CIMAT Centro de Investigación en Matemáticas, A.C.

Análisis topológico de datos: Estudio de la movilidad poblacional en la ciudad de Hermosillo.

T E S I S

Que para obtener el grado de

Maestro en Ciencias

con especialidad en

Probabilidad y Estadística

Presenta

Martin Alonso Flores Gonzalez

Director de Tesis:

Dra. Lilia Leticia Ramírez Ramírez

Co-director de Tesis:

Dr. Jesús Francisco Espinoza Fierro

Autorización de la versión final

Guanajuato, Gto a 26 de Noviembre de 2021

Agradecimientos

A mi familia, que a lo largo del tiempo y a pesar de las adversidades, han sido una fuente de motivación constante a lo largo de mis estudios y formación profesional.

A la Dra. Leticia Ramírez y al Dr. Jesús Espinoza, por la paciencia y confianza que me tuvieron a lo largo de este tiempo. Su atención, explicaciones y guía fueron fundamentales para llevar a buen término este trabajo.

Al Dr. José Montoya cuyos comentarios siempre fueron de gran ayuda para aclarar las ideas y el camino que el trabajo realizado debería tomar.

Al Dr. Miguel Nakamura, por tomarse el tiempo de leer y aportar su experiencia en la revisión del trabajo que se presenta a continuación.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por aportar el financiamiento necesario durante mis estudios de maestría. Al Centro de Investigación en Matemáticas (CIMAT) por darme la oportunidad de seguir preparándome profesionalmente y también a los profesores, que pese a las dificultades de la pandemia, siempre estuvieron atendiendo sus obligaciones de la mejor manera.

Resumen

A lo largo del presente trabajo se introduce al lector en el Análisis Topológico de Datos y a su aplicación en gráficas no dirigidas. Se propone una metodología para la construcción y análisis de redes de movilidad poblacional, se construyen las redes de movilidad para la ciudad de Hermosillo a partir de datos georreferenciados de la población obtenidos mediante aplicaciones para celulares. Se muestra una metodología en la que se propone utilizar el Análisis Topológico de Datos para estudiar el cambio de la movilidad en la ciudad a lo largo del tiempo, esto al comparar las persistencias de las características topológicas de cada red de movilidad y se muestran los resultados obtenidos al aplicar esta metodología a las redes construidas. Por último, se argumenta la importancia de conocer los cambios en la movilidad de la ciudad cuando se estudian fenómenos de contacto y en particular, en el contexto de la pandemia de covid19, se argumenta el posible uso para obtener alertas tempranas de contagios.

Palabras Clave

ATD, filtración, persistencia, redes, movilidad, población.

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1

Introducción

Gracias a las plataformas digitales, hoy en día se puede generar y almacenar una gran cantidad de información, más que en cualquier otra época de la historia. Pero al igual que algún recurso natural, poco sirve tener muchos datos sin haberlos procesado de alguna manera y, es gracias al procesamiento de los datos que se puede llegar a obtener información sustancialmente útil para algún fin en particular.

A lo largo del siglo pasado se crearon varias herramientas estadísticas con la finalidad de poder estudiar la escasa información que se recopilaba en dicha época. Hoy en día, incluso con equipos de cómputo de gama media, la información que se puede manejar es mayor en varios órdenes de magnitud, lo cual pone de manifiesto la necesidad de contar con distintas herramientas eficientes para el análisis de datos masivos y complejos.

En los últimos 40 años se han desarrollando varias herramientas matemáticas que logran brindar información importante de un conjunto masivo de datos, algunas de las herramientas desarrolladas muestran diferentes características de los datos, haciendo

que se tengan diferentes opciones para obtener información valiosa de un conjunto de datos. Una de tales herramientas es conocida como Análisis Topológico de Datos (ATD), la cual se desarrolló formalmente a inicios del siglo XXI, aún cuando los principales conceptos matemáticos en los que se fundamenta aparecieron en la literatura a finales de los años 90.

El ATD combina elementos de topología algebraica (homología persistente) y combinatoria (gráficas y estructuras simpliciales) para el estudio de la forma (esferas homológicas) de bases de datos (nubes de puntos), dotando a estas últimas de una familia de estructuras geométricas no triviales (filtración simplicial) e identificando aquellas características geométricas "más significativas" (clases de homología persistente). Las aplicaciones de este enfoque han permitido identificar características en conjuntos de datos que no son fácilmente identificables por las herramientas estándar de procesamiento de señales (Wang, Ombao, y Chung (2019)) o del análisis de agrupamiento (Nicolau, Levine, y Carlssona (2011)). Tales aplicaciones incluyen áreas de la salud (Dey y Mandal, 2018; Nielson y cols., 2015; Oyama y cols., 2019), finanzas (Gidea y Katz, 2018; Goel, Pasricha, y Mehra., 2020), ingeniería (Clough y cols., 2020; Smith, Dłotko, y Zavala, 2021; Sørensen, Biscio, Bauchy, Fajstrup, y Smedskjaer, 2020.), meteorología (Dongjin, Bresten, Youm, Seo, y Jung, 2020; Muszynski, Kashinath, Kurlin, y Wehner, 2019; Ofori-Boateng, Lee, Gorski, Garay, y Gel, 2021), etc.

En el presente trabajo se crean redes de movilidad, en las que se desarrolla un análisis topológico de datos con el objetivo de identificar cambios significativos en la movilidad poblacional, entre distintas ventanas de tiempo (con longitud de 1 semana) enmarcadas en el contexto de la contingencia sanitaria por covid19, a fin de contar con un indicador que permita obtener señales tempranas de un potencial aumento en el número de contagios. Para realizar el estudio se utiliza una base de datos de movilidad de la población en la ciudad de Hermosillo; estos datos contienen la fecha y ubicación registrada por los equipos de telefonía celular que registran algunas de las

aplicaciones instaladas en los equipos con permisos de geolocalización habilitados.

La gráfica subyacente de cada una de las redes de movilidad (construidas para cada ventana de tiempo de 1 semana, entre el 21 de septiembre de 2020 y el 13 de diciembre de 2020), está definida por el conjunto de vértices dado por la colección de las áreas geoestadísticas básicas (AGEB) de la ciudad de Hermosillo (582 en total) y existe una arista entre dos AGEB si existe una movilidad poblacional directa entre estas. Luego, la red de movilidad, definida en una ventana de tiempo concreta, consiste en ponderar las aristas de la gráfica subyacente con el número de personas que se desplazan directamente entre las correspondientes AGEB, en la ventana de tiempo en cuestión.

Aún cuando dos redes de movilidad tengan la misma gráfica subyacente (o bien, sean isomorfas), éstas pueden representar condiciones de movilidad poblacional muy distintas. El enfoque planteado en el presente trabajo para distinguir distintas redes de movilidad, es a través de los grupos de homología persistentes asociados a las filtraciones inducidas en el complejo clique (complejo simplicial de subgráficas completas) por las ponderaciones definidas en las aristas de la red de movilidad poblacional.

Aunque es posible intuir que la estructura de cada red tomará un rol importante en la dinámica de transmisión de un agente infeccioso, el enfoque de este trabajo es la creación de las redes de movilidad, la revisión de las herramientas necesarias de ATD en redes, para luego proponer filtraciones que puedan auxiliar a la identificación de algunas de las características más relevantes en una red. Se propone el uso de medidas de diferencias topológicas para la identificación de cambios de patrones de movilidad.

Conocer la similitudes entre la movilidad de diferentes semanas podría ser importante para poder planear servicios o eventos masivos en la ciudad. Particularmente en el contexto de la actual pandemia por covid19, se puede medir el impacto de la movilidad en el incremento de los contagios y con ello el gobierno podría anticipar la

preparación de insumos y aparatos necesarios para evitar el colapso de los servicios médicos, así como implementar las medidas justas para el control y eventual levantamiento de las restricciones.

El presente trabajo se encuentra organizado de la siguiente manera. En el Capítulo 2 se presentan los conceptos básicos usados en ATD. Se introducen los complejos simpliciales abstractos y geométricos, se exponen los complejos simpliciales de Čech y Vietoris-Rips, que son clásicos en el análisis de datos. Se introduce a la homología simplicial y se desarrolla el contenido para llegar a definir los grupos de homología del complejo simplicial. Por último, se presentan las herramientas utilizadas para el estudio de la homología persistente y se acompañan con un sencillo ejemplo para ilustrar la persistencia de las clases de homología.

