Universidad Politécnica
de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Máster Universitario en Inteligencia Artificial

Trabajo Fin de Máster

Análisis Topológico de Datos Espacial de Covid 19 en España

Autor: Iván Andrada Mendoza
Tutores: Alfonso Mateos Caballero, Arminda Moreno Díaz

Madrid, julio de 2024

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

Trabajo de Fin de Máster
Máster Universitario en Inteligencia Artificial

Título: Análisis Topológico de Datos Espacial de Covid 19 en España

julio de 2024

Autor: Iván Andrada Mendoza
Tutores: Alfonso Mateos Caballero, Arminda Moreno Díaz
Departamento de Inteligencia Artificial
ETSI Informáticos
Universidad Politécnica de Madrid

Agradecimientos

Quiero agradecer a mi madre y a mis abuelos por su inquebrantable apoyo y esfuerzos incansables en cada paso de mi camino hacia la realización de mis sueños y metas.

Quisiera expresar mi más sincero agradecimiento a las personas que han contribuido de manera significativa a la realización de este proyecto. En primer lugar, deseo agradecer a mis tutores, Alfonso Mateos y Arminda Moreno, cuyo apoyo, orientación y valiosos consejos han sido fundamentales para el desarrollo y la finalización de este proyecto. Su experiencia y conocimientos me han guiado en cada etapa del proceso, desde la conceptualización inicial hasta la elaboración final.

A Alfonso, gracias por su incansable dedicación y paciencia, así como por su habilidad para proporcionar claridad y dirección en momentos críticos. Su enfoque riguroso y su pasión por la investigación han sido una inspiración constante para mí.

A Arminda, agradecer su apoyo continuo y su capacidad para ofrecer perspectivas innovadoras y soluciones prácticas. Su entusiasmo y compromiso con mi progreso han sido invaluables y han enriquecido enormemente mi experiencia de aprendizaje.

Por otro lado, quiero agradecer al Ministerio de Ciencia e Innovación la financiación del proyecto PID2021-122209OB-C31.

Quiero también expresar mi profunda gratitud a Dios, Nuestro Señor, fuente de todo conocimiento y sabiduría en mi vida. Agradezco su infinita bondad y gracia, que me han permitido llegar hasta donde estoy.

Ad maiorem Dei gloriam.

Resumen

Este proyecto aborda la aplicación del Análisis Topológico de Datos (TDA) para caracterizar la propagación espacial del COVID-19 en España. La pandemia de COVID-19 ha desafiado a la comunidad científica a desarrollar métodos innovadores para entender y mitigar su impacto. En este contexto, el TDA emerge como una herramienta poderosa para analizar conjuntos de datos de alta dimensionalidad, revelando estructuras y patrones que las técnicas tradicionales pueden pasar por alto.

El estudio se centra en utilizar el TDA para analizar datos geoespaciales y temporales de casos confirmados y muertes por COVID-19 en España. Mediante la construcción de una nube de puntos en un espacio de cuatro dimensiones (latitud, longitud, fecha y número acumulado de casos o muertes) y la aplicación del algoritmo Mapper, se obtiene una representación topológica de la propagación del virus. Los descriptores topológicos, como la homología persistente y los diagramas de barras de persistencia, permiten identificar características clave en la dinámica de la pandemia a diversas escalas.

Los resultados de este análisis topológico revelan patrones de transmisión y puntos críticos de infección que proporcionan una comprensión más profunda de la propagación del COVID-19 en el contexto geográfico español. Estos hallazgos pueden ayudar a mejorar las estrategias de control y prevención y a evaluar la efectividad de las medidas de salud pública implementadas.

El proyecto no sólo valida la utilidad del TDA en el estudio de pandemias, sino que también ofrece un marco metodológico robusto para futuras investigaciones en epidemiología y salud pública. La integración del TDA en el análisis epidemiológico puede proporcionar herramientas valiosas para la toma de decisiones informadas y la gestión efectiva de enfermedades infecciosas.

Abstract

This project addresses the application of Topological Data Analysis (TDA) to characterize the spatial spread of COVID-19 in Spain. The COVID-19 pandemic has challenged the scientific community to develop innovative methods to understand and mitigate its impact. In this context, TDA emerges as a powerful tool for analyzing high-dimensional data sets, revealing structures and patterns that traditional techniques may overlook.

The study focuses on using TDA to analyze geospatial and temporal data of confirmed cases and deaths from COVID-19 in Spain. By constructing a point cloud in a four-dimensional space (latitude, longitude, date, and cumulative number of cases or deaths) and applying the Mapper algorithm, a topological representation of the virus's spread is obtained. Topological descriptors, such as persistent homology and persistence barcodes, allow identifying key characteristics in the pandemic's dynamics across various scales.

The results of this topological analysis reveal transmission patterns and critical infection points that provide a deeper understanding of the COVID-19 spread in the Spanish geographical context. These findings can help improve control and prevention strategies and evaluate the effectiveness of implemented public health measures.

The project not only validates the utility of TDA in pandemic studies but also offers a robust methodological framework for future research in epidemiology and public health. Integrating TDA into epidemiological analysis can provide valuable tools for informed decision-making and effective management of infectious diseases.

Tabla de contenidos

Capítulo 1

Introducción

La pandemia de COVID-19 ha representado uno de los desafíos más significativos para la salud pública global en la historia reciente [1, 2]. Desde su aparición en diciembre de 2019, el virus SARS-CoV-2 ha desencadenado una crisis sanitaria sin precedentes, afectando a millones de personas en todo el mundo y causando profundas repercusiones en todos los aspectos de la vida humana [3]. En España, al igual que en muchos otros países, la propagación del virus ha sido rápida y extendida, provocando una necesidad urgente de desarrollar métodos eficaces para entender y mitigar la propagación de la enfermedad [4].

En este contexto, el análisis de datos se ha convertido en una herramienta fundamental para comprender la dinámica de la pandemia y apoyar la toma de decisiones informadas. Particularmente, el Análisis Topológico de Datos (TDA, por sus siglas en inglés) ha emergido como una metodología poderosa en la ciencia de datos para describir conjuntos de datos de alta dimensionalidad mediante la forma y estructura que estos datos preservan [5, 6]. El TDA permite capturar y analizar la geometría y topología de los datos, proporcionando una visión más profunda de las relaciones complejas dentro de los conjuntos de datos.

El presente Trabajo de Fin de Máster (TFM) tiene como objetivo aplicar el Análisis Topológico de Datos para caracterizar la dinámica espacial del COVID-19 en España. Mediante esta metodología, se pretende identificar patrones y descriptores topológicos que capturen información relevante sobre la propagación del virus y sus características particulares en el contexto geográfico español. Este enfoque no solo busca entender mejor la distribución espacial del COVID-19, sino también revelar información visual valiosa que pueda contribuir al desarrollo de estrategias de control y prevención más efectivas.

1.1. Objetivos del TFM

El objetivo principal de este TFM es caracterizar la dinámica espacial del COVID-19 a nivel de España mediante el Análisis Topológico de Datos. Para alcanzar este objetivo, se establecen los siguientes objetivos específicos:

- **Recolección de datos:** Obtener datos de series temporales de casos confirmados y muertes por COVID-19 a nivel de comunidades autónomas y ciudades principales en España.

- **Construcción de la nube de puntos:** Crear una nube de puntos en un espacio de cuatro dimensiones ($\mathbb{R}^4$), considerando las variables de latitud, longitud, fecha y número acumulado de casos. Este procedimiento se repetirá también para el número de fallecidos.

- **Aplicación del algoritmo Mapper:** Utilizar el algoritmo Mapper de TDA para describir la información topológica de las relaciones entre los casos y muertes por COVID-19 en España.

- **Análisis de resultados:** Interpretar los resultados obtenidos del análisis topológico para extraer conclusiones sobre la propagación del virus y su dinámica espacial en el territorio español.

1.2. Relevancia y contribuciones en epidemiología

El Análisis Topológico de Datos ha emergido como una herramienta poderosa en el campo de la epidemiología, especialmente en el estudio de enfermedades infecciosas como el COVID-19. Esta metodología, que se centra en la estructura subyacente y las características topológicas de los conjuntos de datos, es capaz de identificar patrones y conexiones en datos de alta dimensionalidad que son difíciles de detectar mediante métodos estadísticos tradicionales. Esto es crucial en epidemiología, donde los patrones de transmisión de enfermedades y las agrupaciones de casos pueden no ser inmediatamente evidentes.

Además, con el aumento de los datos de salud a gran escala, el TDA proporciona un método eficaz para manejar y analizar grandes volúmenes de información. Esto ayuda a los epidemiólogos a obtener insights valiosos sobre la dinámica de las enfermedades sin perderse en la complejidad de los datos. El TDA permite modelar la propagación de enfermedades de manera que se preservan las relaciones geoespaciales y temporales, lo que es esencial para comprender cómo las intervenciones de salud pública afectan la dinámica de una epidemia.

Entre las referencias clave que ilustran la aplicación del TDA en este campo está el estudio de Carlsson et al. [7], que introdujo el algoritmo Mapper. Esta técnica de TDA ha sido adaptada para analizar datos epidemiológicos, permitiendo la visualización de conjuntos de datos complejos en formas simplificadas que conservan información esencial sobre la estructura de los datos. Otros estudios incluyen el trabajo de Nielson y Nock [8], que emplearon TDA para analizar las referencias científicas en Wikipedia, mostrando cómo esta técnica puede aplicarse para descubrir estructuras y patrones en datos no tradicionalmente epidemiológicos, pero relevantes para la salud pública.

Petri y otros colaboradores [9] han utilizado TDA para examinar las redes complejas, proporcionando un marco para el estudio de redes de transmisión de enfermedades y la identificación de comunidades altamente interconectadas susceptibles a brotes. Además, el estudio de Topaz et al. [10] sobre modelos de agregación biológica aplicó TDA para entender cómo los organismos interactúan y forman clústeres, con aplicaciones directas en el estudio de la propagación de enfermedades en poblaciones humanas.

1.3. Estructura

El Trabajo Fin de Máster (TFM) está organizado en seis capítulos principales, cada uno diseñado para explorar diferentes aspectos del análisis topológico de la propagación del COVID-19 en España. La estructura detallada es la siguiente:

- **Capítulo 1: Introducción.** Este capítulo establece el contexto del estudio, delineando los objetivos generales y específicos del TFM. Se aborda la importancia del Análisis Topológico de Datos (TDA) y se explica por qué es pertinente para estudiar la propagación del COVID-19. Además, se destacan las contribuciones clave del trabajo y su relevancia en el contexto actual.
- **Capítulo 2: Análisis Topológico de Datos.** Se detalla la teoría subyacente del TDA, describiendo los conceptos y métodos fundamentales como complejos simpliciales, homología persistente y el algoritmo Mapper. Este capítulo también incluye una revisión del estado del arte en la aplicación del TDA a problemas de salud pública y otras áreas.
- **Capítulo 3: Metodología.** Aquí se describe el proceso metodológico adoptado en el estudio, incluyendo la recolección de datos, el preprocesamiento, la construcción de la nube de puntos en cuatro dimensiones y la implementación del algoritmo Mapper. Se explican con detalle los parámetros utilizados y los métodos de clustering aplicados.
- **Capítulo 4: Resultados.** Se presentan los hallazgos del análisis topológico. Este capítulo detalla la estructura del grafo obtenido a través del Mapper, la interpretación de los descriptores topológicos y cómo estos reflejan la dinámica de la pandemia a nivel geográfico y temporal.
- **Capítulo 5: Discusión.** En este capítulo se compara el enfoque topológico con otras técnicas analíticas y se discuten las ventajas y limitaciones del método utilizado. También se evalúa la aplicabilidad de los resultados obtenidos en la formulación de estrategias de control y prevención.
- **Capítulo 6: Conclusiones y trabajos futuro.** Se resumen los principales hallazgos y se discuten las implicaciones del estudio para la salud pública y la investigación futura. Se identifican las limitaciones del estudio actual y se proponen líneas futuras de investigación basadas en los resultados obtenidos.

Capítulo 2

Análisis Topológico de Datos

El estudio y análisis de datos complejos ha avanzado significativamente con el desarrollo de técnicas de análisis topológico de datos (TDA) y algoritmos de clustering. Entre estos, el algoritmo Mapper y el algoritmo DBSCAN destacan por su capacidad para revelar estructuras y patrones subyacentes en datos de alta dimensión. En esta sección, exploraremos el estado del arte de estas metodologías, revisando las principales contribuciones, avances y aplicaciones en diversos dominios, incluyendo la epidemiología, biología computacional, y análisis de redes. Nos enfocaremos en cómo estas técnicas han sido aplicadas para comprender la propagación de enfermedades como el Covid-19, y cómo han mejorado la capacidad de identificar clusters y anomalías en datos complejos. Este análisis proporciona el contexto necesario para entender la relevancia y la eficacia de las herramientas TDA en la investigación científica y aplicada.

2.1. Análisis topológico de datos - TDA

El Análisis Topológico de Datos es una metodología emergente en la Ciencia de Datos que ha ganado reconocimiento por su capacidad para analizar y extraer información relevante de conjuntos de datos de alta dimensión. A diferencia de las técnicas tradicionales que se centran en aspectos estadísticos o geométricos de los datos, el TDA se enfoca en la topología, es decir, en las propiedades que permanecen invariantes bajo transformaciones continuas. Esta perspectiva permite capturar la estructura global y las relaciones intrínsecas dentro de los datos, proporcionando una visión más completa y profunda de su organización.

Uno de los principales atractivos del TDA es su habilidad para manejar datos complejos de alta dimensionalidad y con ruido, una característica común en muchos dominios científicos y de ingeniería. Utilizando herramientas matemáticas avanzadas, el TDA puede identificar patrones y estructuras que otras técnicas podrían pasar por alto. Esto se logra mediante la construcción de representaciones topológicas, como complejos simpliciales (0-simplex, 1-simplex, 2-simplex), que preservan la conectividad y la forma de los datos. A través de estos complejos, es posible analizar la forma en que los datos se agrupan y se conectan, lo que puede revelar características ocultas y relaciones significativas.

