.23, was calculated using only spatial data, specifically the `crime_lat` and `crime_lon` coordinates. The second heatmap, presented in figure 4.24, incorporates both spatial and temporal data, utilizing `crime_lat`, `crime_lon`, and `days` to account for the temporal component. Each heatmap was adjusted by subtracting a respective random value from the distances to ensure consistency in the comparison.

The obtained distances are positive and relatively high, indicating that we are dealing with datasets that exhibit significant aggregation and differing topologies. In the spatial distance analysis, we found a value of 197, while the temporal distance yielded a higher value of 279. This suggests that the datasets from 2018 and 2019 possess distinct characteristics, reflecting variations in their spatial and temporal distributions.

Reflecting on the addition of a third dimension in the data analysis, incorporating this extra dimension could provide a better understanding of the datasets. The inclusion of a temporal component allows us to visualize aspects that may remain hidden in two-dimensional projections.

However, despite this added complexity, the high degree of aggregation is still well-preserved in both analyses, as indicated by the relatively high distances obtained. This emphasizes the robustness of the aggregation patterns present in the data, regardless of the dimensionality considered.

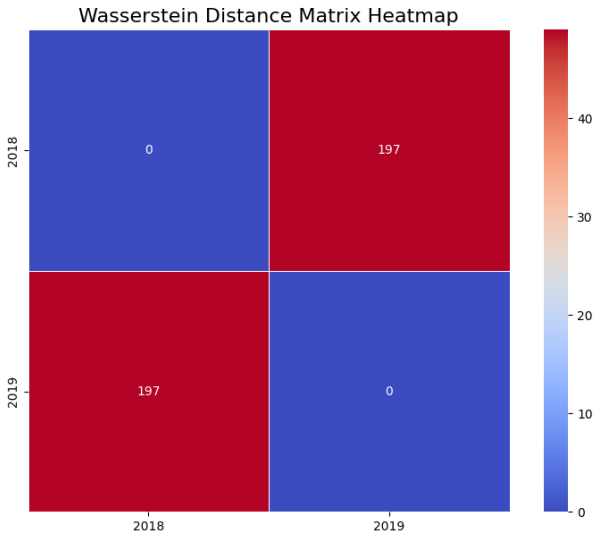


Figure 4.23: Spatial Topological Analysis.

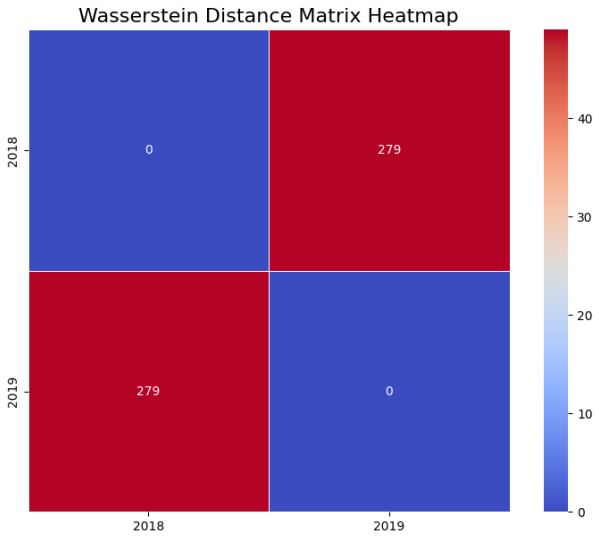


Figure 4.24: Spatio-Temporal Topological Analysis.

## 4.4 Conclusions

As we have seen throughout the article, especially in Section 4.2 from the perspective of spatiotemporal analysis, and in Section 4.3 from the perspective of topological analysis, using the example of the AlarmasMujer 2018 dataset, these two approaches can be related and lead to similar conclusions regarding the aggregation or dispersion of the data.

We observed how the generation of a Poisson random process served as our reference point in both analyses, allowing us to determine how far or close the data is from randomness. In the case of topological analysis, we established a metric where the value indicates the degree of aggregation or dispersion (negative values indicating dispersion, positive values indicating aggregation, and values near zero representing complete randomness).

Thus, not only have we demonstrated the utility of topological data analysis, but we have also obtained valuable and interesting insights (particularly with respect to topological distances between datasets) that could not have been achieved through spatiotemporal point process analysis alone.

In our case, we identified four scenarios generated by two variables and their two possible outcomes (aggregation/dispersion) and (shorter/longer topological distance). As a result, the spectrum of possible solutions is reduced to combinations of these cases, providing a clearer structure for interpreting the relationships between different datasets and their spatial and topological characteristics.