En el Capítulo 3 se presenta una introducción a la teoría de gráficas, en donde se muestran las herramientas necesarias para la construcción de las redes de movilidad. También se expone la relación que las gráficas tienen con los complejos simpliciales geométricos, se define al complejo simplicial clique y algunas filtraciones que son clásicas en el ATD en redes. Por último, se presenta un ejemplo sencillo donde se explica cómo se desarrolla el estudio de la homología persistente en una gráfica ponderada.

En el capítulo 4 se detallan las principales características de los datos con los que se cuenta para realizar el trabajo, el proceso con el que se depuran y el algoritmo utilizado para la construcción de una matriz de "saltos" con la cual se genera la matriz de adyacencia de cada red. Por último, se expone el contexto de la ciudad de Hermosillo y se compara la red obtenida a lo largo del periodo de estudio con la realidad de la ciudad.

En el Capítulo 5 se conjunta el material presentado hasta entonces, con el fin de realizar el análisis topológico de los datos de movilidad poblacional de la ciudad de Hermosillo entre el 21 de septiembre de 2020 y el 13 de diciembre de 2020. En este

capítulo se explica la metodología propuesta para comparar la movilidad de cada semana. Se definen las filtraciones que se utilizan en el estudio y se presentan algunos ejemplos que ilustran la evolución de cada filtración; también se argumenta sobre el significado y la importancia de las características topológicas en las redes. Por último, se presentan y discuten los resultados obtenidos sobre la evolución de la movilidad en ventanas de tiempo de 1 semana.

Finalmente, en el Capítulo 6 se discuten más ampliamente los resultados, limitaciones y posibles líneas de investigación futura sobre el enfoque presentado en el presente trabajo para el estudio de la movilidad poblacional.

CAPÍTULO 2

Análisis topológico de datos

2.1. Conceptos básicos

Antes de comenzar a definir los objetos con los que se trabaja en el análisis topológico de datos, es conveniente definir algunos conceptos básicos de topología que serán útiles para el correcto desarrollo de los ejemplos que se mostrarán en las diferentes secciones de este capítulo.

Una de las maneras más recurrentes de representar los resultados de un experimento es por medio de arreglos de la forma $\{x_1, x_2, \dots\}$ para $x_i \in \mathbb{R}$, pero difícilmente este tipo de arreglos dicen algo por sí mismos ya que la información que realmente ayuda a la descripción de los resultados se obtiene a partir de comparaciones realizadas dentro del mismo contexto. Matemáticamente, este tipo de configuración se captura mediante la noción de espacio métrico, que se define a continuación (Rudin, 1980).

***Definición* 2.1.** *Un conjunto X cuyos elementos llamaremos puntos, se dice que es un espacio métrico si a cada dos puntos p y q de X hay asociado un número real*

$d(p,q)$ llamado distancia de p a q, tal que

1. *$d(p,q) > 0$ si $p \neq q$ y $d(p,q) = 0$ sí y solo sí $p = q$.*

2. *$d(p,q) = d(q,p)$.*

3. *$d(p,q) \leq d(p,r) + d(r,q)$ para todo $r \in X$.*

A cualquier función que cumpla con las tres condiciones anteriores se le conoce como función distancia o métrica. Además, si (X,d) es un espacio métrico e $Y \subset X$, entonces $(Y, d|_Y : Y \times Y \to \mathbb{R})$ es un espacio métrico también y a $d|_Y$ se le conoce como la métrica inducida sobre Y.

El ejemplo estándar de un espacio métrico es el formado por $\mathbb{R}^n$ y la distancia euclidiana, en donde los puntos en $\mathbb{R}^n$ son de la forma $p = (p_1, p_2, \ldots, p_n)$ y por lo tanto si $x, y \in \mathbb{R}^n$ la distancia euclidiana está dada por

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

En general, es posible utilizar una métrica d para crear subconjuntos de X; este tipo de subconjuntos tienen especial importancia en muchas áreas de las matemáticas ya que a partir de ellos se pueden definir distintas estructuras.

***Definición* 2.2.** *Si (X,d) es un espacio métrico, entonces para todo $x \in X$ y $r > 0$ el conjunto*

$$B(x,r) = \{y \in X; d(x,y) < r\}$$

es llamado bola abierta de radio r y centrada en x. La bola cerrada de radio r centrada en x se define como sigue

$$B_c(x,r) = \{y \in X; d(x,y) \leq r\}.$$

Con lo anterior es posible definir formalmente a un conjunto abierto en un espacio métrico tal y como se muestra a continuación (Rudin, 1980).

***Definición* 2.3.** *Sea (X, d) un espacio métrico, un subconjunto $A \subset X$ se dice abierto si para cada $a \in A$, existe $r_a > 0$ (que depende sólo de a) tal que $B(a, r_a) \subset A$.*

Ahora se definirá el concepto de convergencia en espacios métricos y para ello se tiene lo siguiente (Rudin, 1980).

***Definición* 2.4.** *Una sucesión $\{x_n\}_{n\in\mathbb{N}}$ de elementos de un espacio métrico (X, d) converge a un límite x si*

$$\lim_{n\to\infty} d(x_n, x) = 0.$$

Como se menciona en Blumberg (2020), la idea que motiva la topología de un conjunto de puntos es la de mostrar equivalencias entre objetos en un sentido mucho más amplio que la concebida por una métrica y definir una noción de cercanía más débil y flexible que aún nos permita formalizar las nociones que conducen al cálculo (es decir, continuidad y convergencia), incluso se puede pensar en un espacio topológico como un conjunto con una colección de subconjuntos que se comportan bien y que actúan como bolas abiertas en los espacios métricos.

***Definición* 2.5.** *Una topología en un conjunto X es una familia $T \subseteq 2^X$ de subconjuntos de X, tal que:*

1. *$\emptyset, X \in T$.*
2. *Si $S_1, S_2 \in T$, entonces $S_1 \cap S_2 \in T$.*
3. *Si $\{S_j; j \in J\} \subseteq T$, siendo J un conjunto de índices, entonces $\bigcup_{j\in J} S_j \in T$.*

El par (X, T) se llama espacio topológico, además a los elementos de la topología T se les conoce como conjuntos abiertos y a los respectivos complementos se les conoce como conjuntos cerrados.

Un ejemplo sencillo de un espacio topológico sería el formado por el conjunto $X = \{1, 2, 3\}$ y la familia de subconjuntos $T = \{\emptyset, \{1\}, \{1, 2\}, \{1, 3\}, \{1, 2, 3\}\}$.

***Definición* 2.6.** *Sea (X, T) un espacio topológico y $A \subseteq X$ con la topología inducida $T_A = \{S \cap A; S \in T\}$, entonces (A, T_A) es un subespacio topológico de (X, T).*

Llegados a este punto se tiene que resaltar que un espacio métrico no es un espacio topológico, sin embargo todo espacio métrico da lugar a un espacio topológico. Para ejemplificar lo anterior, tomemos al espacio métrico formado por $(\mathbb{R}, d)$, donde d es la distancia euclidiana. Si se toman a $x, y \in \mathbb{R}$ tales que $x < y$, entonces es posible definir a conjuntos abiertos de la forma $A_{x,y} = \{a \in \mathbb{R}; x < a < y\}$, los cuales servirían de base para la topología usual de $\mathbb{R}$, esto es, la topología formada por $\mathbb{R}$, $\emptyset$ y las uniones e intersecciones finitas de algunos conjuntos de la forma $A_{x,y}$ con diferentes x e y.

Teniendo claro lo anterior, resulta conveniente mostrar las siguientes definiciones que introducen al concepto de cubierta en espacios topológicos, estas son definiciones importantes para el teorema del nervio que se mostrará en la segunda sección de este capítulo (Jean-Daniel Boissonnat y Yvinec, 2018).