La capacidad del TDA para trabajar con datos de alta dimensión es particularmente

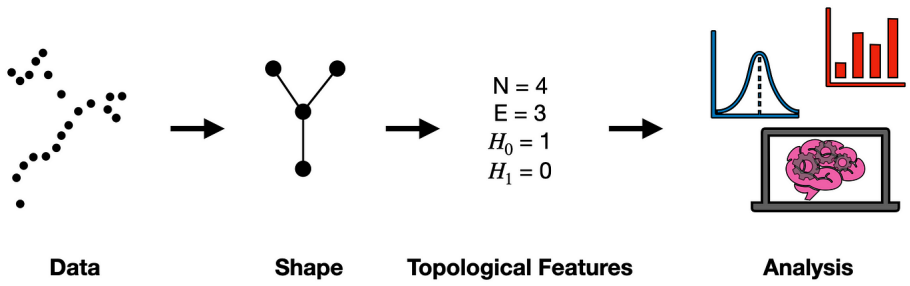

Figura 2.1: Proceso del Análisis Topológico de Datos: Los datos se transforman en una forma, de la cual se extraen características topológicas, que luego se analizan para obtener resultados significativos. [11]

relevante en la era del Big Data, donde los conjuntos de datos masivos y heterogéneos son la norma. Los enfoques tradicionales a menudo enfrentan dificultades para escalar adecuadamente y mantener la precisión en estos contextos. En contraste, el TDA ofrece una forma robusta y escalable de analizar datos complejos, lo que ha llevado a su adopción en una variedad de campos, desde la biología y la medicina hasta la física y las ciencias sociales. En cada uno de estos dominios, el TDA ha demostrado su valor al proporcionar perspectivas novedosas y a menudo inesperadas que mejoran nuestra comprensión de los fenómenos estudiados.

El proceso de TDA típicamente involucra varias etapas clave, comenzando con la recolección y preprocesamiento de los datos. A continuación, se construye una representación topológica de los datos, como un complejo simplicial o un grafo. Esta representación se analiza utilizando técnicas de homología persistente, que permiten identificar y cuantificar características topológicas en múltiples escalas de resolución. Los resultados del análisis se visualizan y se interpretan para extraer conclusiones relevantes (véase Figura 2.1). Este enfoque sistemático y matemáticamente riguroso es lo que permite al TDA ofrecer una visión única y profunda de la estructura de los datos.

El campo del TDA comenzó a tomar forma con los trabajos pioneros de Herbert Edelsbrunner y sus colaboradores [12] a mediados de los años 2000, especialmente en el área de la homología persistente. La homología persistente se popularizó con un artículo clave de Gunnar Carlsson [7], que destacó su potencial para aplicaciones prácticas en el análisis de datos. Desde entonces, TDA ha crecido rápidamente tanto en términos de desarrollo teórico como de aplicaciones prácticas.

2.1.1. Conceptos fundamentales

2.1.1.1. Estructura topológica y geométrica de los datos

En el núcleo de TDA se encuentra la idea de que la topología y la geometría proporcionan una forma potente de inferir información cualitativa robusta sobre la estructura de los datos. La topología se ocupa de propiedades que no cambian bajo deformaciones continuas, como la conectividad y la presencia de agujeros o ciclos. La geometría, por otro lado, se enfoca en las propiedades cuantitativas de los espacios, como la distancia y el ángulo. TDA combina estos enfoques para proporcionar una visión más completa y robusta de los datos.

La estructura topológica y geométrica de los datos es fundamental en el Análisis de Datos Topológicos (TDA). Este enfoque se basa en la idea de que los datos tienen una estructura subyacente que puede ser analizada utilizando conceptos de la topología y la geometría. En esta sección, exploraremos estos conceptos con mayor detalle, utilizando la notación matemática necesaria para una comprensión rigurosa. [9, 13]

Espacios métricos

Un espacio métrico (M, ρ) es un conjunto M junto con una función de distancia $\rho : M \times M \to \mathbb{R}_+$ que satisface las siguientes propiedades para todos $x, y, z \in M$:

1. *No negatividad y separación:*

$$\rho(x, y) \geq 0 \quad \text{y} \quad \rho(x, y) = 0 \iff x = y. \tag{2.1}$$

2. *Simetría:*

$$\rho(x, y) = \rho(y, x). \tag{2.2}$$

3. *Desigualdad triangular:*

$$\rho(x, z) \leq \rho(x, y) + \rho(y, z). \tag{2.3}$$

Asimismo, la métrica ρ induce una estructura topológica en M permitiendo definir conceptos como bolas abiertas, convergencia y continuidad.

Algunos ejemplos de espacios métricos que pueden darse son los siguientes:

1. *Espacio euclidiano:*
   El espacio euclidiano $\mathbb{R}^n$ con la distancia euclidiana es un ejemplo clásico de espacio métrico. La distancia euclidiana entre dos puntos $x = (x_1, x_2, \ldots, x_n)$ e $y = (y_1, y_2, \ldots, y_n)$ en $\mathbb{R}^n$ se define como:

$$\rho(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.4}$$

2. *Espacio discreto:*
   En un espacio discreto, la métrica se define como:

$$\rho(x, y) = \begin{cases} 0, & si \quad x = y, \\ 1, & si \quad x \neq y \end{cases} \tag{2.5}$$

Propiedades de los espacios métricos

1. *Bolas abiertas y cerradas:* Una bola abierta centrada en un punto $x \in M$ con radio $r > 0$ se define como el conjunto de puntos $y \in M$ tales que $\rho(x, y) < r$:

$$B(x, r) = \{y \in M : \rho(x, y) < r\}. \tag{2.6}$$

   Una bola cerrada se define de manera similar, pero con $\leq$ en lugar de $<$:

$$\overline{B}(x, r) = \{y \in M : \rho(x, y) \leq r\}. \tag{2.7}$$

2. *Convergencia:* Una secuencia (x_n) en M converge a un punto $x \in M$ si, para cualquier $\epsilon > 0$, existe un número natural N tal que para todos los $n \geq N$, $\rho(x_n, x) < \epsilon$:
$$\forall \epsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, \rho(x_n, x) < \epsilon. \tag{2.8}$$

3. *Continuidad:* Una función $f : M \to N$ entre espacios métricos (M, ρ_M) y (N, ρ_N) es continua en un punto $x \in M$ si, para cualquier $\epsilon > 0$, existe un $\delta > 0$ tal que si $\rho_M(x, y) < \delta$, entonces $\rho_N(f(x), f(y)) < \epsilon$:
$$\forall \epsilon > 0, \exists \delta > 0, \forall x, y \in M, \rho_M(x, y) < \delta \Rightarrow \rho_N(f(x), f(y)) < \epsilon. \tag{2.9}$$

Distancia de Hausdorff

La distancia de Hausdorff es una métrica definida en el conjunto de subconjuntos compactos de un espacio métrico. Dado un espacio métrico (M, ρ) y dos subconjuntos compactos $A, B \subseteq M$, la distancia de Hausdorff $d_H(A, B)$ se define como:
$$d_H(A, B) = \text{máx} \left\{ \sup_{a \in A} \inf_{b \in B} \rho(a, b), \sup_{b \in B} \inf_{a \in A} \rho(a, b) \right\}. \tag{2.10}$$

Esta métrica mide la mayor distancia que uno debe desplazarse desde un punto en un conjunto hasta el punto más cercano en el otro conjunto, y viceversa. Es útil para comparar la proximidad de diferentes conjuntos de datos en el contexto del TDA.

Distancia de Gromov-Hausdorff

Para comparar espacios métricos que no necesariamente están embebidos en el mismo espacio, se utiliza la distancia de Gromov-Hausdorff. Dos espacios métricos compactos (M_1, ρ_1) y (M_2, ρ_2) son isométricos si existe una biyección $\varphi : M_1 \to M_2$ que preserva distancias:
$$\rho_2(\varphi(x), \varphi(y)) = \rho_1(x, y) \quad \text{para todo } x, y \in M_1. \tag{2.11}$$

La distancia de Gromov-Hausdorff $d_{GH}(M_1, M_2)$ mide cuán cerca están dos espacios métricos de ser isométricos. Se define como el ínfimo de los números reales $r \geq 0$ tales que existen un espacio métrico (Z, ρ) y dos embebimientos isométricos $\varphi_i : M_i \to Z$ (para $i = 1, 2$) tal que la distancia de Hausdorff entre $\varphi_1(M_1)$ y $\varphi_2(M_2)$ en Z es menor o igual a r:
$$d_{GH}(M_1, M_2) = \inf\{d_H(\varphi_1(M_1), \varphi_2(M_2))\}. \tag{2.12}$$

En la Figura 2.2 se visualiza un ejemplo ilustrativo de ambas distancias: Hausdorff y Gromov-Hausdorff.

2.1.1.2. Representación de datos como nubes de puntos

En TDA, los datos a menudo se representan como nubes de puntos en un espacio métrico. Una nube de puntos es simplemente un conjunto finito de puntos en un espacio euclidiano o métrico general. La elección del espacio y la métrica depende de la naturaleza de los datos y del problema en cuestión. La estructura topológica y geométrica de la nube de puntos se infiere construyendo complejos simpliciales, que son combinaciones de simpleses (puntos, líneas, triángulos, etc.) que reflejan la proximidad y la interconexión entre los puntos.

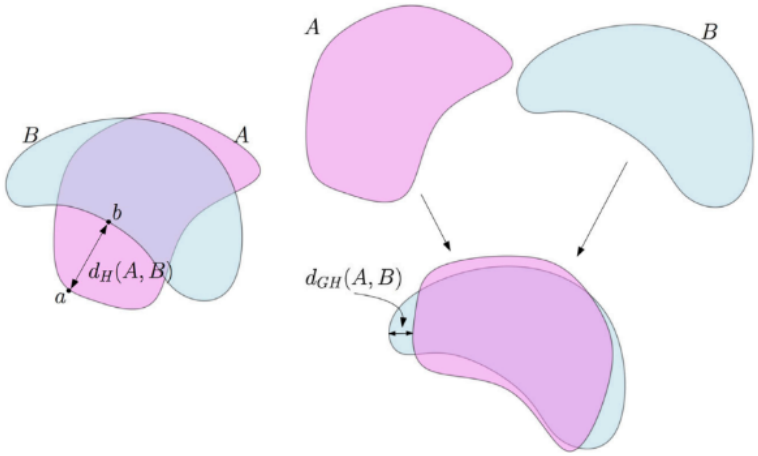

Figura 2.2: Visualización de la distancia de Hausdorff (d_H) y la distancia de Gromov-Hausdorff (d_{GH}) entre dos conjuntos A y B. $d_H(A,B)$ mide la mayor distancia de un punto en un conjunto al conjunto más cercano. $d_{GH}(A,B)$ compara formas de los conjuntos en un espacio métrico. Fuente: [13]

Nubes de puntos

Una nube de puntos es simplemente un conjunto finito de puntos en un espacio métrico, y se utiliza para representar datos de manera geométrica. Formalmente, una nube de puntos X en un espacio métrico (M,ρ) se define como:

$$X = \{x_1, x_2, \ldots, x_n\}, \quad x_i \in M, \quad i = 1, 2, \ldots, n \tag{2.13}$$

donde M es el espacio métrico y ρ es la métrica que define la distancia entre los puntos en M.

Propiedades de las nubes de puntos

Entre las propiedades de las nubes de puntos, podemos encontrar:

1. *Cardinalidad:*
   La cardinalidad de una nube de puntos X es el número de puntos en la nube, denotado por $|X|$. En nuestro caso, $|X| = n$.

2. *Dimensionalidad:*
   La dimensionalidad de la nube de puntos está determinada por el espacio métrico M en el que se localizan los puntos. Por ejemplo, si $M = \mathbb{R}^d$, entonces la nube de puntos tiene una dimensionalidad d.

3. *Disposición geométrica:*
   La disposición geométrica de los puntos en la nube es crucial para el análisis topológico, ya que las relaciones de proximidad entre los puntos se utilizan para construir complejos simpliciales y analizar características topológicas.

2.1.1.3. Simplex y complejos simpliciales

Un k-simplex σ es la envoltura convexa de $k+1$ puntos afínmente independientes $\{v_0, v_1, \ldots, v_k\}$ en $\mathbb{R}^d$. Formalmente, un k-simplex se define como:

$$\sigma = \left\{ \sum_{i=0}^{k} \lambda_i v_i : \lambda_i \geq 0, \sum_{i=0}^{k} \lambda_i = 1 \right\}. \tag{2.14}$$

Los puntos v_i se denominan vértices del simplex, y cualquier subconjunto de $m+1$ vértices define un m-simplex llamado cara del simplex original.

Un complejo simplicial K es una colección de simplex que satisface dos propiedades:

1. Inclusión de caras: Si un simplex σ pertenece a K, entonces todas sus caras también pertenecen a K.
2. Intersección: La intersección de dos simplex en K es vacía o es una cara común de ambos.

El complejo simplicial K puede representarse tanto geométricamente en $\mathbb{R}^d$ como combinatorialmente mediante la descripción de sus vértices y caras.

2.1.1.4. Construcción de complejos simpliciales a partir de nubes de puntos

Una de las principales técnicas en TDA es construir complejos simpliciales a partir de nubes de puntos. A continuación, se describen dos métodos comunes para esta construcción: el complejo Čech y el complejo Vietoris-Rips.