The first scenario involves aggregation and shorter topological distance. In this case, the data points exhibit a high degree of spatial aggregation, meaning that the events are concentrated in specific regions. Additionally, the shorter topological distance between datasets implies that the underlying topological features are similar, suggesting similar clustering patterns over different datasets or time periods. For example, this could represent repeated crime incidents in the same urban hotspots, which might indicate persistent social or environmental factors driving these events in those areas.

The second scenario is characterized by aggregation and longer topological distance. Here, the data is spatially aggregated, but the topological distance between datasets is large, indicating that while the events are concentrated, the topological features differ significantly across datasets. This could occur if crime hotspots shift between datasets, with each dataset showing distinct clusters in different geographic areas or at different spatial scales. This pattern might reflect shifting socio-economic or urban dynamics that influence crime locations over time.

The third scenario shows dispersion and shorter topological distance. This means that the data points are more spatially dispersed, with events being more evenly distributed across the study region. However, the shorter topological distance suggests that the distribution patterns remain topologically similar between datasets. In this case, although crimes are not concentrated in specific areas, the overall spatial structure is consistent across different datasets, potentially reflecting general uniformity in the occurrence of incidents across a wider area, such as evenly distributed low-intensity crime throughout a city.

Finally, the fourth scenario involves dispersion and longer topological distance. Here, the data points are spatially dispersed, and the topological distance between datasets is large. This suggests a marked difference in the spatial distribution patterns between datasets. The crime incidents could be widespread and occur in distinct and different patterns across the datasets, showing no clear clustering or consistency over time. This scenario might represent a city-wide dispersion of incidents with no persistent hotspots, potentially indicating random or unpredictable occurrences of crime over time and space.

Furthermore, the topological analysis has allowed us to identify which variable among a set of variables has the most significant topological impact in terms of aggregation or dispersion relative to other fixed variables (in our case, the spatial variables). This provides a deeper insight into the relationships between variables, offering a more complex understanding of their interactions.

In contrast, the spatiotemporal analysis is limited to a maximum of three variables and is constrained in its input, typically focusing only on spatial and temporal dimensions. This limitation reduces its ability to capture the full complexity of data interactions compared to the broader scope of topological analysis.

In conclusion, Topological Data Analysis serves as a valuable example of how to enhance traditional global analyses of spatiotemporal point processes. By offering unique perspectives on data, TDA can be crucial for obtaining more insightful information. This enhanced analysis not only aids in understanding the underlying patterns of crime but also equips decision-makers with the tools needed to better control and anticipate criminal activity in urban environments.

Ultimately, the insights derived from TDA can contribute to saving lives and improving the overall quality of life for residents in cities. Moreover, the versatility of TDA allows it to be applied creatively in a wide range of fields and contexts, limited only by the imagination of those who use it.

## 4.5 Next Steps

I trust that you have found the concepts presented in this article both engaging and insightful. In this section, I aim to offer several ideas and potential next steps for further exploration in this dynamic field of study, while also discussing innovative ways to integrate topological analysis into various research domains. The investigation of topological data analysis opens up exciting avenues for enhancing our understanding of complex datasets and refining traditional analytical methodologies.

Numerous tools in topological analysis remain to be explored and assessed for their utility. For example, the generation of persistence diagrams can benefit from incorporating multiple variables and employing dimension reduction techniques such as Principal Component Analysis (PCA). Additionally, there are innovative approaches to constructing simplices that surpass the computational efficiency of traditional methods like Vietoris-Rips. These include techniques that offer advantages through cubic or square complexes. Such methods can effectively reduce noise within a dataset and stabilize the resulting diagrams by employing techniques like kernel density estimators and distance measures.

Moreover, vectorizations and the utilization of landscapes present compelling opportunities for exploring their connections with statistics and machine learning. This exploration can extend to concepts such as silhouettes and time series analysis, further enriching our understanding of data dynamics.

Ultimately, there remains a vast array of concepts to delve into and comprehend more deeply. An intriguing avenue for future research could involve the inverse mapping of the most persistent classes within persistence diagrams back to the specific points that generate those classes. This approach could illuminate which points contribute most significantly to the most persistent classes and enable us to identify the spatial areas or hotspots that are particularly valuable for analysis.

It is my aspiration that if any aspect of this discussion resonates with you, you will take the initiative to further investigate and broaden your knowledge. Ultimately, I hope to have encouraged a re-evaluation of our analytical practices and to inspire a consideration of the complementary use of these tools to enhance our performance in research.

# Bibliography


[1] Varun Chandola and Vipin Kumar. Topological data analysis for anomaly detection in network traffic. *Proceedings of the Workshop on Big Data Analytics and Machine Learning for Data Communication Networks*, 2018.