***Definición* 2.7.** *Sea I un conjunto de índices. Una cubierta abierta de un conjunto U en un espacio topológico (X, T) es una colección $\{U_i\}_{i \in I}$ de subconjuntos abiertos $U_i \subset X$, tales que $U \subseteq \bigcup_{i \in I} U_i$.*

***Definición* 2.8.** *Sea I un conjunto de índices. Una cubierta abierta de un espacio topológico (X, T) es una colección $\{U_i\}_{i \in I}$ de subconjuntos abiertos $U_i \subseteq X$, tales que $X = \bigcup_{i \in I} U_i$.*

***Definición* 2.9.** *Un espacio topológico (X, T) es un espacio compacto si cualquier cubierta abierta de X admite una subcubierta finita, es decir, para cualquier familia $\{U_i\}_{i \in I}$ de conjuntos abiertos tales que $X = \bigcup_{i \in I} U_i$ existe un subconjunto finito $J \subseteq I$ tal que $X = \bigcup_{j \in J} U_j$.*

Cabe señalar que en un espacio métrico, la compacidad se puede caracterizar usando secesiones, en donde un espacio métrico X es compacto si y sólo si cualquier sucesión en X tiene una subsucesión convergente (Rudin, 1980).

Otro concepto fundamental en la topología es el de conexidad, el cual se define a continuación (Jean-Daniel Boissonnat y Yvinec, 2018).

***Definición* 2.10.** *Un espacio topológico (X, T) es conexo si no es la unión de dos conjuntos abiertos disjuntos, es decir, si O_1 y O_2 son dos conjuntos abiertos disjuntos tales que $X = O_1 \cup O_2$, entonces $O_1 = \emptyset$ u $O_2 = \emptyset$.*

***Definición* 2.11.** *Sea $[0, 1] \subset \mathbb{R}$ el subespacio topológico de $\mathbb{R}$ con la topología usual. Un espacio topológico (X, T) es conexo por caminos si para cualquier $x, y \in X$ existe una función continua $\gamma : [0, 1] \to X$ tal que $\gamma(0) = x$ y $\gamma(1) = y$.*

Para formalizar la noción de equivalencia entre espacios topológicos, es necesario el concepto de función continua, el cual se define como sigue (Jean-Daniel Boissonnat y Yvinec, 2018).

***Definición* 2.12.** *Una función $f : X \to X'$ entre dos espacios topologicos (X, T) y (X', T') es continua si y sólo si, la preimagen $f^{-1}(O') = \{x \in X; f(x) \in O'\}$ de cualquier conjunto abierto $O' \subset X'$ es un conjunto abierto en X. Equivalentemente, f es continua si y sólo si, la preimagen de un conjunto cerrado en X' es un conjunto cerrado en X.*

Existen varias formas de medir la cercanía entre dos objetos y en particular se puede distinguir entre los criterios topológicos y los criterios geométricos. En topología, dos espacios topológicos se consideran iguales, o equivalentes, cuando son homeomorfos (Jean-Daniel Boissonnat y Yvinec, 2018).

***Definición* 2.13.** *Dos espacios topológicos (X, T), (X', T') son homeomorfos si existe una función continua y biyectiva $h : X \to X'$ tal que la función inversa $h^{-1} : X' \to X$ también es continua. La función h es llamada homeomorfismo.*

Para distinguir entre espacios, cuando éstos son subespacios de $\mathbb{R}^d$, la noción de isotopía es más fuerte que la noción de homeomorfismo.

***Definición* 2.14.** *Una isotopía entre $X \subset \mathbb{R}^d$ y $X' \subset \mathbb{R}^d$ es una función continua $F : \mathbb{R}^d \times [0, 1] \to \mathbb{R}^d$ tal que $F(\cdot, 0) : \mathbb{R}^d \to \mathbb{R}^d$ es la función identidad y $F(X, 1) = X'$, y para cualquier $t \in [0, 1]$, $F(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^d$ es un homeomorfismo de $\mathbb{R}^d$.*

Que dos espacios topológicos X y X' sean isotópicos significa que X se puede deformar continuamente en X' sin crear auto intersecciones o cambios topológicos. La noción de isotopía es más fuerte que la de homeomorfismo en el sentido de que si X y X' son isotópicos también son homeomorfos, mientras que dos subespacios de $\mathbb{R}^d$ homeomorfos pueden no ser isotópicos. Este hecho es el que da pie a la teoría de nudos.

En general es difícil decidir qué espacios son homeomorfos, por ello a veces es conveniente trabajar con una noción más débil de equivalencia llamada equivalencia homotópica (Jean-Daniel Boissonnat y Yvinec, 2018).

***Definición* 2.15.** *Sean (X, T) y (X', T') dos espacios topológicos. Dos funciones continuas $f, g : X \to X'$ son homotópicas si existe una función continua $h : X \times [0, 1] \to X'$, llamada homotopía, tal que*

1. $h(x, 0) = f(x)$,
2. $h(x, 1) = g(x)$.

Se denota por $f \simeq g$ cuando f y g son homotópicas.

***Definición* 2.16.** *Sean (X, T) y (X', T') dos espacios topológicos. Decimos que X y X' son homotópicamente equivalentes si existen funciones continuas $f : X \to X'$ y $g : X' \to X$ tales que $f \circ g \simeq id_{X'}$ y $g \circ f \simeq id_X$, donde $id_{X'}$ e id_X denotan las funciones identidad sobre X' y X, respectivamente. En este caso se denota por $X \simeq X'$ para referirse a que X y X' son homotópicamente equivalentes.*

En particular, son de interés aquellos espacios topológicos que sean homotópicamente equivalentes a un punto, ya que estos permiten definir una cubierta buena como se mostrará en la siguiente sección.

***Definición* 2.17.** *Un espacio contráctil es un espacio que es homotópicamente equivalente a un punto.*

En general, es difícil probar la equivalencia homotópica a partir de la definición, pero cuando $X' \subset X$ los siguientes criterios son útiles para verificar la equivalencia homotópica entre X y X'.

***Proposición* 2.1** (Jean-Daniel Boissonnat y Yvinec 2018)**.** *Si $X' \subset X$ y existe una función continua $H : X \times [0,1] \to X$ tal que*

1. $\forall x \in X, H(x,0) = x$,
2. $\forall x \in X, H(x,1) \in X'$,
3. $\forall x' \in X', \forall t \in [0,1], H(x',t) \in X'$,

entonces X y X' son homotópicamente equivalentes.

***Definición* 2.18.** *Si en la anterior proposición se cambia la tercer propiedad por la siguiente*

$$\forall x' \in X', \forall t \in [0,1], H(x',t) = x',$$

entonces a H se le conoce como retracto por deformación de X a X'.

Una manera clásica de caracterizar e identificar las propiedades de los espacios topológicos, es considerando los invariantes topológicos. Los invariantes son objetos matemáticos asociados a cada espacio topológico que tienen la propiedad de ser equivalentes para espacios homeomorfos.

2.2. Complejos simpliciales

En el análisis topológico de datos se utilizan complejos simpliciales, los cuales permiten introducir un modelo combinatorio de espacios topológicos adaptados a un conjunto de datos mediante la creación de subconjuntos de datos de acuerdo a algún criterio en específico. Más adelante se muestran algunos ejemplos concretos de complejos simpliciales, pero antes es necesario definirlos.

Veamos primero una definición de complejo simplicial de un modo combinatorio, para verlo luego con un enfoque más geométrico (Jean-Daniel Boissonnat y Yvinec, 2018).

***Definición* 2.19.** *Sea $V = \{v_1, v_2, ..., v_n\}$ un conjunto finito. Un complejo simplicial abstracto Δ con vértices V es una colección de subconjuntos de V que satisfacen las siguientes condiciones:*

1. *Los elementos de V pertenecen a Δ.*

2. *Si $\sigma \in \Delta$ y $\tau \subset \sigma$, entonces $\tau \in \Delta$.*

Cada elemento $\sigma \in \Delta$ se llama simplejo de dimensión k o k-simplejo si $card(\sigma) = k + 1$, la dimensión del complejo simplicial Δ es igual a la dimensión de su simplejo con mayor dimensión, a $\tau \subset \sigma$ se le llama cara del simplejo σ y se denota $\tau \leq \sigma$, cabe señalar que un k-simplejo tiene 2^{k+1} caras, contando el vacío y al propio simplejo.

Para ilustrar la anterior definición, se puede tomar en cuenta al conjunto $V = \{v_1, v_2, v_3, v_4\}$ junto con el respectivo conjunto potencia de V, de manera que todos los posibles subconjuntos de V están en el complejo simplicial Δ. Si se toma en cuenta al 2-simplejo formado por $\{v_1, v_3, v_4\}$, las posibles caras serían las siguientes

- un 2-simplejo: $\{v_1, v_3, v_4\}$,
- tres caras que a su vez son 1-simplejos: $\{v_1, v_3\}, \{v_1, v_4\}, \{v_3, v_4\}$,
- tres caras que son a la vez 0-simplejos: $\{v_1\}, \{v_3\}, \{v_4\}$,
- el conjunto vacio: $\emptyset$.