Complejo Čech

El *complejo Čech* se basa en la intersección de bolas centradas en los puntos de la nube. Para un conjunto de puntos $X = \{x_1, x_2, \ldots, x_n\}$ y un parámetro de escala $\alpha \geq 0$, el complejo Čech $\check{C}_\alpha(X)$ se define como:

$$\check{C}_\alpha(X) = \left\{ \{x_{i_0}, x_{i_1}, \ldots, x_{i_k}\} : \bigcap_{j=0}^{k} B(x_{i_j}, \alpha) \neq \emptyset \right\}, \tag{2.15}$$

donde $B(x, \alpha)$ denota la bola abierta de radio α centrada en x:

$$B(x, \alpha) = \{y \in M : \rho(x, y) < \alpha\}. \tag{2.16}$$

Complejo Vietoris-Rips

El *complejo Vietoris-Rips* se basa en la distancia entre pares de puntos. Para el mismo conjunto de puntos y parámetro de escala, el complejo Vietoris-Rips, $\text{Rips}_\alpha(X)$, se define como:

$$\text{Rips}_\alpha(X) = \left\{ \{x_{i_0}, x_{i_1}, \ldots, x_{i_k}\} : \rho(x_{i_j}, x_{i_l}) \leq 2\alpha, \forall j, l \right\}. \tag{2.17}$$

El complejo Vietoris-Rips es más fácil de calcular que el complejo Čech, ya que no requiere verificar intersecciones de bolas, sino solo la distancia entre pares de puntos.

2.1.1.5. Filtraciones

Una *filtración* es una familia de complejos simpliciales parametrizada por un valor de escala que refleja la estructura de los datos a diferentes niveles de resolución. Formalmente, una filtración $\{K_\alpha\}_{\alpha\in\mathbb{R}_+}$ es una familia de complejos simpliciales tal que para $\alpha \leq \beta$:

$$K_\alpha \subseteq K_\beta. \tag{2.18}$$

A medida que α aumenta, los complejos simpliciales K_α capturan más detalles de la estructura de los datos (ver Figura 2.3).

Ejemplos de filtraciones

1. *Filtración Čech:* La filtración Čech se construye incrementando el parámetro α y formando el complejo Čech correspondiente:

$$\{\check{C}_\alpha(X)\}_{\alpha\in\mathbb{R}_+}. \tag{2.19}$$

2. *Filtración Vietoris-Rips:* Similarmente, la filtración Vietoris-Rips se construye incrementando α y formando el complejo Vietoris-Rips correspondiente:

$$\{\text{Rips}_\alpha(X)\}_{\alpha\in\mathbb{R}_+}. \tag{2.20}$$

Distancia a la medida

La *distancia a la medida* es una técnica robusta para manejar datos ruidosos y atípicos. Dado un conjunto de datos X y una distribución de probabilidad (medida) P en M, la distancia a la medida $d_{P,m,r}$ con parámetros $m \in [0,1]$ y $r \geq 1$ se define como:

$$d_{P,m,r}(x) = \left(\frac{1}{m}\int_0^m \delta_{P,u}(x)^r\, du\right)^{1/r}, \tag{2.21}$$

donde $\delta_{P,u}(x)$ es la mínima distancia t tal que $P(B(x,t)) \geq u$. Esta métrica es más robusta frente a outliers comparada con las métricas tradicionales.

2.1.1.6. Homología persistente

La homología persistente es una herramienta fundamental en TDA que permite analizar la evolución de las características topológicas de un conjunto de datos a lo largo de una filtración de complejos simpliciales. La homología persistente rastrea la aparición y desaparición de características topológicas como componentes conexas, ciclos y cavidades a medida que varía un parámetro de escala. Este enfoque permite capturar la estructura topológica de los datos de manera robusta y a múltiples escalas (como se puede apreciar en la Figura 2.4).

Grupos de homología

Para un complejo simplicial K, los grupos de homología $H_k(K)$ se definen en términos de k-cadenas, k-ciclos y k-bordes.

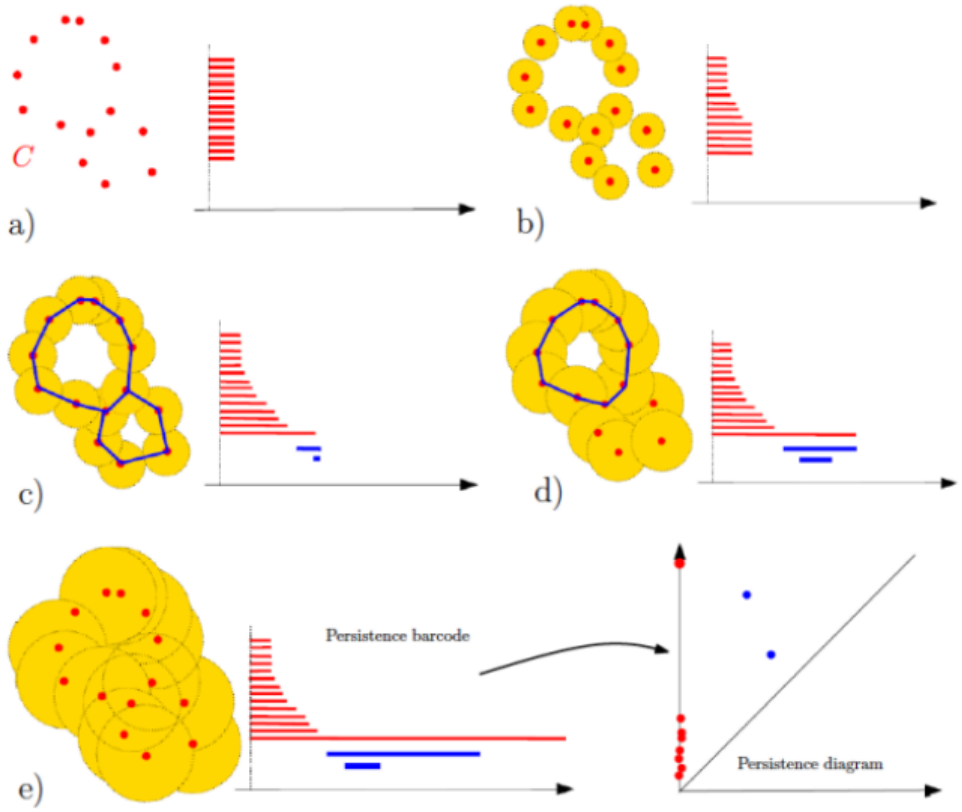

Figura 2.3: Filtración del conjunto de nivel inferior de la función de distancia a una nube de puntos y construcción de su diagrama de persistencia a medida que aumenta el radio de las esferas. Las curvas azules representan ciclos unidimensionales asociados con las barras azules en los códigos de barras. El diagrama de persistencia se define a partir de los códigos de barras de persistencia, mostrando la evolución de las características topológicas a través de los valores del radio. Fuente: [13].

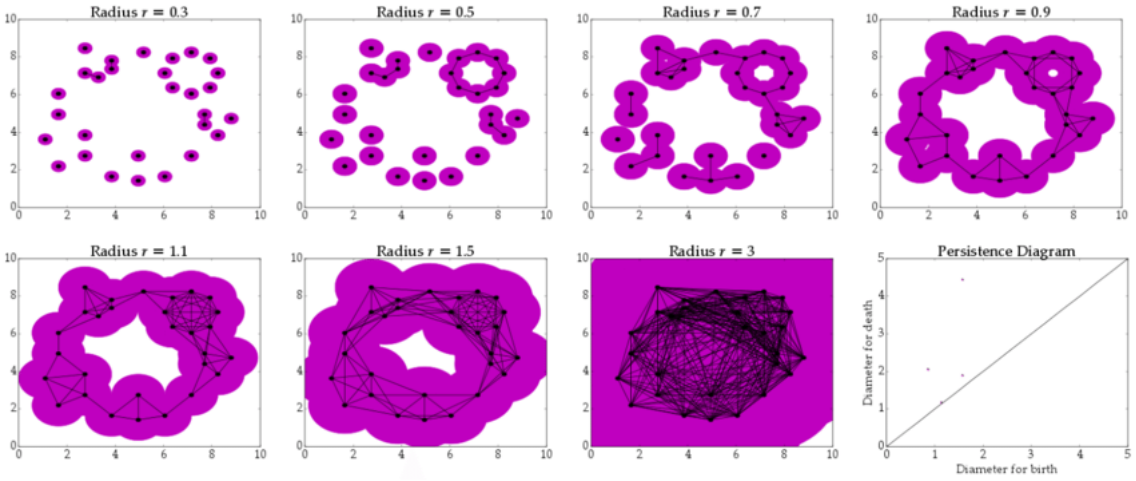

Figura 2.4: Ejemplo de uso de la homología persistente para investigar un conjunto de datos de nube de puntos mediante la construcción del complejo de Rips. Las aristas del complejo de Rips se dibujan en negro. El diagrama de persistencia (abajo a la derecha) resume la aparición y desaparición de ciclos en el espacio a medida que cambia el parámetro del complejo de Rips. Fuente: [14].

- k**-cadena:** Una k-cadena es una combinación lineal formal de k-simplex en K, denotada por $C_k(K)$.
- k**-ciclo:** Un k-ciclo es una k-cadena cuya frontera es cero, es decir, $\partial\sigma = 0$.
- k**-borde:** Un k-borde es la frontera de un $(k+1)$-simplex.

El k-ésimo grupo de homología $H_k(K)$ se define como el cociente del espacio de k-ciclos $Z_k(K)$ por el espacio de k-bordes $B_k(K)$:

$$H_k(K) = \frac{Z_k(K)}{B_k(K)}. \tag{2.22}$$

La homología persistente analiza cómo cambian los grupos de homología a lo largo de una filtración. Para una filtración $\{K_\alpha\}_{\alpha\in\mathbb{R}_+}$, se considera una secuencia de grupos de homología:

$$H_k(K_\alpha) \to H_k(K_\beta) \quad \text{para } \alpha \leq \beta. \tag{2.23}$$

Módulo de persistencia

Un *módulo de persistencia* es una secuencia de grupos de homología $H_k(K_\alpha)$ con mapas inducidos por la inclusión:

$$H_k(K_{\alpha_1}) \to H_k(K_{\alpha_2}) \to \cdots \to H_k(K_{\alpha_n}). \tag{2.24}$$

El rango de estos mapas proporciona información sobre la persistencia de las características topológicas.

2.2. Motivación para caracterizado de la propagación espacial del COVID-19 en España

La pandemia de COVID-19 ha resaltado la necesidad urgente de desarrollar herramientas analíticas capaces de comprender y mitigar la propagación de enfermedades infecciosas. La complejidad de la propagación del COVID-19, influenciada por una multitud de factores biológicos, sociales y geográficos, exige enfoques innovadores para su análisis. En este contexto, el Análisis Topológico de Datos surge como una metodología prometedora debido a su capacidad para manejar datos complejos y de alta dimensionalidad, y para revelar patrones ocultos en los datos.

Aplicar el TDA para caracterizar la propagación espacial del COVID-19 en España tiene varias motivaciones clave. En primer lugar, España ha sido uno de los países más afectados por la pandemia en Europa, experimentando múltiples olas de infección que han variado significativamente en términos de intensidad y distribución geográfica. Comprender cómo el virus se ha propagado espacialmente en diferentes momentos puede proporcionar conocimientos valiosos para la gestión de futuras pandemias y la implementación de medidas de control más efectivas.

En segundo lugar, la propagación de enfermedades infecciosas como el COVID-19 es un fenómeno inherentemente topológico. La forma en que los casos de infección se agrupan y dispersan en diferentes regiones geográficas puede influir en la efectividad de las intervenciones de salud pública. El TDA, con su capacidad para analizar la estructura y la forma de los datos, es particularmente adecuado para capturar estas dinámicas espaciales y temporales. Al utilizar el TDA, es posible identificar patrones recurrentes, puntos críticos de infección y trayectorias de propagación que pueden ser difíciles de detectar con métodos tradicionales.

Además, la metodología TDA puede proporcionar una visión integrada de múltiples factores que influyen en la propagación del COVID-19. Por ejemplo, al incluir dimensiones adicionales como la temporalidad (fecha de los casos) y datos demográficos, es posible construir una representación multidimensional de la pandemia. Esto permite un análisis más completo que puede considerar cómo las características topológicas cambian a lo largo del tiempo y en diferentes contextos sociales y económicos.

En suma, la aplicación del TDA en el análisis de la propagación del COVID-19 puede contribuir a la validación y mejora de modelos epidemiológicos existentes. Los descriptores topológicos obtenidos del TDA pueden ser utilizados para calibrar y ajustar modelos matemáticos, proporcionando una base más sólida para la predicción y el control de la enfermedad. Esta integración de enfoques topológicos y epidemiológicos tiene el potencial de mejorar significativamente nuestra capacidad para responder a pandemias actuales y futuras.

Capítulo 3

Metodología

3.1. Algoritmo Mapper

El algoritmo Mapper fue introducido por Singh, Memoli y Carlsson en 2007 [7] como una técnica para visualizar y analizar datos de alta dimensión. Mapper crea una representación gráfica de los datos que facilita la identificación de patrones, clusters y otras características topológicas que pueden no ser evidentes en la visualización directa de los datos en su espacio original.

El propósito de Mapper es proyectar datos complejos en un espacio más manejable, cubrir este espacio proyectado con conjuntos abiertos (bins), y luego usar la topología para analizar las preimágenes de estos conjuntos. El resultado es un grafo que retiene información sobre la conectividad y otras características topológicas del conjunto de datos original.

3.2. Pasos del algoritmo Mapper

La implementación del algoritmo consta de varios pasos:

1. *Proyección de los datos:* La proyección es una función $f : X \to \mathbb{R}^d$ que transforma el conjunto de datos X en un espacio de menor dimensión $\mathbb{R}^d$. Esta proyección puede ser una función de características significativas de los datos, como la densidad del núcleo, la distancia a una medida, o el laplaciano del grafo.