[2] Kathleen E. Turner, Sayan Mukherjee, and Doug M. Boyer. A survey of topological data analysis methods in machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2020.

[3] Peter Bubenik. Perspectives on persistence landscapes. *Journal of Applied and Computational Topology*, 2021.

[4] Robert Ghrist. Topological methods for spatial data analysis. *arXiv preprint arXiv:1806.06232*, 2018.

[5] T.J. Nelson et al. Topological data analysis for temporal data. *arXiv preprint arXiv:1902.03924*, 2019.

[6] J.M.P. Pratola et al. Persistent homology of spatial point processes. *arXiv preprint arXiv:2001.00585*, 2020.

[7] Jonatan A. González, Francisco J. Rodríguez-Cortés, Ottmar Cronie, and Jorge Mateu. Spatio-temporal point process statistics: A review. *Stochastic Processes and their Applications*, 2021.

[8] Simon Zhang, Mengbai Xiao, and Hao Wang. Gpu-accelerated computation of vietoris-rips persistence barcodes. *Journal of Computational Geometry*, 2022.

[9] David Payares-Garcia, Javier Platero, and Jorge Mateu. A dynamic spatio-temporal stochastic modeling approach of emergency calls in an urban context. *Statistical Modelling*, 2022.

# Appendix A

# Appendix I

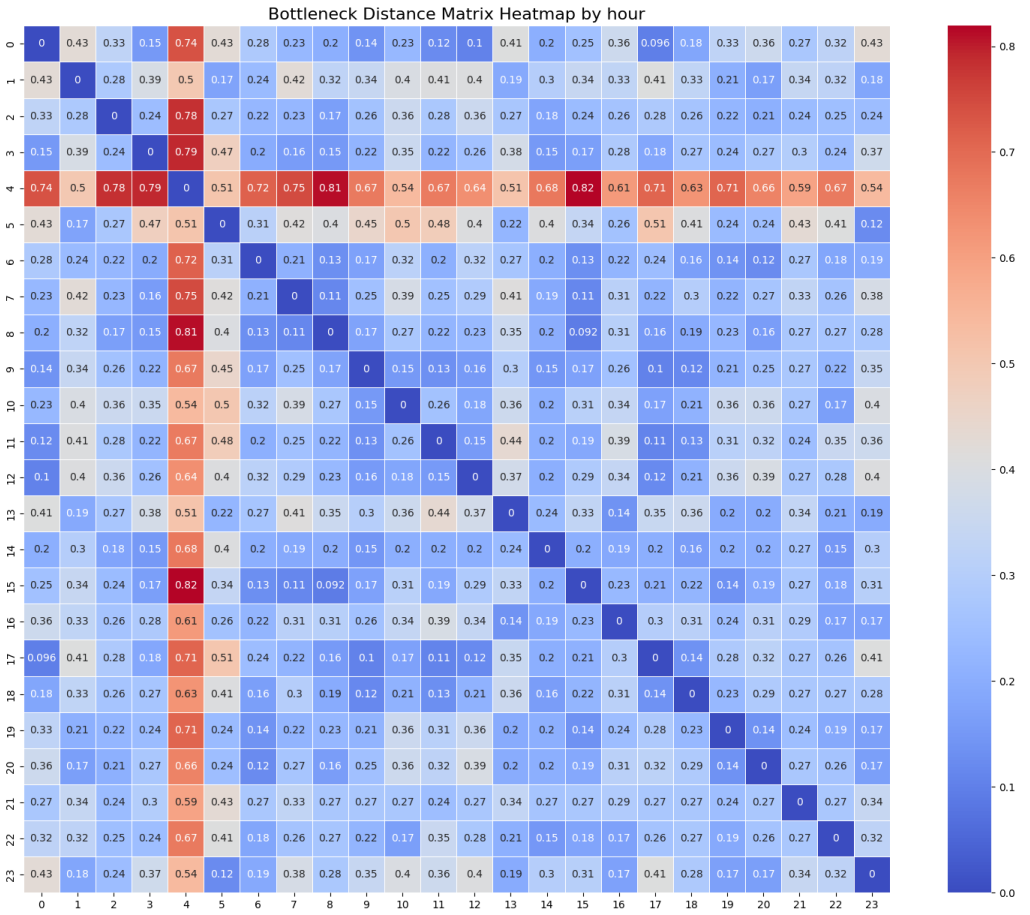


Figure A.1: Bottleneck Heatmap of Maximum Distances: Hourly Analysis in Valencia.