Habiendo definido lo anterior, resulta conveniente definir formalmente el concepto de subcomplejo simplicial ya que se estará utilizando en las siguientes secciones y es un concepto fundamental para el estudio de la homología persistente.

***Definición* 2.20.** *Sea Δ un complejo simplicial. Un complejo simplicial Δ' se dice que es un subcomplejo simplicial de Δ si $\Delta' \subseteq \Delta$.*

Es necesario mencionar que la anterior definición también aplica para complejos simpliciales geométricos (Def 2.31). A continuación se presenta la definición de un subcomplejo simplicial particular, el cual será de vital importancia más adelante.

***Definición* 2.21.** *Sea $n \in \mathbb{N}$. El $n-$ésimo esqueleto ($n-$esqueleto) de un complejo simplicial Δ de dimensión mayor o igual a n, consiste en un subcomplejo simplicial de Δ formado por todos los simplejos de dimensión menor o igual que n, es decir,*

$$K_n = \{\sigma \in \Delta; \dim(\sigma) \leq n\}.$$

En el contexto del análisis topológico de datos, es usual trabajar con datos que pueden considerarse como pertenecientes a $\mathbb{R}^n$, de manera que es posible construir complejos simpliciales que utilizan las bolas cerradas sobre el espacio métrico euclidiano en el que se sitúan los datos (Blumberg, 2020) .

***Definición* 2.22.** *Sea $X \subset \mathbb{R}^n$ un subespacio finito y sea $\epsilon \geq 0$ un número fijo. El complejo de Čech $C_\epsilon(X)$ es el complejo simplicial abstracto con*

1. *los puntos en X como $0-$simplejos,*
2. *un $k-$simplejo $[v_0, v_1, ..., v_k] \in C_\epsilon(X)$ cuando el conjunto de puntos $\{v_0, v_1, ..., v_k\}$ satisface*

$$\bigcap_i B_c(v_i, \epsilon) \neq \emptyset.$$

En la anterior definición, $B_c(v_i, \epsilon)$ representa la bola cerrada de radio ϵ centrada en v_i (ver Definición 2.2) y cabe señalar que el radio de las bolas juega un papel fundamental, ya que la dimensión de los simplejos está directamente relacionada con el tamaño de ϵ. El complejo de Čech proporciona una manera de asignar un complejo simple a un subespacio métrico finito de $\mathbb{R}^n$, sin embargo para construir el complejo de Čech se necesita poder decidir si la intersección de las bolas es no vacía y esta es una tarea no trivial en grandes dimensiones, debido al costo computacional que

conlleva realizar ésta verificación.

La Figura 2.1 muestra una representación de un complejo de Čech para puntos en $\mathbb{R}^2$, en donde se puede observar cómo cada punto v_i para $i = 0, 1, \ldots, 7$ es parte del complejo, además se forman $1-$simplejos cuando $B_c(v_i, \epsilon) \cap B_c(v_j, \epsilon) \neq \emptyset$ para $i \neq j$. También se muestra un $2-$simplejo formado por $\{v_1, v_6, v_7\}$ y se puede observar claramente que $B_c(v_1, \epsilon) \cap B_c(v_6, \epsilon) \cap B_c(v_7, \epsilon) \neq \emptyset$.

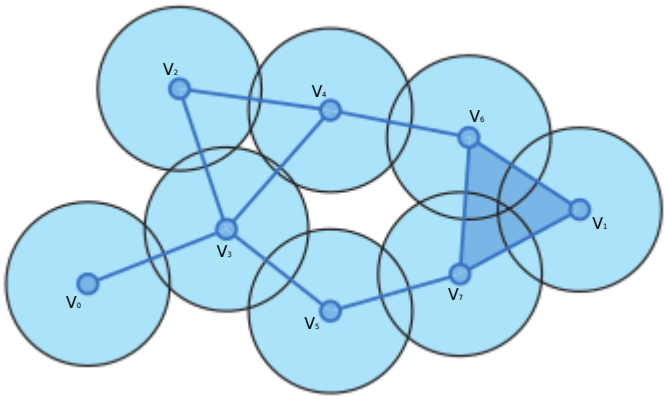


Figura 2.1: Complejo de Čech para un conjunto de puntos en $\mathbb{R}^2$

El complejo de Vietoris-Rips es el complejo simplicial máximal determinado por los vértices y los $1-$simplejos especificados por una gráfica G (Blumberg, 2020).

***Definición* 2.23.** *Sea (X, d_X) un espacio métrico finito y sea $\epsilon \geq 0$ un número fijo. El complejo de Vietoris-Rips $VR_\epsilon(X, d_X)$ es el complejo simplicial abstracto con*

1. *los puntos en X como $0-$simplejos,*

2. *un $k-$simplejo $[v_0, v_1, ..., v_k] \in VR_\epsilon(X, d_X)$ cuando*

$$d_X(v_i, v_j) \leq 2\epsilon \text{ para todo } 0 \leq i, j \leq k.$$

Al igual que en el complejo de Čech, el valor de ϵ se encuentra íntimamete relacionado a la dimensión de los simplejos en el complejo de Vietoris-Rips, sin embargo

en este último se pueden formar simplejos de dimensión mayor a uno aún cuando las intersecciones de conjuntos de más de dos bolas estén vacías. Para ejemplificar lo anterior véase la Figura 2.2 en donde se forma un $2-$simplejo con los puntos $\{v_2, v_3, v_4\}$, pero $B_c(v_2, \epsilon) \cap B_c(v_3, \epsilon) \cap B_c(v_4, \epsilon) = \emptyset$.

A pesar de las diferencias que tienen los complejos de Čech y de Vietoris-Rips, estos tienen una relación estrecha y cumplen las siguientes inclusiones.

***Lema* 2.1** (Blumberg 2020)**.** *Sea $X \subset \mathbb{R}^n$ un subespacio finito y sea $\epsilon \geq 0$ un número fijo. Se cumplen las siguiente inclusiones simpliciales:*

$$C_\epsilon(X) \subseteq VR_\epsilon(X, d_X) \subseteq C_{2\epsilon}(X).$$

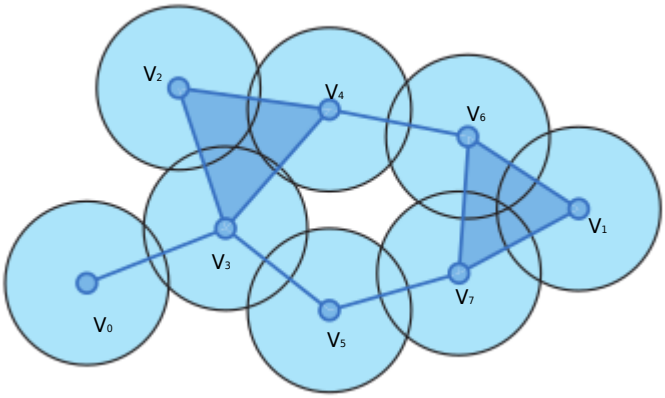


Figura 2.2: Complejo de Vietoris-Rips para un conjunto de puntos en $\mathbb{R}^2$.

Es importante resaltar que el complejo Čech es un caso especial de una construcción estándar de topología algebraica que asocia un complejo simplicial a la cubierta de un espacio. Tomando en cuenta la definición de una cubierta abierta para el espacio topológico (X, T) (Definición 2.8), se extiende la definición al nervio de la cubierta como sigue (Blumberg, 2020).

***Definición* 2.24.** *Sea $U = \{U_i\}_{i \in I}$ una cubierta abierta de un espacio topológico (X, T). Se define al nervio de la cuvierta U, como el complejo simplicial abstracto $C(U)$ con conjunto de vértices U y $k-$simplejos*

$$\sigma_k = \{U_{i_0}, U_{i_1}, \ldots, U_{i_k}\} \in C(U) \text{ si y sólo si } \bigcap_{j=0}^{k} U_{i_j} \neq \emptyset.$$

***Definición* 2.25.** *Sea $U = \{U_i\}_{i\in I}$ una cubierta abierta de (X, T). Se dice que U es una cubierta abierta buena si para cada subconjunto finito $J \subset I$, el conjunto $\bigcap_{j\in J} U_j$ es vacío o contráctil.*

Cuando se trabaja con una cubierta abierta buena, es posible relacionar al espacio topológico (X, T) con el complejo $C(U)$ mediante el siguiente teorema.