   Matemáticamente, si X es un conjunto de puntos en un espacio de alta dimensión, la proyección f podría ser:

   $$f(x) = \sum_{i=1}^{n} w_i \cdot x_i \tag{3.1}$$

   donde w_i son pesos y x_i son las coordenadas del punto x en X.

2. *Cobertura del espacio proyectado:* Se cubre la imagen de los datos proyectados $f(X)$ con un conjunto de cubos superpuestos $\{U_i\}$. Estos cubos están definidos en el espacio proyectado $\mathbb{R}^d$ y la cantidad de solapamiento entre ellos es un parámetro ajustable. Este paso crea una partición de los datos en conjuntos más pequeños y manejables.

Si consideramos $f(X)$ en $\mathbb{R}^2$, la cobertura podría ser:

$$\{U_i\} = \{\text{Cubos en } \mathbb{R}^2 \text{ con longitud de lado } l \text{ y solapamiento } \delta\}. \tag{3.2}$$

3. *Clustering en las preimágenes:* Para cada cubo U_i, se calcula la preimagen $f^{-1}(U_i)$, que contiene los puntos del conjunto de datos original que son mapeados en U_i. Se aplica un algoritmo de clustering a cada preimagen para agrupar los puntos de datos que son similares.

   Por ejemplo, usando DBSCAN para clustering:

   $$\text{DBSCAN}(\epsilon, \text{minPts}), \tag{3.3}$$

   donde ϵ es la distancia máxima entre puntos para considerarlos en el mismo cluster y minPts es el número mínimo de puntos para formar un cluster.

4. *Construcción del grafo Mapper:* Se crea un grafo donde los nodos representan los clusters $\{C_{ij}\}$. Se añade una arista entre dos nodos C_{ij} y C_{kl} si sus preimágenes tienen intersección no vacía $C_{ij} \cap C_{kl} \neq \emptyset$.

   Matemáticamente, el grafo $M(X, U, f)$ se define como:

   $$M(X, U, f) = \{(C_{ij}, C_{kl}) : C_{ij} \cap C_{kl} \neq \emptyset\}. \tag{3.4}$$

Como ejemplo ilustrativo, consideremos un conjunto de datos X en $\mathbb{R}^2$ con una proyección f que olvida la primera coordenada:

$$f(x, y) = y. \tag{3.5}$$

Si cubrimos $f(X)$ con intervalos U_i en $\mathbb{R}$ y aplicamos clustering en cada $f^{-1}(U_i)$, el grafo resultante mostrará cómo se agrupan los puntos en función de su segunda coordenada, revelando la estructura topológica subyacente del conjunto de datos original.

3.3. Parámetros de algoritmo Mapper e implementación

El algoritmo Mapper transforma datos de alta dimensión en un grafo que preserva muchas de sus características topológicas (ver en Figuras 3.1 y 3.2). Este proceso involucra varios pasos y parámetros clave que deben configurarse adecuadamente para obtener resultados significativos. A continuación, se presenta una explicación detallada de estos argumentos y cómo se configuran en la implementación del algoritmo Mapper.

Datos de entrada

El conjunto de datos a analizar debe estar estructurado en una matriz donde cada fila representa un punto de datos y cada columna una característica. Estos datos pueden ser de cualquier tipo, incluyendo valores numéricos, categóricos, temporales, espaciales, etc.

Formato: Matriz de *numpy* o dataframe en *pandas*.

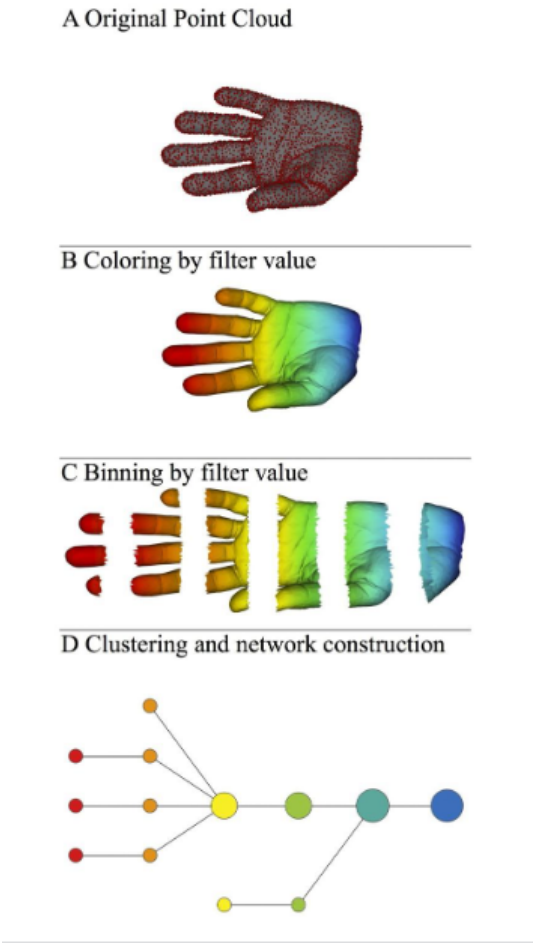

Figura 3.1: Proceso de análisis topológico aplicado a una nube de puntos de una mano: se colorea según un valor de filtro, se clasifica, y se agrupan y construyen redes para analizar la estructura. Fuente: [15].

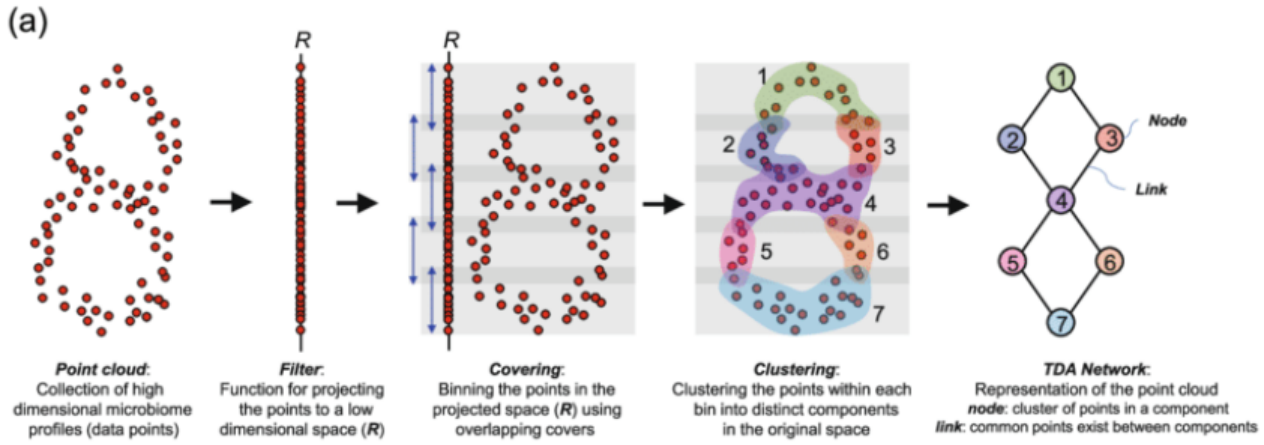

Figura 3.2: Proceso desde la nube de puntos de perfiles microbiológicos hasta la construcción de una red TDA, que incluye la proyección a un espacio de baja dimensión, la cobertura con cubiertas superpuestas, la agrupación en componentes distintos y la representación final como una red de nodos y enlaces. Fuente: [16].

Función de proyección

La función de proyección transforma los datos originales de alta dimensión en un espacio de dimensión más baja, facilitando su análisis. La proyección puede ser lineal o no lineal, y se selecciona en función de las características más relevantes del conjunto de datos y los objetivos del análisis.

Tipos de proyección

- *Proyección lineal:*
  Utiliza métodos como el Análisis de Componentes Principales (PCA) para reducir dimensionalidad manteniendo relaciones lineales.
- *Proyección de coordenadas:*
  Selecciona ciertas dimensiones de los datos originales para análisis, ignorando otras.
- *Proyección basada en características:*
  Utiliza características específicas de los datos, como la densidad del núcleo o la distancia a una medida.
- *Proyección no lineal:*
  Métodos como t-SNE o UMAP que capturan relaciones no lineales en los datos.

Cobertura (cover)

La cobertura del espacio proyectado implica dividir este espacio en un conjunto de intervalos (cubos) superpuestos. Esto facilita el análisis topológico de los datos, ya que cada cubo agrupa subconjuntos de los datos proyectados.

Parámetros

- `n_cubes`:
  Número de intervalos (cubos) en cada dimensión del espacio proyectado. Este parámetro determina la resolución de la cobertura.

- `perc_overlap`:
  Porcentaje de solapamiento entre los intervalos en cada dimensión. Un mayor solapamiento incrementa la conectividad entre clusters, mientras que un menor solapamiento puede generar más componentes disjuntos.

Clustering

El clustering agrupa los puntos de datos en cada preimagen de los conjuntos de la cobertura. Esto ayuda a identificar clusters locales en los datos proyectados y a construir la estructura del grafo Mapper.

Parámetros:

- **clusterer:** Algoritmo de clustering utilizado para agrupar los puntos de datos. Algunos de los algoritmos comunes son:

  - **DBSCAN** *(Density-Based Spatial Clustering of Applications with Noise):* Identifica clusters de cualquier forma y maneja outliers eficientemente. Parámetros clave incluyen:

    - `eps`: Distancia máxima entre puntos para considerarlos en el mismo cluster.

    - `min_samples`: Número mínimo de puntos para formar un cluster.

  - **K-means:** Agrupa los datos en k clusters mediante la minimización de la varianza dentro de cada cluster. Parámetros clave incluyen:

    - `n_clusters`: Número de clusters a formar.

  - **Hierarchical Clustering:** Construye una jerarquía de clusters mediante la fusión o división iterativa de clusters. Parámetros clave incluyen:

    - `n_clusters`: Número de clusters a formar.

    - `linkage`: Método para calcular la distancia entre clusters (por ejemplo, `ward`, `complete`, `average`).

3.3.1. Algoritmo de clustering: DBSCAN

El algoritmo DBSCAN *(Density-Based Spatial Clustering of Applications with Noise)* es un método de clustering basado en la densidad que identifica clusters de puntos densamente agrupados y puede manejar eficientemente el ruido (outliers). Fue propuesto por Ester et al. [17] y se ha convertido en una de las técnicas más populares para el clustering de datos espaciales y de alta dimensión.

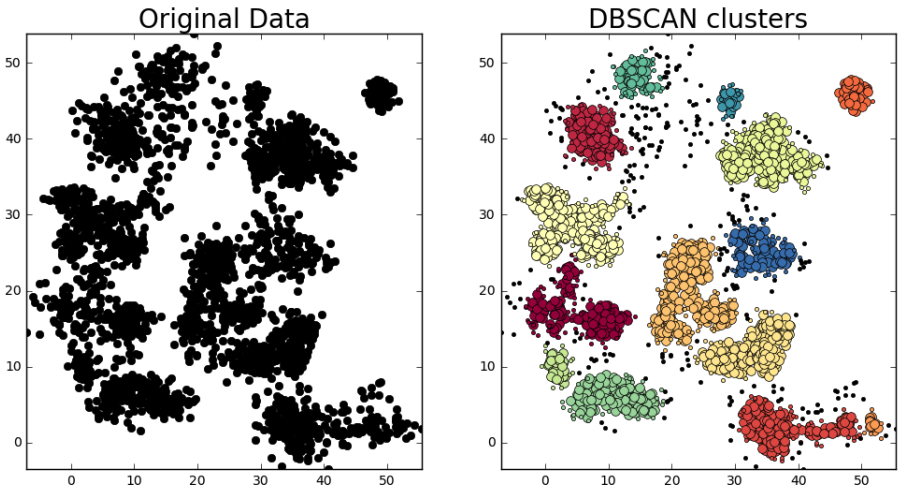

Figura 3.3: Visualización de datos originales frente a los clusters obtenidos con el algoritmo DBSCAN. Los datos originales (izquierda) se agrupan en varios clusters diferenciados por colores (derecha) tras la aplicación del algoritmo.

3.3.1.1. Principios básicos del DBSCAN

DBSCAN clasifica los puntos de datos en tres categorías: puntos del núcleo, puntos alcanzables directamente y puntos fronterizos. Los puntos del núcleo son aquellos que tienen al menos un número mínimo de vecinos dentro de un radio especificado. Los puntos alcanzables directamente son aquellos que están dentro del radio especificado de un punto del núcleo. Los puntos fronterizos están dentro del radio especificado de un punto del núcleo, pero no tienen suficientes vecinos para ser un punto del núcleo. Además, existen los outliers, que son puntos que no se clasifican ni como puntos del núcleo ni como puntos alcanzables directamente o fronterizos.

Parámetros clave

DBSCAN utiliza dos parámetros: epsilon (ϵ) y MinPts. Epsilon es el radio de búsqueda alrededor de cada punto, y dos puntos se consideran vecinos si la distancia entre ellos es menor o igual a ϵ. MinPts es el número mínimo de puntos requeridos para formar un cluster (incluyendo el punto en sí mismo). Un punto es considerado un punto del núcleo si tiene al menos MinPts puntos dentro de su radio ϵ.

Funcionamiento del algoritmo

El algoritmo DBSCAN comienza con un punto aleatorio que no haya sido visitado y obtiene el conjunto de puntos vecinos a este punto utilizando el radio ϵ. Si el número de vecinos es mayor o igual a MinPts, el punto se clasifica como un punto del núcleo y se inicia un nuevo cluster. Si el número de vecinos es menor a MinPts, el punto se clasifica como ruido temporalmente, ya que podría ser parte de un cluster si se encuentra en el radio de otro punto del núcleo más adelante. Para cada punto del núcleo, se agrega al cluster todos los puntos alcanzables directamente y se repite el proceso para los puntos vecinos, expandiendo el cluster hasta que no se puedan agregar más puntos. Este proceso (Figura 3.3) se repite para todos los puntos no visitados hasta que todos los puntos hayan sido procesados.