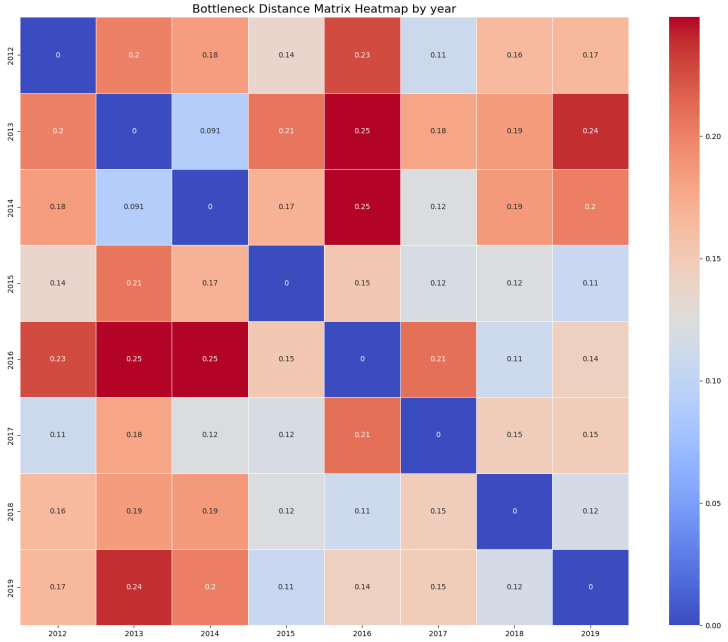


Figure A.2: Bottleneck Heatmap of Maximum Distances: Monthly Analysis in Valencia.

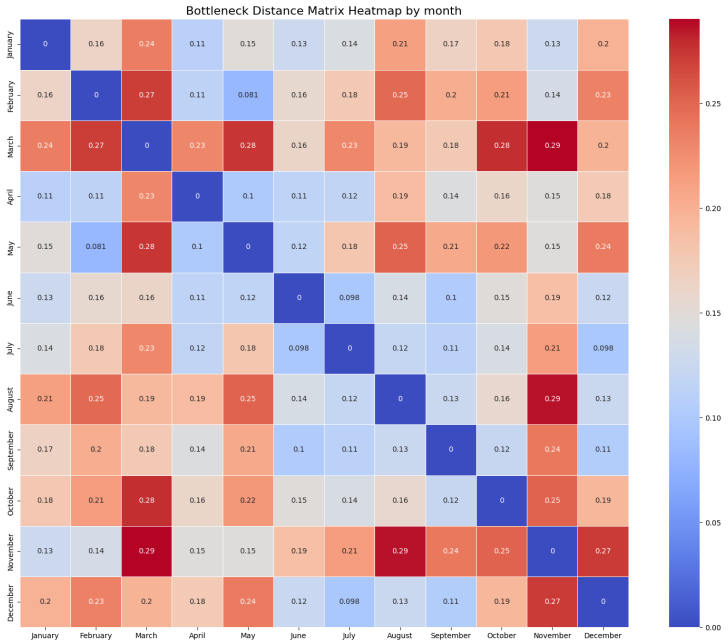


Figure A.3: Bottleneck Heatmap of Maximum Distances: Weekly Analysis in Valencia.

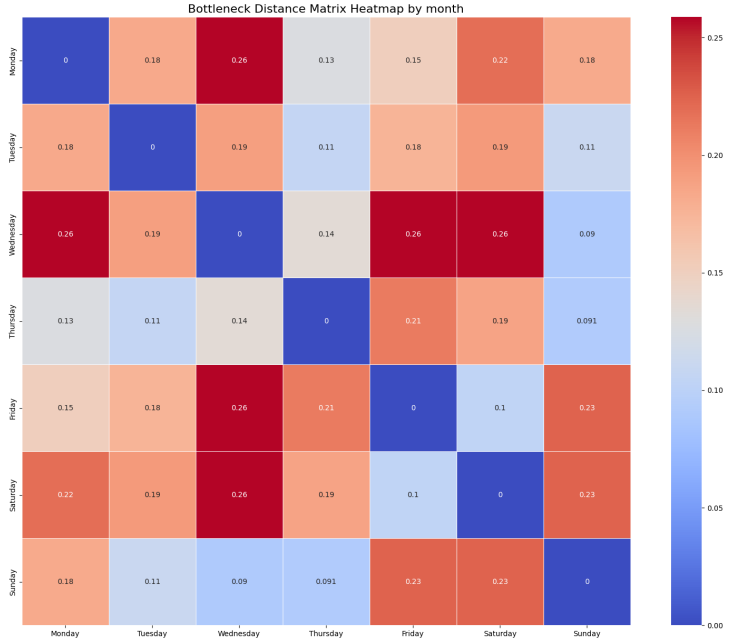


Figure A.4: Bottleneck Heatmap of Maximum Distances to Landmarks in Valencia.