***Teorema* 2.1** (Jean-Daniel Boissonnat y Yvinec 2018)**.** *Sea $U = \{U_i\}_{i\in I}$ una cubierta abierta buena de un subconjunto $X \subseteq \mathbb{R}^d$, entonces X y $C(U)$ son homotópicamente equivalentes.*

El anterior teorema se conoce como teorema del nervio, el cual es muy importante para la topología computacional y la inferencia geométrica, ya que proporciona una forma de codificar el tipo de homotopía del espacio topológico X mediante un complejo simple que describe el patrón de intersección de una cubierta buena. En particular, cuando X es la unión (finita) de bolas (cerradas o abiertas) en $\mathbb{R}^d$, es homotópicamente equivalente al nervio de esta unión de bolas.

En muchas ocasiones pensar en el complejo simplicial desde el punto de vista geométrico nos permite trabajar con mayor intuición, pero antes es conveniente recordar algunas definiciones de espacios afines (Baer, 1966).

***Definición* 2.26.** *Sea $\mathbb{M} \subset \mathbb{R}^n$, diremos que $\mathbb{M}$ es un espacio afín si dados $x, y \in \mathbb{M}$, la recta que pasa por x e y está contenida en $\mathbb{M}$, esto es,*

$$tx + (1-t)y \in \mathbb{M}, \text{ para cada } t \in \mathbb{R}.$$

***Definición* 2.27.** *Sea $\mathbb{M} \subset \mathbb{R}^n$, diremos que $\mathbb{M}$ es un conjunto convexo si dados $x, y \in \mathbb{M}$ el segmento que une a x e y está contenido en $\mathbb{M}$, esto es,*

$$tx + (1-t)y \in \mathbb{M}, \text{ para todo } t \in [0, 1].$$

***Teorema* 2.2** (Baer 1966)**.** *Sea $\mathbb{M} \subset \mathbb{R}^n$ un espacio afín con $p_0, p_1, ..., p_r \in \mathbb{M}$ y sean $a_0, a_1, ..., a_r \in \mathbb{R}$ tales que $\sum_{i=0}^{r} a_i = 1$, entonces $\sum_{i=0}^{r} a_i p_i \in \mathbb{M}$.*

El resultado es cierto para conjuntos convexos con la hipótesis adicional $a_i \geq 0$ para cada i con $0 \leq i \leq r$.

***Definición* 2.28.** *Sea $\mathbb{M} \subset \mathbb{R}^n$ un conjunto afín con $p_0, p_1, ..., p_r \in \mathbb{M}$, decimos que $p_0, p_1, ..., p_r$ son afínamente independientes si los vectores $p_1 - p_0$, $p_2 - p_0, ..., p_r - p_0$ son linealmente independientes.*

***Definición* 2.29.** *Sean $p_0, p_1, ..., p_r$ puntos afínamente independientes, en un espacio afín. El conjunto de todas las combinaciones afines es llamado el espacio generado por $p_0, p_1, ..., p_r$, y es denotado por $Span(p_0, p_1, ..., p_r)$.*

Cada $p \in Span(p_0, p_1, ..., p_r)$ tiene una expresión única dada por una combinación afín

$$p = \sum_{i=0}^{r} a_i p_i,$$

donde el vector de coordenadas $(a_0, a_1, ..., a_r)$ es llamado vector de coordenadas baricéntricas con respecto al conjunto $\{p_0, p_1, ..., p_r\}$.

Con las anteriores definiciones, es posible definir a un simplejo geométrico como se muestra a continuación (Jean-Daniel Boissonnat y Yvinec, 2018).

***Definición* 2.30.** *Dados $p_0, p_1, ..., p_n$ puntos afínamente independientes, el subconjunto de $Span(p_0, p_1, ..., p_n)$ de todos los puntos con coordenadas baricéntricas positivas es llamado simplejo geométrico $n-$dimensional generado por $p_0, p_1, ..., p_n$. Se denota por $\Delta(p_0, p_1, ..., p_n)$ y por simplicidad se le suele llamar $n-$simplejo geométrico.*

Los simplejos geométricos se representan por gráficas completas y por ende conviene dar ejemplos una vez se haya dado la introducción a la teoría de gráficas de la Sección 3.1, sin embargo a continuación se muestra la definición de un complejo simplicial geométrico y algunas relaciones interesantes entre el complejo simplicial abstracto y su realización geométrica.

***Definición* 2.31.** *Un complejo simplicial geométrico es una colección finita K de simplejos geométricos que cumplen lo siguiente*

- *Si $\sigma_q \subseteq \sigma_r$ $(q \leq r)$ y $\sigma_r \in K$, entonces $\sigma_q \in K$.*
- *Para cualesquiera dos simplejos geométricos $\sigma_q, \sigma_r \in K$, si $\sigma_q \cap \sigma_r \neq \emptyset$, entonces $\sigma_q \cap \sigma_r$ es una cara común de σ_q y σ_r, y $\sigma_q \cap \sigma_r \in K$.*

***Definición* 2.32.** *El conjunto de puntos de los simplejos geométricos de un complejo simplicial geométrico K se denomina poliedro subyacente a K y lo denotaremos por $|K|$, es decir*

$$|K| = \bigcup_{\sigma \in K} \sigma \subset \mathbb{R}^n.$$

Los simplejos geométricos quedan determinados por sus vértices y, por lo tanto, se pueden ver como simplejos abstractos (simplemente considerando su conjunto de vértices). Más generalmente (Jean-Daniel Boissonnat y Yvinec, 2018), si K es un complejo simplicial geométrico en $\mathbb{R}^n$, entonces identificando los simplejos geométricos $\Delta(v_0, ..., v_k)$ con el conjunto finito $\{v_0, ..., v_k\}$, se obtiene un complejo simplicial abstracto. Recíprocamente, dado un simplejo abstracto, puede asociársele un simplejo geométrico en algún espacio euclidiano, eligiendo puntos afínmente independientes correspondientes a sus vértices y tomando su envolvente convexa. Por supuesto, esta elección no es única, pero lo es salvo homeomorfismo.

***Definición* 2.33.** *Dos complejos simpliciales abstractos Δ y Δ' con conjunto de vértices V y V' respectivamente, son isomorfos si existe una biyección $\Phi : V \to V'$ tal que $\{v_0, ..., v_k\} \in \Delta$ si, y sólo si, $\{\Phi(v_0), ..., \Phi(v_k)\} \in \Delta'$.*

La relación de isomorfismo entre dos complejos simpliciales abstractos induce un homeomorfismo entre sus realizaciones geométricas.

***Proposición* 2.2** (Blumberg 2020)**.** *Si dos complejos simpliciales geométricos K y K' son las realizaciones geométricas de dos complejos simpliciales abstractos isomorfos Δ y Δ', entonces $|K|$ y $|K'|$ son espacios topológicos homeomorfos. En particular, los espacios subyacentes de dos realizaciones geométricas cualesquiera de un complejo simplicial abstracto son homeomorfos.*

Una cosa a resaltar de lo anterior es que un complejo simplicial abstracto Δ tiene más de una posible realización geométrica K, sin embargo todas estas realizaciones serán homeomorfas y por lo tanto tendrán los mismos invariantes topológicos.

2.3. Homología simplicial

En esta sección se definirán los objetos que nos permitirán construir los grupos de homología asociados a un complejo simplicial, pero antes de comenzar con las definiciones es necesario mencionar que la teoría de homología simplicial está definida de forma general para grupos abelianos de coeficientes, sin embargo, en el contexto del análisis topológico de datos usualmente se trabaja con el campo de coeficientes $\mathbb{F}_2 := (\{0,1\}, +, \cdot)$ y por ende se utiliza la aritmética módulo 2 en las operaciones.

Dicho lo anterior, ya es posible comenzar a definir los objetos que son de interés en esta sección. El primer objeto que se definirá será el espacio de $d-$cadenas, definición que se muestra a continuación (Aktas Mehmet E. y El, 2019; Deo, 2018).