3.3.1.2. Ejemplo de uso de DBSCAN en análisis de datos

Imaginemos un conjunto de datos de coordenadas geográficas que representan ubicaciones de casos de Covid-19 y queremos identificar clusters de alta densidad de casos para estudiar la propagación del virus en ciertas regiones. Primero, se seleccionan los parámetros, donde ϵ puede ser una distancia geográfica como 0.01 grados (aproximadamente 1 km) y MinPts puede ser el número mínimo de al menos 5 casos cercanos para considerar un área como un brote significativo. Luego, se ejecuta el algoritmo DBSCAN sobre los datos de ubicaciones geográficas para identificar los clusters de alta densidad de casos. Los clusters identificados representan áreas donde hay una alta concentración de casos de Covid-19, y los puntos de ruido representan áreas aisladas que no forman parte de ningún cluster significativo.

3.4. Uso del algoritmo Mapper para estudiar la propagación del Covid-19

En el estudio de la propagación del Covid-19, el algoritmo Mapper se aplica a un conjunto de datos de múltiples dimensiones que incluyen información geográfica, temporal y datos epidemiológicos acumulados (casos y fallecidos) de Covid-19.

3.4.1. Fuentes de datos del Covid-19

Para este estudio, los datos de COVID-19 en España se obtuvieron de dos fuentes principales:

- *Instituto de Salud Carlos III (ISCIII):* El ISCIII [18] proporciona datos detallados sobre la evolución de la pandemia, incluyendo el número de casos confirmados, hospitalizaciones, ingresos en unidades de cuidados intensivos (UCI), recuperaciones y fallecimientos. Los datos se recopilan y actualizan diariamente, ofreciendo un seguimiento exhaustivo del virus en todo el país.
- *Ministerio de Sanidad:* El Ministerio de Sanidad [19] ofrece informes diarios y semanales sobre la situación de la pandemia, incluyendo datos agregados a nivel nacional y autonómico. Estos informes también incluyen información sobre la capacidad hospitalaria, la distribución de vacunas y las medidas de control implementadas en diferentes momentos.

Los datos recopilados abarcan varios aspectos clave de la pandemia, permitiendo un análisis detallado y multifacético. A continuación, se describe la estructura y las dimensiones de los datos utilizados:

Campos utilizados:

- **Coordenadas geográficas:** Longitud y latitud en formato decimal (más apropiado para introducirlo como argumento del algoritmo Mapper) de las Comunidades Autónomas de España.
- **Tiempo:** Número de días desde una fecha de referencia (datos disponibles a partir del 1 de enero de 2020). Datos hasta el 13 de junio de 2022.
- **Datos epidemiológicos del Covid-19:** número de casos y fallecidos diarios acumulado en cada Comunidad Autónoma en un día específico.

Cada punto de datos p se representa como un vector en $\mathbb{R}^4$:

$$p = (x, y, z, w),$$

donde x e y son las coordenadas geográficas, z es el número de casos o fallecidos, y w es la fecha (diferencia de días con respecto a una fecha de referencia).

Por ejemplo, Cantabria se localiza aproximadamente en $43°N$, $4°W$ y reportó 2277 casos acumulados a fecha de 18 de mayo de 2020. Por tanto, ese punto de datos se representa en la nube de puntos de la siguiente forma:

$$(43.1595664, -4.0878382, 2277, 138).$$

La nube de puntos registra los casos diarios desde el 1 de enero del 2020 hasta el 13 de junio de 2022 (diferencia de 895 días) para cada Comunidad Autónoma, incluyendo las Ciudades Autónomas de Ceuta y Melilla. Por tanto, la nube de puntos consta de:

$$\text{(número de regiones)} \cdot \text{(número de días)} = 895 \cdot 19 = 17005.$$

Por lo tanto, se dispone de 17005 vectores pertenecientes a $\mathbb{R}^4$.

3.4.2. Procesamiento de los datos

3.4.2.1. Limpieza de datos

- *Eliminación de datos faltantes:* Se han tratado las entradas con datos faltantes para evitar sesgos y asegurar la integridad del análisis. En los casos donde faltaban datos para días específicos, se utilizó la interpolación lineal para estimar los valores faltantes. Este método se aplicó especialmente cuando los datos faltantes representaban menos del 2% del total de observaciones en una serie temporal, permitiendo una estimación continua y consistente basada en los valores adyacentes.
- *Corrección de errores:* Se han corregido posibles errores en los datos, como fechas incorrectas y valores fuera de rango.
- *Adecuación de los datos:* Se han convertido los datos originales a la estructura estándar esperada por el algoritmo Mapper $(latitud, longitud, casos, dia)$. Cabe indicar que las fechas se convirtieron a una escala numérica, donde el primer día de registro (1 de enero de 2020) se asigna a un valor de 0, y los días sucesivos se incrementan en consecuencia (hasta el último registro recogido para el día 13 de junio de 2022). Asimismo, los datos de partida para los casos y fallecidos por Covid-19, publicados diariamente, estaban desagregados. Se ha realizado una labor de hacer estos casos acumulativos.
- *Nivel de granularidad:* Dada la disponibilidad de datos ofrecida por las dos fuentes de datos utilizadas, se han agregado los datos a niveles autonómicos para permitir un análisis granularizado y facilitar la comparación entre diferentes CCAA.

3.4.2.2. Agregación de los datos

Los datos utilizados contemplan que el número de casos y fallecidos sea acumulado, y no desagregado. Esto se fundamenta en varias consideraciones prácticas y metodológicas que optimizan la representación y análisis de la propagación del Covid-19.

Primero, los datos acumulados permiten una visualización continua y coherente de la evolución de la pandemia. Al utilizar casos acumulados, se puede seguir el desarrollo y la progresión de los contagios en el tiempo sin perder información sobre los eventos pasados. Cada punto de datos refleja el estado actual y la historia de infecciones en un lugar específico, proporcionando una perspectiva completa del crecimiento de los casos. Esto es crucial para identificar tendencias y patrones a lo largo del tiempo, lo que sería más difícil si los datos estuvieran desagregados por días o semanas, ya que implicaría un análisis más segmentado y posiblemente inconsistente.

Además, los casos acumulados ayudan a suavizar las fluctuaciones diarias que pueden resultar de retrasos en la notificación, variaciones en la capacidad de prueba, y otros factores operativos. Las cifras diarias pueden presentar picos y caídas abruptas que podrían complicar la identificación de patrones a largo plazo. Al acumular los datos, se mitigan estas variaciones y se obtiene una curva más representativa del progreso de la pandemia, facilitando un análisis más robusto y preciso.

Otra razón es la simplicidad y claridad en la interpretación de los datos. Los datos acumulados son intuitivamente más fáciles de entender para los investigadores y el público en general, ya que siempre representan el total de casos hasta una fecha específica. Esto contrasta con los datos desagregados, que requerirían una suma acumulativa adicional para obtener una visión completa del impacto de la pandemia.

Consideremos un ejemplo simplificado con datos desagregados y acumulados:

Datos desagregados (diarios):

Día	Casos en CCAA A	Casos en CCAA B
1	5	2
2	3	1
3	4	3

Datos acumulados:

Día	Casos en CCAA A	Casos en CCAA B
1	5	2
2	8	3
3	12	6

En los datos desagregados, la variabilidad diaria puede dificultar la identificación de patrones claros. Sin embargo, en los datos acumulados, es evidente que la Comunidad Autónoma A está experimentando un crecimiento más rápido en el número de casos en comparación con la Comunidad Autónoma B.

3.4.2.3. Consistencia de los datos entre diferentes fuentes

En este proyecto, uno de los pasos cruciales en el procesamiento de datos ha sido la comprobación de la consistencia entre diferentes fuentes de datos que reportan

información sobre COVID-19 en España. Concretamente, se ha realizado una comparación entre los datos de las dos fuentes consultadas: Ministerio de Sanidad e Instituto de Salud Carlos III (ISCIII), véase en la Figura 3.4. Este proceso es esencial para asegurar que los análisis posteriores se basen en información coherente y fiable.

```
# Cargar los datos desde un archivo CSV
data = pd.read_csv('ccaa_covid19_datos_sanidad_nueva_serie - ccaa_covid19_datos_sanidad_nueva_serie_SIMPLIFICADO.csv')

# Filtrar las columnas de interés (casos y fallecidos diarios por Comunidad Autónoma)
data = data[['Fecha', 'CCAA', 'Casos', 'Fallecidos']]
data
```

	Fecha	CCAA	Casos	Fallecidos
0	2020-01-01	No consta	0	0
1	2020-01-01	Andalucía	0	0
2	2020-01-01	Aragón	0	0
3	2020-01-01	Asturias	0	0
4	2020-01-01	Baleares	0	0
...	...	...	...	...
17895	2022-06-13	Navarra	0	0
17896	2022-06-13	País Vasco	24	0
17897	2022-06-13	La Rioja	0	0
17898	2022-06-13	Ceuta	0	0
17899	2022-06-13	Melilla	0	0

17900 rows × 4 columns

Figura 3.4: Carga de datos procesados para algoritmo Mapper.

Para llevar a cabo esta tarea, se seleccionaron y limpiaron los datos de alcance autonómico. Se transformaron las fechas y se calcularon datos acumulados diarios, asegurando que los datos estuvieran en un formato adecuado para el análisis. Estas tareas de manipulación y análisis de datos se realizaron en un entorno *Jupyter Notebook* utilizando la librería *Pandas*. Esta elección facilitó el manejo interactivo de los datos y la ejecución de diversas operaciones de limpieza, fusión y análisis de manera eficiente.

Los datasets del Ministerio de Sanidad y del ISCIII se fusionaron basándose en dos criterios: la fecha y la Comunidad Autónoma (véase Figura 3.5). Esta fusión permitió alinear los datos de ambas fuentes y realizar comparaciones directas. Se calculó la diferencia diaria en el número de casos reportados entre ambas fuentes. Para cada día y cada Comunidad Autónoma, se computó la diferencia absoluta entre el número de casos reportados por el Ministerio de Sanidad y el ISCIII. Este paso reveló las discrepancias entre las dos fuentes de datos, lo cual es importante para entender la variabilidad y posibles fuentes de error en los reportes de datos durante la pandemia.

Finalmente, se extrajo la diferencia media diaria para evaluar cuán similares eran los números de casos reportados por cada institución. Para el rango de fechas desde enero de 2020 hasta marzo de 2022, se encontró que la diferencia media diaria era de aproximadamente 151 casos. Este valor cuantifica la divergencia promedio entre las dos fuentes y subraya la necesidad de una revisión crítica de los datos reportados

```
merged_df = pd.merge(df_spain_iscii, df_spain_ministerio_sanidad, on=['Fecha', 'CCAA'], how='outer')
merged_df
```

	Fecha	cod_ine_x	CCAA	Casos_x	num_casos_prueba_pcr	num_casos_prueba_test_ac	num_casos_prueba_ag	num_casos_prueba_elisa	num_casos_prueba_desconocida	cod_ine_y	Casos_y	Fallecidc
0	2020-01-01	1.0	Andalucía	0.0	0.0	0.0	0.0	0.0	0.0	1	0	
1	2020-01-02	1.0	Andalucía	0.0	0.0	0.0	0.0	0.0	0.0	1	0	
2	2020-01-03	1.0	Andalucía	0.0	0.0	0.0	0.0	0.0	0.0	1	0	
3	2020-01-04	1.0	Andalucía	0.0	0.0	0.0	0.0	0.0	0.0	1	0	
4	2020-01-05	1.0	Andalucía	0.0	0.0	0.0	0.0	0.0	0.0	1	0	
...	...	...	...	...	...	...	...	...	...	...	...	
16375	2022-03-25	NaN	No consta	NaN	NaN	NaN	NaN	NaN	NaN	0	24	
16376	2022-03-26	NaN	No consta	NaN	NaN	NaN	NaN	NaN	NaN	0	19	
16377	2022-03-27	NaN	No consta	NaN	NaN	NaN	NaN	NaN	NaN	0	9	
16378	2022-	NaN	No	NaN	NaN	NaN	NaN	NaN	NaN	0	26	

Figura 3.5: Fusión entre datasets del ISCIII y Ministerio de Sanidad en base a la fecha y CCAA.

durante la pandemia.

El análisis mostró que, aunque los datos del Ministerio de Sanidad y del ISCIII están generalmente alineados, existe una diferencia promedio diaria de 151 casos (véase Figura 3.6) en el periodo estudiado. Este hallazgo es significativo porque refleja las posibles variaciones en la metodología de recolección y reporte de datos entre las instituciones, algo que fue común durante la pandemia.

```
merged_df['Diferencia'] = abs(merged_df['Casos_y'] - merged_df['Casos_x'])
merged_df['Diferencia']
```

```
0        0.0
1        0.0
2        0.0
3        0.0
4        0.0
        ...
16375    NaN
16376    NaN
16377    NaN
16378    NaN
16379    NaN
Name: Diferencia, Length: 16380, dtype: float64
```

```
# Calcular diferencia de número de casos entre datasets del ISCIII y Ministerio de Sanidad
average_error_by_day = merged_df['Diferencia'].mean()
average_error_by_day
```

```
150.77938435833173
```

Figura 3.6: Diferencia media de casos entre los datasets del ISCIII y Ministerio de Sanidad una vez fusionados. Rango de fechas comunes entre enero de 2020 hasta marzo de 2022.

Implementación en el algoritmo Mapper

Al aplicar el algoritmo Mapper a los datos acumulados, el proceso de proyección, cobertura y clustering se realiza sobre una serie temporal acumulativa, facilitando la construcción de un grafo que refleje fielmente la progresión de la pandemia. Esto es esencial para identificar patrones de crecimiento y declive, así como para visualizar la conectividad entre diferentes regiones en función de su evolución epidemiológica.

3.4.2.4. Normalización de los datos

Debido a las diferentes magnitudes de los datos en cada dimensión, es necesario normalizar cada coordenada para evitar que ciertas características dominen el análisis. La normalización se realiza escalando cada coordenada a una norma unitaria por columna, garantizando que cada dimensión tenga una contribución equilibrada en el clustering y la construcción del grafo.

3.4.2.5. Proyección y cobertura

La proyección es el primer paso en la implementación del algoritmo Mapper y consiste en mapear los datos de alta dimensión a un espacio de dimensión más baja, generalmente $\mathbb{R}^d$, para facilitar su análisis y visualización. La elección de la función de proyección depende del objetivo del análisis y las características más relevantes de los datos.

Debido a que los datos originales no son de dimensiones extremadamente altas y se quiere preservar la mayor cantidad de información posible, se utilizó la función de identidad como proyección. Esto significa que los datos no se redujeron en dimensionalidad antes de ser procesados por el algoritmo Mapper, permitiendo que todas los campos originales (latitud, longitud, casos confirmados, fechas, etc.) se utilicen en el análisis.

Por ende, la proyección utilizada es la identidad $f : X \to \mathbb{R}^4$. Los valores de cada una de las cuatro variables en cada eje de $\mathbb{R}^4$ se cubren con N ($N = 8$, $N = 10$, etc) intervalos solapados y, por tanto, cada cubo en el espacio original se forma como el producto cartesiano de esos cuatro intervalos. Es en este cubo donde se aplica el algoritmo de clustering que conduce a la construcción del gráfico final. El grado de solapamiento de esos N intervalos es de un $10\,\%$ o $19\,\%$.

Capítulo 4

Resultados

En esta sección, se presentan los resultados obtenidos a partir del análisis de los datos de COVID-19 en España utilizando el algoritmo Mapper. El objetivo principal es proporcionar una visualización detallada de la evolución temporal y espacial de la pandemia a través de un grafo que encapsula la información relevante sobre la propagación del virus en diferentes Comunidades Autónomas.

4.1. Aplicación de algoritmo Mapper

4.1.1. Grafo de datos

El grafo generado a partir del análisis del algoritmo Mapper se compone de nodos y aristas que representan agrupaciones de datos similares y sus relaciones, respectivamente. Los datos utilizados incluyen la ubicación geográfica (latitud y longitud de CCAA), el tiempo (días transcurridos desde el inicio de la pandemia), y el número de casos acumulados de Covid-19.

- Nodos: Cada nodo en el grafo representa un conjunto de datos de diferentes CCAA que comparten características similares en términos de epidemiología (número de casos o fallecidos) y temporalidad.
- Aristas: Las aristas conectan nodos que tienen intersecciones significativas en los datos, indicando similitudes o relaciones entre ellos.

4.1.2. Descriptores topológicos

Los descriptores topológicos proporcionan información sobre la estructura y características del grafo. Los principales descriptores utilizados en este análisis incluyen:

- **Componentes conexas:** Estas representan conjuntos de nodos que están interconectados, indicando regiones o periodos de tiempo donde los casos de COVID-19 muestran similitudes significativas.
- **Ramas:** Las ramas que emergen del grafo principal representan focos de brotes específicos en ciertas CCAA, mostrando cómo la pandemia se expande y afecta distintas regiones.

- **Tamaño de nodos:** El tamaño de cada nodo está relacionado con el número de datos (miembros) que contiene, lo que a su vez refleja la densidad de casos reportados en las CCAA correspondientes.

4.2. Análisis e interpretación de los resultados en contexto de propagación del Covid-19

4.2.1. Componentes conexas

Los componentes conexas del grafo revelan agrupaciones de datos que muestran similitudes significativas en términos de número de casos y temporalidad. Estas agrupaciones permiten identificar cómo la pandemia afectó a diferentes regiones en momentos específicos.

- *Componentes principales:* La mayoría de los nodos se agrupan en un componente principal o tronco, lo que refleja la tendencia general de la pandemia en España. Este componente muestra un crecimiento constante de casos, con picos asociados a oleadas específicas de contagio.
- *Componentes aislados:* Algunos nodos forman componentes aislados, que representan brotes localizados en ciertas Comunidades Autónomas. Estos brotes están asociados a eventos específicos o políticas regionales que afectaron a la propagación del virus.

4.2.2. Temporalidad y dinámica de la pandemia

La evolución temporal de los nodos en el grafo proporciona una visión clara de cómo la pandemia se ha desarrollado a lo largo del tiempo en diferentes CCAA.

- *Primeras fases:* Los nodos correspondientes a los primeros días de la pandemia están concentrados en un segmento del grafo, mostrando una menor cantidad de casos y una propagación más lenta.
- *Oleadas subsecuentes:* A medida que avanza el tiempo, se observa un aumento en el número de casos y una mayor dispersión de los nodos. Esto refleja las oleadas sucesivas de contagio y la expansión geográfica del virus.
- *Estabilización y rebrotes:* En las fases más recientes, se observa una estabilización en algunos nodos, mientras que otros muestran signos de rebrotes, lo que indica la necesidad de monitoreo continuo y adaptación de las políticas de salud pública.

4.2.3. Influencia de políticas autonómicas

Las diferencias en las políticas autonómicas, como confinamientos, restricciones de movilidad y campañas de vacunación, han tenido un impacto significativo en la propagación del COVID-19. El análisis del grafo permite correlacionar estas políticas con cambios en la estructura del grafo.

- *Confinamientos localizados:* En regiones donde se implementaron confinamientos estrictos, se observa una disminución en el tamaño y número de nodos, reflejando una reducción en los casos.

- *Efectos de la vacunación:* En las fases más recientes, las campañas de vacunación masiva han contribuido a una estabilización en la propagación del virus, lo cual se refleja en una menor expansión de nuevos nodos y ramas.

4.3. Obtención de resultados con Mapper

Para la obtención de resultados con el algoritmo Mapper, se emplea la librería *KeplerMapper* [20], bajo parámetros de la cobertura como el número de intervalos (cubos) N y el porcentaje de solapamiento entre intervalos.

4.3.1. Resultados con N = 10 cubos y perc overlap = 19% para el número de casos

En la Figura 4.1 se muestra el grafo Mapper, proporcionando una visualización detallada de la propagación de casos de Covid-19 en España desde el 1 de enero hasta el 13 de junio de 2022. De la imagen, pueden interpretarse varios troncos principales y componentes aislados.

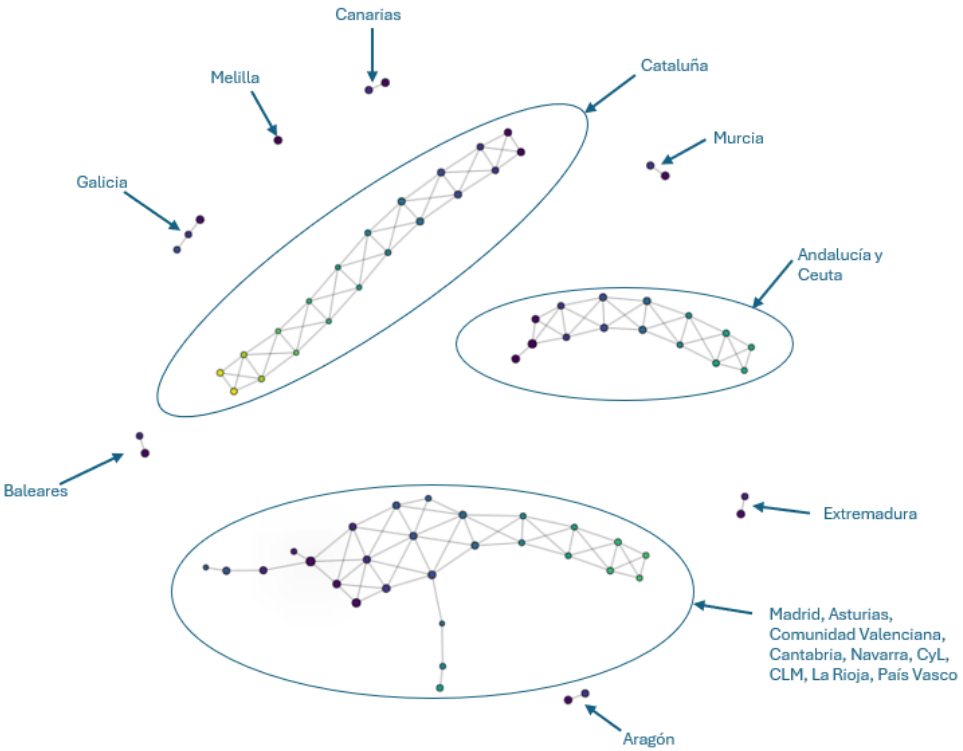

Figura 4.1: Grafo obtenido para el número de casos de Covid-19 reportados mediante algoritmo Mapper con parámetros N = 10, perc_overlap = 19% para todas las Comunidades Autónomas. Periodo: 01/01/2020-13/06/2022.

Troncos principales

El tronco principal de Cataluña es uno de los más densos y alargados en el gráfico. Representa la evolución de los casos de COVID-19 en la región de Cataluña, una

de las más afectadas. La densidad y longitud del tronco indican una alta tasa de infección y una propagación sostenida a lo largo del tiempo. Las múltiples conexiones dentro del tronco sugieren la existencia de varios picos de infección y su dispersión en diferentes áreas de Cataluña.

El tronco que incluye a Madrid, Asturias, Comunidad Valenciana, Cantabria, Navarra, Castilla y León (CyL), Castilla-La Mancha (CLM), La Rioja y País Vasco se presenta como uno de los troncos más prominentes y densamente conectados en el gráfico Mapper. Este clúster refleja una alta interconectividad y similitud en la evolución de los casos de COVID-19 en estas regiones. Madrid, siendo el nodo central y una de las ciudades más grandes y conectadas de España, actúa como un punto focal desde el cual se han dispersado los casos hacia las otras regiones del tronco.

La densa conectividad dentro de este tronco sugiere que estas regiones han experimentado una propagación de casos similar, posiblemente debido a factores como la alta movilidad de personas entre estas áreas y la interdependencia económica y social. La Comunidad de Madrid, al ser un centro neurálgico, ha influido significativamente en la dinámica de la pandemia, afectando a las regiones circundantes.

Asturias, Comunidad Valenciana, Cantabria, Navarra, CyL, CLM, La Rioja y País Vasco, aunque tienen sus propias particularidades, muestran patrones de contagio relacionados, lo que se visualiza en las múltiples conexiones dentro del tronco principal del cluster (véase Figura 4.2). Esta estructura sugiere que las oleadas de infección en una región han tenido repercusiones en las demás, creando un patrón de propagación interregional. Las políticas de confinamiento, las restricciones de movilidad y las estrategias de vacunación también han jugado un papel crucial en la evolución de la pandemia dentro de este tronco.

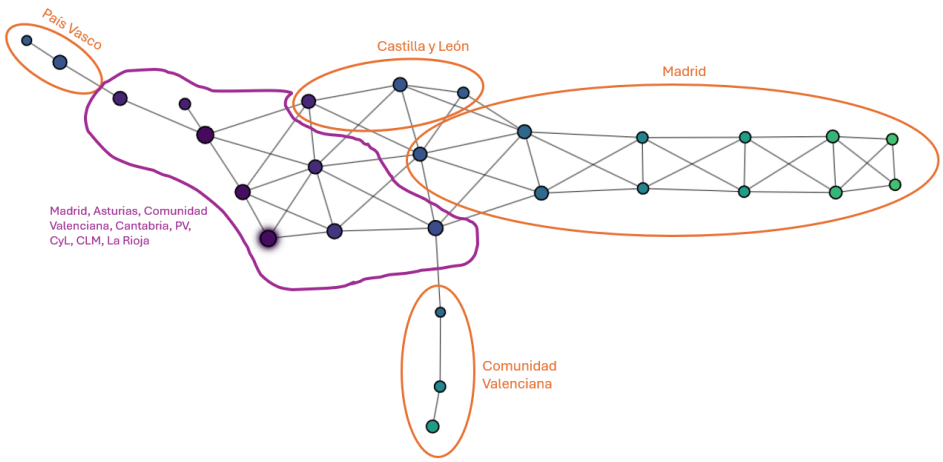

Figura 4.2: Detalle sobre el tronco principal que involucra a varias CCAA a partir de la Figura 4.1.

Ceuta y Andalucía aparecen en el gráfico Mapper como componentes interesantes debido a su particular evolución en la pandemia. Ceuta, una ciudad autónoma con características geográficas únicas, se presenta como un componente aislado. Esta separación se debe a su ubicación en el norte de África y sus conexiones limitadas con la península ibérica. La propagación de casos de COVID-19 en Ceuta ha seguido un

patrón distinto, influenciado por su aislamiento geográfico y las medidas de control específicas implementadas en la ciudad. Este componente aislado en el gráfico refleja cómo la ubicación y las políticas locales han influido en la evolución de la pandemia en esta región.

Por otro lado, Andalucía se muestra como una región con un tronco más conectado, pero con algunas ramificaciones que indican variaciones en la propagación de los casos dentro de la comunidad autónoma. Andalucía, siendo una de las regiones más extensas y pobladas de España, ha experimentado diferentes picos de infección a lo largo del tiempo. La estructura del gráfico para Andalucía refleja estas fluctuaciones, mostrando un tronco principal con varias ramas que representan subregiones dentro de la comunidad. Estas ramas indican cómo la pandemia ha afectado de manera diversa a distintas áreas dentro de Andalucía, con algunas zonas experimentando brotes más intensos y otras manteniendo tasas de infección más bajas.

Componentes aislados

Las Islas Canarias y Baleares aparecen como clústeres aislados en el gráfico. La separación geográfica natural de estas islas del resto de España se refleja en el gráfico Mapper. La propagación de casos en estas islas ha seguido un patrón distinto, posiblemente influenciado por factores locales como el turismo y las medidas de confinamiento específicas de las islas.