***Definición* 2.34.** *Sea Δ un complejo simplicial abstracto y $d \in \mathbb{N}$. El espacio de $d-$cadenas de Δ, denotado por $C_d(\Delta)$, es el conjunto de sumas formales finitas de $d-$simplejos de Δ con coeficientes en $\mathbb{F}_2$. Es decir, si $\{\sigma_1, \sigma_2, ..., \sigma_n\}$ es el conjunto de $d-$simplejos abstractos de Δ, entonces las $d-$cadenas de Δ se pueden escribir de la siguiente forma*

$$c = \sum_{i=1}^{n} t_i \sigma_i,$$

con $t_i \in \mathbb{F}_2$.

Ahora que se han definido a las $d-$cadenas o cadenas de dimensión d, es posible definir la frontera (algebraica) de un simplejo por medio del operador frontera tal y como se muestra a continuación (Munkres, 1984).

***Definición* 2.35.** *Sea $\sigma = \{v_0, v_1, \ldots, v_d\}$ un simplejo de dimensión d. Se define una función lineal $\partial_d : C_d(K) \to C_{d-1}(K)$ llamada el operador (o función) frontera de nivel d como se muestra a continuación*

$$\partial_d(\sigma) = \sum_{i=0}^{d} (-1)^i \{v_0, \ldots, \hat{v}_i, \ldots, v_d\},$$

donde $\hat{v}_i$ significa que el vértice v_i se omite.

Se tiene que aclarar que la anterior definición es valida para cualquier campo de coeficientes, donde los coeficientes de las cadenas son tomados de un campo diferente a $\mathbb{F}_2$. Como ya se mencionó anteriormente, nosotros estaremos trabajando en $\mathbb{F}_2$ y por ello en la anterior definición el operador frontera quedaría de la siguiente manera

$$\partial_d(\sigma) = \sum_{i=0}^{d} \{v_0, \ldots, \hat{v}_i, \ldots, v_d\}.$$

Con la definición del operador frontera y la frontera de un simplejo, es posible hablar de dos tipos de cadenas particularmente importantes para generar los grupos de homología, dichas cadenas son los ciclos y las fronteras del complejo simplicial (Aktas Mehmet E. y El, 2019; Deo, 2018).

***Definición* 2.36.** *Los ciclos de dimensión d de Δ están dados por el conjunto*

$$Z_d(\Delta) = \ker(\partial_d) = \{c \in C_d(\Delta); \partial_d(c) = 0\},$$

también es conocido como el espacio de $d-$ciclos de Δ.

***Definición* 2.37.** *Las fronteras de dimensión d de Δ están dadas por el conjunto*

$$B_d(\Delta) = \mathrm{Im}(\partial_{d+1}) = \{c \in C_d(\Delta); \exists t \in C_{d+1}(\Delta), \partial_{d+1}(t) = c\},$$

también es conocido como el espacio de $d-$fronteras de Δ.

Cabe señalar que los ciclos y las fronteras del complejo simplicial están estrechamente relacionados, ya que se cumplen las siguientes contenciones (Aktas Mehmet E.

y El, 2019; Deo, 2018)

$$B_d(\Delta) \subseteq Z_d(\Delta) \subseteq C_d(\Delta),$$

lo anterior es claro si se toma en cuenta que el operador frontera aplicado consecutivamente es igual a cero (Aktas Mehmet E. y El, 2019).

Para ejemplificar el proceso con el cual se pueden identificar a los ciclos y fronteras de un complejo simplicial, se tomará en cuenta el conjunto de vértices $\{a, b, c, d, e\}$ junto con el siguiente complejo simplicial abstracto $\{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{a, b\}, \{a, e\}, \{b, e\}, \{b, c\}, \{c, d\}, \{d, e\}, \{a, b, e\}\}$. Lo primero a notar en el complejo anterior es que tiene dimensión 2 ya que el simplejo de mayor dimensión en él es $\{a, b, e\}$, un 2-simplejo. Además, podemos notar que las caras que aparecen en el complejo son las siguientes

- 0-simplejos: $\{a\}, \{b\}, \{c\}, \{d\}, \{e\}$,
- 1-simplejos: $\{a, b\}, \{a, e\}, \{b, e\}, \{b, c\}, \{c, d\}, \{d, e\}$,
- 2-simplejos: $\{a, b, e\}$.

Si se toma en cuenta la 1-cadena formada por los simplejos $\{a, b\}, \{a, e\}, \{b, e\}$ tal y como se muestra a continuación

$$q = \{a, b\} + \{a, e\} + \{b, e\},$$

se puede verificar que q forma un ciclo de dimensión 1, ya que al aplicar el operador frontera se obtiene lo siguiente:

$$\begin{aligned}
\partial_1(q) &= \partial_1(\{a, b\}) + \partial_1(\{a, e\}) + \partial_1(\{b, e\}) \\
&= \{a\} + \{b\} + \{a\} + \{e\} + \{b\} + \{e\} \\
&= 2\{a\} + 2\{b\} + 2\{e\} \\
&= 0\{a\} + 0\{b\} + 0\{e\} \quad \text{(con la aritmética de módulo 2).}
\end{aligned}$$

También se puede notar que q es la frontera de $\{a, b, e\}$ y se puede verificar con lo siguiente:

$$\partial_2(\{a, b, e\}) = \{a, b\} + \{a, e\} + \{b, e\} = q.$$

Con lo anterior se puede ver cómo una frontera siempre es un ciclo, pero un ciclo no necesariamente es una frontera y prueba de ello es tomar la 1-cadena formada por $\{\{b, e\}, \{b, c\}, \{c, d\}, \{d, e\}\}$ como se muestra a continuación

$$q' = \{b, e\} + \{b, c\} + \{c, d\} + \{d, e\},$$

aplicando el operador frontera a q' encontramos lo siguiente:

$$\begin{aligned}
\partial_1(q') &= \partial_1(\{b, e\}) + \partial_1(\{b, c\}) + \partial_1(\{c, d\}) + \partial_1(\{d, e\}) \\
&= \{b\} + \{e\} + \{b\} + \{c\} + \{c\} + \{d\} + \{d\} + \{e\} \\
&= 2\{b\} + 2\{c\} + 2\{d\} + 2\{e\} \\
&= 0\{b\} + 0\{c\} + 0\{d\} + 0\{e\} \quad \text{(con la aritmética de módulo 2).}
\end{aligned}$$

probando que q' es un ciclo, pero no existe un $2-$simplejo σ en nuestro complejo simplicial tal que $\partial_2(\sigma) = q'$.

En este punto ya se tienen definidos todos los objetos necesarios para hablar acerca de los grupos de homología del complejo simplicial y cuya definición se muestra a continuación.

***Definición* 2.38.** *El $d-$ésimo grupo de homología del complejo simplicial Δ es el $\mathbb{F}_2-$espacio vectorial cociente*

$$H_d(\Delta) = \frac{Z_d(\Delta)}{B_d(\Delta)}.$$

Es necesario notar que en este espacio cociente sólo hay clases de equivalencia de $d-$ciclos que no son $d-$frontera de un $(d+1)-$simplejo, en otras palabras sólo están los ciclos que encierran vacío. También es conveniente mencionar que los grupos de

homología tienen una interpretación inmediata ya que en $H_0(K)$ se encuentran las componentes conexas, en $H_1(K)$ se encuentran los ciclos que rodean huecos bidimensionales, en $H_2(K)$ se encuentran los ciclos que rodean huecos tridimensionales y así sucesivamente (Aktas Mehmet E. y El, 2019).

Con lo anterior mencionado queda claro que la dimensión del espacio cociente $H_d(K)$ es igual a los números de Betti β_d cuya interpretación también es inmediata, ya que β_0 es el número de componentes conexas, β_1 es el número de huecos bidimensionales, β_2 es el número de huecos tridimensionales y así sucesivamente.

2.4. Homología persistente

Los complejos simpliciales con frecuencia tienen un ordenamiento específico de simplejos, el cual es factor importante para el estudio de homología persistente (Jean-Daniel Boissonnat y Yvinec, 2018).

***Definición* 2.39.** *Sea K un complejo simplicial. Una filtración $\mathcal{F}$ del complejo K es una colección de subcomplejos $K_0, K_1, ..., K_n$ tales que*

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

La anterior definición de filtración es un poco ambigua en el sentido de que pareciera que cada subcomplejo simplicial se podría crear de manera casi arbitraria, sólo se tienen que cumplir las contenciones. Generalmente se puede encontrar un orden natural en los subcomplejos simpliciales de la filtración y para poder dar un ordenamiento correcto, se utilizan pesos sobre cada simplejo en el complejo simplicial original (Zomorodian, 2010).