Melilla también aparece como componentes aislados. Similar a las islas, esta Ciudad Autónoma tiene dinámicas de propagación del virus únicas debido a su ubicación geográfica y conexiones limitadas con la península. Esto justifica su representación como componente separado en el gráfico Mapper. A diferencia de Melilla, Ceuta está conectada al cluster de Andalucía debido a su proximidad geográfica.

Extremadura y Aragón se presentan como componentes relativamente aislados con menor conectividad en comparación con otras regiones. Estas regiones pueden haber tenido menores tasas de infección o diferentes patrones de propagación del virus, lo que las hace menos conectadas con los troncos principales. Esto podría deberse a factores como la densidad de población, la movilidad interna y las políticas regionales de salud.

Análisis temporal y geográfico

El gráfico Mapper refleja la evolución de la pandemia a lo largo de más de dos años. Los troncos más largos y densos indican regiones con una propagación sostenida y múltiples picos de casos a lo largo del tiempo. La estructura del gráfico también muestra cómo la geografía influye en la propagación del virus. Las regiones geográficamente aisladas o menos conectadas tienden a aparecer como componentes separados.

4.3.2. Resultados con N = 10 cubos y perc overlap = 19% para el número de fallecidos

Se puede ver en la Figura 4.3 una visualización, que ofrece una comprensión profunda de la evolución temporal y geográfica de los fallecimientos en diferentes CCAA de España.

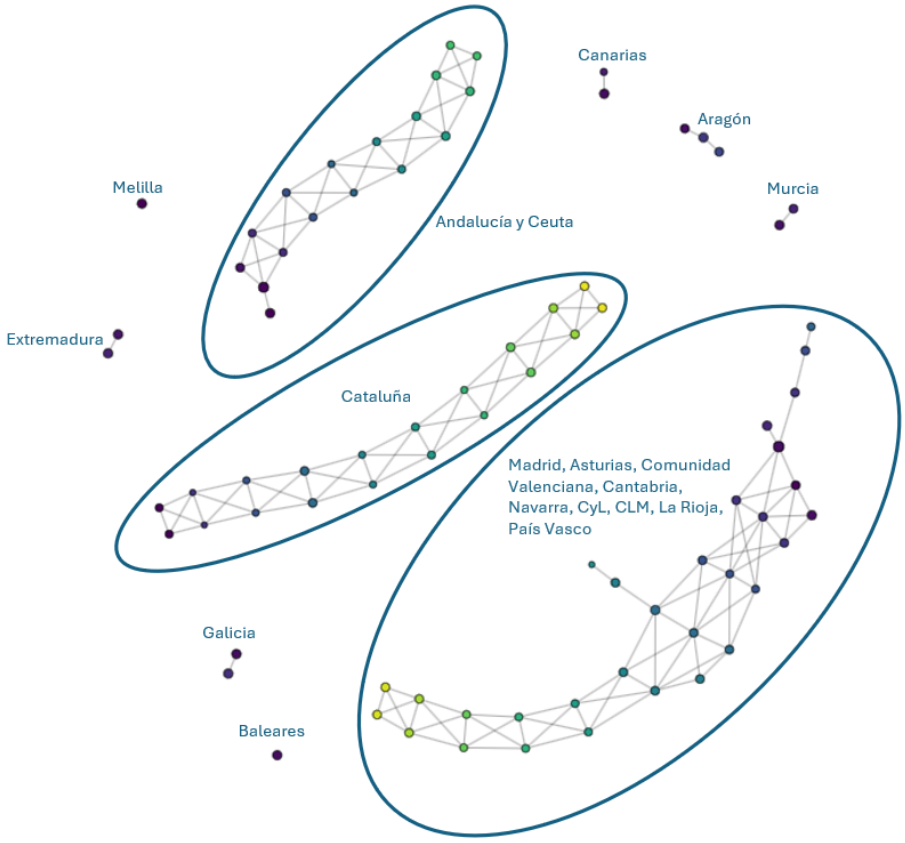

Figura 4.3: Grafo obtenido para el número de fallecidos de Covid-19 reportados mediante algoritmo Mapper con parámetros N = 10, perc_overlap = 19 % para todas las Comunidades Autónomas. Periodo: 01/01/2020-13/06/2022.

Troncos principales y clústeres

El gráfico muestra varios troncos principales que representan las regiones con mayor conectividad y casos de fallecidos. Uno de los troncos más destacados corresponde a la región de Cataluña. Este tronco alargado y denso indica una alta tasa de fallecimientos y una propagación sostenida del virus en el tiempo. Las múltiples conexiones dentro de este tronco sugieren varios picos de mortalidad, reflejando la gravedad de la situación en esta Comunidad Autónoma.

Otro tronco prominente incluye a Madrid, Asturias, Comunidad Valenciana, Cantabria, Navarra, Castilla y León , Castilla-La Mancha, La Rioja y País Vasco. Este clúster muestra una alta interconectividad y similitud en la evolución de los fallecimientos por Covid-19 en estas regiones. Madrid, actuando como un nodo central, influye significativamente en la dinámica de la pandemia, dispersando los casos hacia las regiones circundantes. Esto se refleja en la estructura del gráfico, donde las regiones mencionadas están estrechamente conectadas.

Componentes aislados

Las Islas Canarias y Baleares aparecen como componentes aislados, lo que subraya su separación geográfica del resto de España. La propagación de fallecimientos en

estas islas sigue un patrón distinto, influenciado por factores locales como el turismo y las medidas de confinamiento específicas. Del mismo modo, Melilla se presenta como nodos aislados, reflejando su ubicación geográfica en el norte de África y sus dinámicas únicas de contagio y mortalidad.

Extremadura y Aragón también se muestran como componentes relativamente aislados con menor conectividad. Esto podría deberse a menores tasas de mortalidad o diferentes patrones de propagación del virus, influenciados por factores como la densidad de población, la movilidad interna y las políticas regionales de salud.

Análisis temporal y geográfico

Desde una perspectiva temporal, el grafo Mapper captura la progresión de la pandemia y sus efectos mortales a lo largo de más de dos años. Los troncos más largos y densos en el gráfico representan regiones con una alta tasa de fallecimientos y una propagación sostenida del virus durante periodos prolongados. Por ejemplo, el tronco correspondiente a Cataluña muestra varias conexiones, lo que sugiere múltiples picos de mortalidad en distintos momentos. Esto indica que Cataluña ha experimentado varios períodos críticos donde los fallecimientos por COVID-19 aumentaron significativamente.

La estructura del grafo permite observar cómo las olas de mortalidad se han propagado temporalmente en diferentes regiones. Las ramificaciones en los troncos principales indican episodios de aumento en los fallecimientos, seguidos por periodos de relativa calma. Esta ciclicidad es una característica común en la evolución de la pandemia, donde se observan picos de mortalidad correlacionados con olas de contagios.

El análisis del gráfico Mapper revela que la evolución de la mortalidad por COVID-19 en España ha estado fuertemente influenciada por factores tanto regionales como temporales. Las diferencias en la densidad de población, las políticas de salud pública, las restricciones de movilidad y otros factores socioeconómicos han contribuido a la variabilidad en los patrones de fallecimientos observados en el gráfico.

4.3.3. Análisis de estructura del grafo y redes

En el ámbito de la construcción del grafo Mapper, una funcionalidad crucial es la capacidad de identificar y resumir las características geográficas y temporales de cada nodo. Esto se logra mediante procesos que recopilan las regiones específicas vinculadas a los miembros de cada nodo, asegurando que se reconozcan todas las áreas geográficas únicas involucradas. Esta recopilación es esencial para entender cómo se distribuyen geográficamente las interacciones o eventos asociados a la red.

Paralelamente, se analiza la dimensión temporal de los eventos relacionados con cada nodo. Se determinan las fechas más tempranas y tardías en las que se registraron actividades para cada conjunto de miembros. Esta información se utiliza para establecer un marco temporal que indica el período de actividad de cada nodo.

Al finalizar estos análisis, se generan resúmenes que proporcionan una vista clara de las regiones y los intervalos temporales para cada nodo (ver Figura 4.4). Estos resúmenes son herramientas valiosas para visualizar y comprender la extensión y duración de las actividades en la red, facilitando decisiones informadas en contextos que van desde la planificación urbana y la gestión de recursos hasta estudios

sociológicos y respuestas a emergencias.

```
Created 141 edges and 76 nodes in 0:00:00.689717.
Wrote visualization to: mapper_visualization_projection=[0, 1]_n_cubes=10_perc_overlap=0.19_clusterer=DBSCAN_scaler=MinMaxScaler_2024_07_23_17_50.html
El análisis topológico y la clusterización se han completado. Se ha generado el archivo 'mapper_visualization_projection=[0, 1]_n_cubes=10_perc_overla
Nodo cube0_cluster0:
  CCAA: Canarias
  Fechas: 2020-01-01 00:00:00 a 2022-02-13 00:00:00
Nodo cube1_cluster0:
  CCAA: Ceuta
  Fechas: 2020-01-01 00:00:00 a 2022-06-13 00:00:00
Nodo cube1_cluster1:
  CCAA: Melilla
  Fechas: 2020-01-01 00:00:00 a 2022-06-13 00:00:00
Nodo cube2_cluster0:
  CCAA: Andalucía, Ceuta
  Fechas: 2020-01-01 00:00:00 a 2022-06-13 00:00:00
Nodo cube3_cluster0:
  CCAA: Andalucía
  Fechas: 2020-01-01 00:00:00 a 2021-01-07 00:00:00
Nodo cube3_cluster1:
  CCAA: Murcia
  Fechas: 2020-01-01 00:00:00 a 2022-01-17 00:00:00
Nodo cube4_cluster0:
  CCAA: Baleares
  Fechas: 2020-01-01 00:00:00 a 2022-04-07 00:00:00
Nodo cube4_cluster1:
  CCAA: Castilla La Mancha, C. Valenciana, Madrid
```

Figura 4.4: Extracto de resultados de un análisis topológico utilizando la técnica de Mapper. Se detalla la creación de nodos y aristas en un grafo, con especificaciones sobre la cantidad de bordes y nodos generados, así como el archivo de visualización resultante. Además, presenta una lista de nodos denominados *clusters*, cada uno asociado a distintas Comunidades Autónomas de España y un rango de fechas que indica el período de los datos analizados para cada cluster. Cada nodo o cluster refleja una combinación única de regiones y fechas, mostrando cómo se agrupan y distribuyen los datos en la estructura de la red.

Capítulo 5

Discusión

5.1. Comparativa con otras técnicas

Al comparar los resultados de este estudio con investigaciones previas que han utilizado enfoques alternativos, como modelos estadísticos tradicionales y simulaciones epidemiológicas, se observan varias diferencias y ventajas clave del TDA. Los métodos tradicionales suelen centrarse en parámetros agregados y promedios, lo que puede llevar a la pérdida de información sobre la estructura subyacente de los datos y las relaciones intrínsecas entre diferentes regiones y periodos temporales.

5.1.1. Comparativa con modelos compartimentales (SIR, SEIR)

Por ejemplo, estudios basados en modelos SIR (Susceptibles, Infectados, Recuperados) [21] o SEIR (Susceptibles, Expuestos, Infectados, Recuperados) [22] han proporcionado importantes insights sobre la dinámica general de la pandemia y la efectividad de distintas intervenciones. Sin embargo, estos modelos a menudo simplifican la complejidad de la propagación del virus y no capturan adecuadamente las variaciones espaciales y las heterogeneidades regionales.

En contraste, el TDA ha permitido identificar patrones más detallados y localizados, revelando cómo la conectividad y la estructura geográfica influyen en la propagación del COVID-19. Por ejemplo, mientras que un modelo SIR podría indicar una tasa de infección promedio para todo el país, el TDA puede mostrar cómo ciertas regiones están más interconectadas y cómo los casos se distribuyen de manera desigual. Este nivel de detalle es crucial para diseñar intervenciones más precisas y efectivas.

5.1.2. Análisis geoespacial

Además, investigaciones previas que han utilizado técnicas de análisis geoespacial [23] y redes complejas [24] han demostrado la importancia de la conectividad y los flujos de movilidad en la propagación del virus. El TDA complementa estos enfoques al proporcionar una representación visual y cuantitativa de la estructura topológica de los datos, lo que facilita la identificación de patrones y la toma de decisiones informadas.

5.1.3. Modelos basados en agentes (Agent-Based Models)

Trabajos de Epstein (2009) [25] utilizan modelos basados en agentes (ABM) para contener pandemias, proporcionando una simulación detallada de la interacción entre individuos y cómo estas interacciones afectan la propagación del virus. A diferencia de los ABM, que se centran en la microdinámica de las interacciones individuales, el enfoque del TDA ofrece una visión macroscópica de la propagación de la enfermedad, capturando patrones globales y puntos críticos de infección. Mientras que los ABM son útiles para explorar escenarios específicos y el impacto de decisiones individuales, el TDA complementa esta perspectiva al revelar la estructura global de la transmisión del virus, lo que es crucial para diseñar estrategias de control a gran escala.

5.1.4. Redes neuronales y Machine Learning

Por otro lado, estudios como el de Yang et al. (2020) [26] implementa un modelo SEIR modificado junto con técnicas de inteligencia artificial para predecir la tendencia de la epidemia de COVID-19 en China bajo intervenciones de salud pública . Este enfoque combina la robustez de los modelos compartimentales con la capacidad predictiva del aprendizaje automático. Sin embargo, el modelo SEIR modificado se basa en supuestos específicos sobre la tasa de transmisión y otros parámetros epidemiológicos que pueden variar regionalmente. En contraste, el TDA no requiere tales supuestos específicos y puede adaptarse a diferentes configuraciones de datos. Esto permite al TDA identificar patrones emergentes y dinámicas espaciales sin la necesidad de ajustar previamente parámetros específicos del modelo, ofreciendo así una flexibilidad adicional en el análisis.