***Definición* 2.40.** *Sean K un complejo simplicial y $\epsilon \in \mathbb{R}$. Un complejo simplicial ponderado es una tupla (K, ν), donde $\nu : K \to \mathbb{R}$ es una función de pesos discreta tal*

que $K_\epsilon = \{\sigma \in K; \nu(\sigma) \leq \epsilon\}$ *genera una filtración.*

La anterior definición nos permite crear filtraciones con subcomplejos simpliciales ordenados mediante una función de pesos. Para dar un par de ejemplos, se tomarán en cuenta los complejos simpliciales de Vietoris-Rips y de Čech (ver Definiciones 2.22 y 2.23).

En la practica, se calcula el complejo de Vietoris-Rips $V_\epsilon(X, d_x)$ para un parámetro de proximidad máximo $\hat{\epsilon} \in \mathbb{R}$ y cada subcomplejo simplicial en la filtración se genera para algún parámetro $\epsilon \leq \hat{\epsilon}$ mediante la función de pesos de Vietoris-Rips $\nu : VR_{\hat{\epsilon}}(X, d_X) \to \mathbb{R}$ definida sobre los simplejos $\sigma \in VR_{\hat{\epsilon}}(X, d_X)$, como se muestra a continuación (Zomorodian, 2010):

$$\nu(\sigma) = \begin{cases} 0, & \text{si } dim(\sigma) \leq 0, \\ d_X(u, v), & \text{si } \sigma = \{u, v\}, \\ \text{máx}_{\tau \subset \sigma}\, \nu(\tau), & \text{en otro caso.} \end{cases}$$

Al igual que en el caso anterior, en la práctica se genera el complejo simplicial de Čech para un $\hat{\epsilon}$ máximo y cada subcomplejo simplicial en la filtración se genera para algún parámetro $\epsilon \leq \hat{\epsilon}$. En este caso la función de pesos de Čech $\nu : C_{\hat{\epsilon}}(X) \to \mathbb{R}$ le asigna a cada simplejo $\sigma \in C_{\hat{\epsilon}}(X)$ el valor del radio en el que se originó, así como se muestra a continuación (Espinoza, Hernández-Amador, Hernandez, y Ramonetti-Valencia, 2019):

$$\nu(\sigma) = \text{ínf} \left\{ r \geq 0; \bigcap_{i=0}^{n} B_c(v_i, r) \neq \emptyset \right\}, \text{ si } \sigma = \{v_0, v_1, ..., v_n\}.$$

La relevancia de la función de pesos en el estudio de la homología persistente se puede observar en la Definición 2.40, en donde la creación de la filtración ya no se hace de manera arbitraria como lo podría sugerir la Definición 2.39, ahora los subcomplejos simpliciales se pueden ordenar de acuerdo a una función de pesos. Como se ha podido observar con los anteriores dos ejemplos, existen funciones de pesos que se asocian naturalmente a ciertos complejos simpliciales, sin embargo dicha asociación

no es ley, siempre se puede generar una filtración con funciones de pesos que utilizan diferentes criterios a la hora de asignar los pesos a cada simplejo.

Teniendo ya las herramientas que nos permitirán generar una filtración para un complejo simplicial, es hora de hablar sobre la relevancia que tiene la filtración en el estudio de la homología simplicial.

Si se trabaja con una colección V de puntos en $\mathbb{R}^n$ y se trata de obtener los grupos de homología de esta nube de puntos tomando como complejo simplicial a V, es decir sólo se tienen $0-$simplejos, entonces nos encontraríamos con tantas componentes conexas (β_0) como puntos en el conjunto V. Al trabajar con un complejo simplicial con una estructura pobre como V, jamás se encontrarán características topológicas que sean relevantes para el estudio de la nube de puntos y tiene sentido si se piensa en que al estudiar sólo $0-$simplejos sería equivalente a querer estudiar cada punto por separado en lugar de estudiarlos en conjunto.

Con lo anterior es posible observar que en realidad para obtener características topológicas interesantes de una nube de puntos es necesario trabajar con un complejo simplicial que tenga una estructura rica. Para crear estructuras que relacionen a ciertos puntos en la nube, se pueden utilizar construcciones como las de Čech o la de Vietoris-Rips, donde se construyen simplejos de diferentes dimensiones de acuerdo a un parámetro de proximidad. A manera de ejemplo tomemos en cuenta de nuevo el conjunto de puntos V, pero en esta ocasión se construirá el complejo simplicial de Vietoris-Rips utilizando algún parámetro de proximidad que sea lo suficientemente grande como para crear algunos simplejos de dimensión mayor o igual que uno. En esta ocasión se tendrán menos componente conexas que el número de elementos en V a causa de que algunas de estas componentes conexas ya contarán con más de un punto, además ya se tienen algunos simplejos de dimensión mayor o igual que uno y por lo tanto es posible que aparezcan algunos hoyos $(\beta_1, \beta_2, ...)$.

Para entender la relevancia de la filtración en el análisis topológico de datos, pensemos de nueva cuenta en el conjunto de puntos V y en el complejo simplicial de Vietoris-Rips que se obtendría con el parámetro de proximidad ϵ. Si nuestra filtración consistiera en una colección de complejos simpliciales de Vietoris-Rips con parámetros de proximidad $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ con $\epsilon_i < \epsilon_j$ para $i < j$, entonces se pueden obtener las características topológicas de cada complejo y se podría observar cómo las características evolucionan según el parámetro de proximidad aumenta, además de poder reconocer aquellas que persisten a pesar del cambio en el parámetro de proximidad.

Con lo anterior, se puede decir que la idea intuitiva del estudio de la homología persistente es ver cómo los grupos de homología evolucionan de acuerdo a cierta filtración del complejo simplicial y para observar dicha evolución típicamente se utiliza un código de barras o un diagrama de persistencia (Aktas Mehmet E. y El, 2019).

- Código de barras: consiste en un conjunto de intervalos $[a, b)$, con $a \in \mathbb{R}_{\geq 0}$ y $b \in \mathbb{R}_{\geq 0} \cup \{\infty\}$, que representa el tiempo en el que "nace" y "muere" una clase de homología persistente.

- Diagrama de persistencia: consiste en los tiempos de nacimiento y muerte de las clases de homología persistente como puntos (nacimiento, muerte) en el plano extendido $\mathbb{R}^2$.

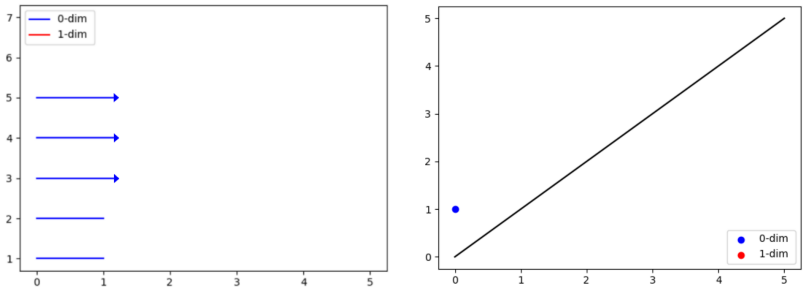
La mejor manera de mostrar cómo se resume la evolución de los grupos de homología en el código de barra y el diagrama de persistencia es mediante un ejemplo, por ello se volverá a tomar en cuenta el complejo simplicial $\Delta = \{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{a, b\}, \{a, e\}, \{b, e\}, \{b, c\}, \{c, d\}, \{d, e\}, \{a, b, e\}\}$ construido con los vértices $V = \{a, b, c, d, e\}$ y además se utilizará la siguiente filtración

- $\Delta_0 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}\}$

- $\Delta_1 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{a, b\}, \{a, e\}\}$

- $\Delta_2 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{a, b\}, \{a, e\}, \{b, e\}, \{c, d\}\}$
- $\Delta_3 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{a, b\}, \{a, e\}, \{b, e\}, \{c, d\}, \{b, c\}, \{d, e\}\}$
- $\Delta_4 = \Delta$

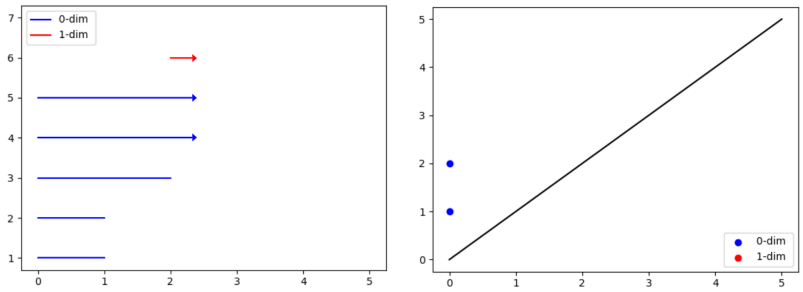
En Δ_0 se tienen sólo los $0-$simplejos de Δ, por lo que el número de componentes conexas es $\beta_0 = 5$, además aún no se tiene ningún hoyo.