5.1.5. Análisis de series temporales

Petropoulos y Makridakis (2020) [27] utilizan análisis de series temporales para prever la evolución de la pandemia de COVID-19, enfocándose en la predicción de la tendencia futura basándose en datos históricos . Mientras que el análisis de series temporales es efectivo para detectar tendencias y cambios en los datos a lo largo del tiempo, este método puede no capturar adecuadamente las complejidades espaciales y la estructura topológica de los datos epidemiológicos. El TDA, por otro lado, proporciona una representación más rica al integrar dimensiones espaciales y temporales, permitiendo visualizar cómo la propagación del virus está interconectada entre diferentes regiones. Esta capacidad de combinar análisis espacial y temporal proporciona una ventaja significativa en la comprensión de la dinámica completa de la pandemia.

5.2. Comparativa con algoritmo Mapper aplicado a otras geografías

El estudio de Chen y Volić (2021) [5] aplicó el algoritmo Mapper a los datos de COVID-19 en Estados Unidos, obteniendo resultados comparables a los de este trabajo. Ambos estudios destacan la capacidad del TDA para identificar clústeres de regiones con patrones de propagación similares y para visualizar la evolución temporal de la pandemia. Sin embargo, este trabajo se centra en el contexto español y utiliza datos a nivel autonómico, lo que permite un análisis más detallado de las diferencias regio-

nales en la propagación del virus. Además, se han incorporado datos de fallecidos, lo que proporciona una visión más completa del impacto de la pandemia en España.

En contraste con el estudio de Chen y Volić, que se centró en los casos confirmados a nivel de condado en Estados Unidos y en un periodo más corto de tiempo, este trabajo amplía el análisis al considerar también los fallecimientos y utiliza datos a nivel autonómico en España durante un periodo de más de dos años. Esta diferencia en el nivel de granularidad y el periodo de tiempo permite una evaluación más precisa de las disparidades regionales en la propagación del virus y sus consecuencias a lo largo del tiempo. Además, al incluir los fallecimientos como una dimensión adicional en el análisis topológico, se obtiene una visión más completa del impacto de la pandemia en términos de morbilidad y mortalidad.

Es importante destacar que, si bien ambos estudios utilizan el algoritmo Mapper, difieren en algunos aspectos metodológicos. En este trabajo, se ha utilizado la función identidad como proyección, lo que permite preservar la mayor cantidad de información posible de los datos originales. Además, se ha empleado una normalización de los datos para garantizar una contribución equilibrada de cada dimensión en el análisis. Estas diferencias metodológicas pueden influir en los resultados y la interpretación de los mismos, y resaltan la importancia de adaptar el enfoque del TDA a las características específicas de los datos y al contexto del estudio.

Capítulo 6

Conclusiones

6.1. Conclusiones

Este estudio ha demostrado la eficacia del Análisis Topológico de Datos (TDA) como herramienta para caracterizar la propagación espacial del COVID-19 en España. Utilizando el algoritmo Mapper, se han identificado patrones significativos en la evolución temporal y geográfica de la pandemia. Los principales hallazgos incluyen la identificación de troncos y clústeres que representan regiones con alta conectividad y tasas de infección elevadas, como Cataluña y Madrid, así como componentes aislados como las Islas Canarias, Baleares y Melilla, que presentan dinámicas de propagación únicas debido a sus características geográficas y políticas locales.

El análisis topológico ha revelado que la propagación del virus en España no es homogénea, sino que está influenciada por factores regionales y temporales. Se observan diferencias significativas en la conectividad de los nodos del grafo Mapper, lo cual indica variaciones en la propagación del virus entre diferentes CCAA. Además, los datos muestran cómo las olas de mortalidad y contagios se han propagado en diferentes momentos, reflejando la influencia de factores como la densidad de población, las políticas de salud pública, y las medidas de confinamiento.

El TDA ha permitido visualizar la estructura global de la pandemia, identificando patrones de transmisión y puntos críticos de infección que otras técnicas podrían haber pasado por alto. Esto ha proporcionado una comprensión más profunda de la dinámica del COVID-19, facilitando el desarrollo de estrategias de control y prevención más efectivas.

6.2. Limitaciones

Este estudio presenta varias limitaciones que deben ser consideradas al interpretar los resultados. Una de las principales limitaciones es la falta de un dataset granularizado a nivel de provincias. La disponibilidad de datos estandarizados sólo a nivel autonómico limita la precisión del análisis topológico, ya que no se pueden captar variaciones más finas dentro de las Comunidades Autónomas.

Otra limitación importante es la frecuencia de actualización de los datos. Algunas de las fuentes consultadas no publican datos con una periodicidad diaria, sino semanal

o en otros intervalos de tiempo. Esta falta de actualización frecuente puede afectar la capacidad para detectar y analizar cambios rápidos en la propagación del virus, lo que a su vez puede influir en la precisión y utilidad de los patrones identificados por el TDA. Asimismo, dado que no se dispone para los datasets empleados datos referidos a los recuperados, se ha acotado el alcance de esta implementación exclusivamente a número de casos y fallecidos.

Además, la integración de datos de diferentes fuentes que pueden tener metodologías de recolección y reporte ligeramente distintas introduce un margen de error que puede afectar la consistencia y comparabilidad de los resultados. Es fundamental continuar trabajando en la mejora de la calidad y la resolución de los datos epidemiológicos para maximizar el potencial del TDA en futuros estudios.

6.3. Trabajos futuros

El presente proyecto ha demostrado la utilidad y eficacia del TDA para comprender la propagación espacial del COVID-19 en España. Sin embargo, existen varias direcciones futuras que podrían explorarse para ampliar y profundizar los hallazgos obtenidos. A continuación, se proponen algunas líneas de trabajo a futuro:

- **Extensión a otras geografías:**
  Ampliar el análisis a otras geografías europeas o a nivel de la Unión Europea [28] sería un paso lógico y significativo. Realizar un estudio comparativo entre diferentes países europeos podría revelar patrones de propagación y puntos críticos de infección que son comunes o distintos entre las diversas regiones. Esta comparación podría ayudar a identificar factores contextuales, como políticas de salud pública, densidad de población y movilidad, que influyen en la propagación del virus.

- **Análisis a nivel provincial/urbano:**
  Obtener y utilizar datos más detallados a nivel de provincias o ciudades permitiría un análisis más granular y preciso. Este enfoque podría revelar variaciones intra-autonómicas en la propagación del Covid-19, proporcionando información valiosa para la implementación de medidas de control más específicas y efectivas a nivel local.

- **Frecuencia de datos más alta:**
  Trabajar con datasets que proporcionen actualizaciones diarias y de alta frecuencia mejoraría significativamente la capacidad de detectar y analizar cambios rápidos en la dinámica de la pandemia. Esto podría facilitar una respuesta más ágil y efectiva a brotes emergentes y ayudar a ajustar las políticas de salud pública en tiempo real.

- **Integración de datos socioeconómicos y demográficos:**
  Incorporar variables socioeconómicas y demográficas adicionales, como densidad de población, edad, ingresos, y niveles de movilidad, podría enriquecer el análisis topológico. Esto permitiría explorar cómo estos factores influyen en la propagación del virus y podrían proporcionar insights adicionales para el diseño de intervenciones de salud pública más equitativas y efectivas.

- **Desarrollo de modelos predictivos:**
  Utilizar los hallazgos obtenidos mediante el TDA para desarrollar modelos pre-

dictivos que puedan anticipar futuras olas de contagio. Estos modelos podrían ser validados y calibrados utilizando datos históricos y aplicados para predecir la propagación del virus bajo diferentes escenarios, mejorando así la planificación y preparación ante futuras pandemias.

- **Aplicación a otras enfermedades infecciosas:**
  Explorar la aplicación del TDA en el análisis de otras enfermedades infecciosas más allá del COVID-19. Este enfoque podría proporcionar nuevas perspectivas y herramientas para el manejo de enfermedades como la gripe, el dengue, o cualquier otro patógeno con un componente de propagación espacial significativo.
- **Colaboración entre disciplinas:**
  Fomentar la colaboración entre expertos en salud pública, epidemiología, ciencia de datos, y matemáticas para mejorar y refinar las metodologías de TDA aplicadas en estudios epidemiológicos. Esta colaboración interdisciplinaria puede conducir a innovaciones metodológicas y a una mejor comprensión de los datos complejos en salud pública.

Bibliografía

[1] W. H. Organization, "World health organization - coronavirus situation report," https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports, 2020.

[2] U. C. for Disease, Control, and Prevention, "Cdc - about covid-19," https://www.cdc.gov/covid/about/index.html, 2019.

[3] C. Pagel and C. A. Yates, "Role of mathematical modelling in future pandemic response policy," *BMJ*, vol. 378, 2022. [Online]. Available: https://www.bmj.com/content/378/bmj-2022-070615

[4] G. de España | Boletín Oficial del Estado, "Medidas urgentes de salud pública y económica para hacer frente al covid-19." https://www.boe.es/buscar/act.php?id=BOE-A-2024-12379, 2024.

[5] Y. Chen and I. Volić, "Topological data analysis model for the spread of the coronavirus," *PloS one*, vol. 16, no. 8, p. e0255584, 2021.

[6] T. B. Oehmke, L. A. Post, C. B. Moss, T. Z. Issa, M. J. Boctor, S. B. Welch, and J. F. Oehmke, "Dynamic panel data modeling and surveillance of covid-19 in metropolitan areas in the united states: Longitudinal trend analysis," *J Med Internet Res*, vol. 23, no. 2, p. e26081, Feb 2021. [Online]. Available: https://www.jmir.org/2021/2/e26081

[7] G. Singh, F. Memoli, and G. Carlsson, "Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition," in *Eurographics Symposium on Point-Based Graphics*, M. Botsch, R. Pajarola, B. Chen, and M. Zwicker, Eds. The Eurographics Association, 2007.

[8] F. Å. Nielsen, "Clustering of scientific citations in wikipedia," in *Wikimania 2008*. Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby: Informatics and Mathematical Modelling, Technical University of Denmark, jun 2008. [Online]. Available: http://www2.compute.dtu.dk/pubdb/pubs/5666-full.html

[9] G. Petri, M. Scolamiero, I. Donato, and F. Vaccarino, "Topological strata of weighted complex networks," *PLOS ONE*, vol. 8, no. 6, pp. 1–8, 06 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0066506

[10] C. M. Topaz, L. Ziegelmeier, and T. Halverson, "Topological data analysis of biological aggregation models," *PLOS ONE*, vol. 10, no. 5, pp. 1–26, 05 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0126383

[11] S. T. TowardsDataScience, "Topological data analysis (tda)," https://towardsdatascience.com/topological-data-analysis-tda-b7f9b770c951, 2022.

[12] H. Edelsbrunner and J. Harer, "Persistent homology—a survey," *Discrete and Computational Geometry - DCG*, vol. 453, 01 2008.

[13] F. Chazal and B. Michel, "An introduction to topological data analysis: Fundamental and practical aspects for data scientists," *Frontiers in Artificial Intelligence*, vol. 4, 2021. [Online]. Available: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2021.667963

[14] E. Munch, "A user's guide to topological data analysis," *Journal of Learning Analytics*, vol. 4, pp. 47–61, 07 2017.

[15] L.-Y. Ma, T. Feng, C. He, M. Li, K. Ren, and J. Tu, "A progression analysis of motor features in parkinson's disease based on the mapper algorithm," *Frontiers in Aging Neuroscience*, vol. 15, p. 1047017, 02 2023.

[16] T. Liao, Y. Wei, M. Luo, G.-P. Zhao, and H. Zhou, "tmap: an integrative framework based on topological data analysis for population-scale microbiome stratification and association studies," *Genome Biology*, vol. 20, 12 2019.

[17] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Knowledge Discovery and Data Mining*, 1996. [Online]. Available: https://api.semanticscholar.org/CorpusID:355163

[18] I. de Salud Carlos III, Instituto de Salud Carlos III, 2024, disponible en: https://www.isciii.es/. [Online]. Available: https://www.isciii.es/

[19] M. de Sanidad, Ministerio de Sanidad, 2024, disponible en: https://www.sanidad.gob.es/. [Online]. Available: https://www.sanidad.gob.es/

[20] H. J. van Veen, N. Saul, D. Eargle, and S. W. Mangham, "Kepler Mapper: A flexible Python implementation of the Mapper algorithm," Oct. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4077395

[21] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 115, no. 772, pp. 700–721, 1927.

[22] L. J. Allen, "An introduction to stochastic epidemic models," in *Mathematical Epidemiology*. Springer, 2008, pp. 81–130.

[23] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Physical Review Letters*, vol. 86, no. 14, p. 3200, 2001.

[24] A. J. Tatem, D. J. Rogers, and S. I. Hay, "Global transport networks and infectious disease spread," *Advances in Parasitology*, vol. 62, pp. 293–343, 2006.

[25] J. M. Epstein, "Modelling to contain pandemics," *Nature*, vol. 460, no. 7256, pp. 687–687, 2009.

[26] Z. Yang, Z. Zeng, K. Wang, S. S.-S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. Mai *et al.*, "Modified seir and ai prediction of the epidemics trend of

covid-19 in china under public health interventions," *Journal of Thoracic Disease*, vol. 12, no. 3, pp. 165–174, 2020.

[27] F. Petropoulos and S. Makridakis, "Forecasting the novel coronavirus covid-19," *PLOS ONE*, vol. 15, no. 3, p. e0231236, 2020.

[28] E. C. for Disease Prevention and Control, "Surveillance and disease data - covid 19 data," European Centre for Disease Prevention and Control, 2024, disponible en: https://www.ecdc.europa.eu/en/covid-19/data. [Online]. Available: https://www.ecdc.europa.eu/en/covid-19/data

Índice de figuras

Anexo

Anexo 1: Repositorio de Github del proyecto

https://github.com/MrAndrada/tda-mapper-covid19-spain