(a) Código de barras. **(b)** Diagrama de persistencia.

Figura 2.3: Evolución de grupos de homología de Δ_0 a Δ_1.

En Δ_1 aparecen dos $1-$simplejos $\{a, b\}$ y $\{a, e\}$, esto hace que los vértices a, b, e formen una sola componente conexa y con ello el número de componentes conexas disminuye a $\beta_0 = 3$ (ver Figura 2.3). Notemos que a pesar de que ya se puede formar la primer cadena de dimensión 1, esta no produce ningún ciclo y por ende no es posible tener algún hoyo.



(a) Código de barras. **(b)** Diagrama de persistencia.

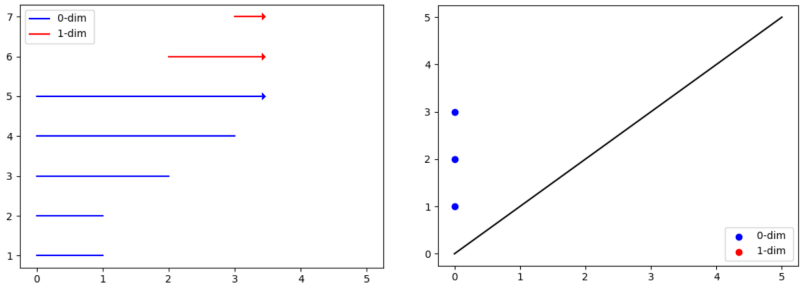
Figura 2.4: Evolución de grupos de homología de Δ_0 a Δ_2.

En Δ_2 aparecen otros dos $1-$simplejos $\{b, e\}$ y $\{c, d\}$, particularmente el simplejo $\{c, d\}$ provoca que los vértices c y d sean una sola componente conexa, de manera que

se reduce el número de componentes conexas a $\beta_0 = 2$. Además se tiene que hacer notar que la cadena

$$q = \{a, b\} + \{a, e\} + \{b, e\}$$

forma un ciclo y como en Δ_2 no hay ningún 2−simplejo, q no puede ser una frontera y genera el primer hueco de dimensión 1, es decir, se tiene que $\beta_1 = 1$ (ver Figura 2.4).



(a) Código de barras. **(b)** Diagrama de persistencia.

Figura 2.5: Evolución de grupos de homología de Δ_0 a Δ_3.

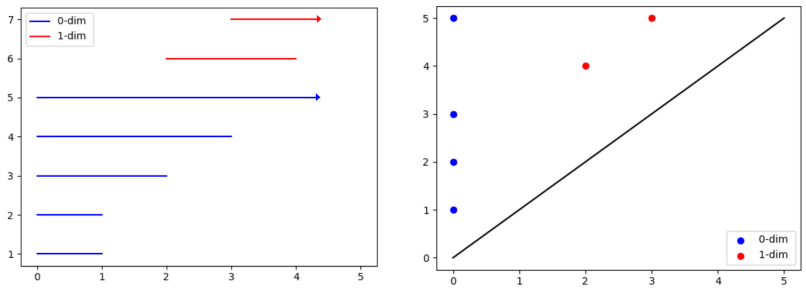
En Δ_3 aparecen los dos últimos 1−simplejos $\{b, c\}$ y $\{d, e\}$, estos hacen que todos los vértices queden unidos y por ende sólo se tiene una componente conexa $\beta_0 = 1$. También se puede notar que la cadena

$$q' = \{b, c\} + \{c, d\} + \{d, e\} + \{b, e\}$$

forma un ciclo y como en Δ_3 tampoco hay 2−simplejos, q' forma el segundo hueco de dimensión uno, es decir, $\beta_1 = 2$ (ver Figura 2.5).

Por último, en Δ_4 el número de componentes conexas sigue siendo $\beta_0 = 1$ y además aparece el único 2−simplejo ($\{a, b, e\}$) que hay en Δ, provocando que q sea ahora una frontera y con ello el número de huecos disminuye a $\beta_1 = 1$ (ver Figura 2.6).

Notemos que en el diagrama de persistencia de la Figura 2.6 aparecen dos puntos con coordenadas (0, 5) y (3, 5), estos corresponden a las características del complejo simplicial original Δ y en el código de barras de la Figura 2.6 se muestran como las



(a) Código de barras. **(b)** Diagrama de persistencia.

Figura 2.6: Evolución de grupos de homología de Δ_0 a Δ_4.

lineas que terminan con una flecha, en muchas ocasiones se suelen representar como lineas que terminan después del valor máximo que el parámetro de la filtración puede tomar.

Como ya se ha mencionado anteriormente, el análisis de la homología persistente depende de la filtración a utilizar, por ende es posible obtener diferentes resultados de la homología persistente si se utilizan diferentes filtraciones para el mismo complejo simplicial. Lo anterior genera la necesidad de saber qué tan distantes pueden ser los resultados obtenidos al generarse a partir de diferentes filtraciones y con la intención de obtener esta información, en el análisis topológico de datos comúnmente se utiliza la métrica de Wasserstein y la métrica de cuello de botella para medir la distancias entre los diagramas de persistencia generados por cada filtración.

***Definición* 2.41.** *Sean P y Q dos diagramas de persistencia. La distancia de cuello de botella entre P y Q se define como*

$$d_B(P,Q) = \inf_{\gamma} \sup_{x \in P} ||x - \gamma(x)||_\infty,$$

donde γ varía sobre todas las biyecciones entre P y Q.

En otras palabras, la distancia de cuello de botella mide la distancia entre dos diagramas de persistencia P y Q con base en la distancia máxima entre dos puntos en una biyección de P a Q. De manera que la distancia de cuello de botella da como resultado la distancia entre el mayor valor atípico, en lugar de la distancia entre todos

los pares de puntos (Aktas Mehmet E. y El, 2019).

***Definición* 2.42.** *Sean P y Q dos diagramas de persistencia. La $p-$ésima distancia de Wasserstein entre P y Q se define como*

$$d_{W_p}(P,Q) = \inf_{\gamma} \left(\sum_{x \in P} ||x - \gamma(x)||_p \right)^{1/p},$$

donde γ varía sobre todas las biyecciones entre P y Q.

En otras palabras, la distancia de Wasserstein considera la distancia total entre el par de puntos emparejados, por lo que cuantifica de manera general la similitud entre dos diagramas de persistencia (Aktas Mehmet E. y El, 2019).

Hasta ahora se han presentado ejemplos relacionados a simplejos y complejos simpliciales puramente abstractos, lo anterior se debe a que los complejos simpliciles geométricos están íntimamente relacionados a la teoría de gráficas y resulta conveniente dar una pequeña introducción a esta para después mostrar cómo se aplica la homología simplicial y la homología persistente a las realizaciones geométricas de los complejos simpliciales.

CAPÍTULO 3

Análisis topológico de datos en redes

En este capítulo se busca introducir el análisis topológico de datos en redes, para ello es conveniente dar una breve introducción a la teoría de gráficas y más específicamente a gráficas no dirigidas.

3.1. Redes no dirigidas

Una red se construye a partir de vértices y aristas, en donde los vértices representan a los objetos cuya interacción se busca estudiar, mientras que las aristas representan la interacción entre los objetos. Como ejemplo pensemos en una red de ciudades, es decir, si se buscara estudiar la interacción entre las diferentes ciudades de un país, se podría representar a cada ciudad como un vértice y si dos ciudades se conectan directamente por carretera, entonces los vértices que representan estas dos ciudades estarían conectados por una arista.

La forma más sencilla de entender y formalizar a las redes es por medio de la teoría de gráficas, en donde es posible distinguir entre dos tipos de gráficas: las dirigidas, que

llamaremos digráficas y las no dirigidas, que llamaremos simplemente gráficas. En las digráficas se suelen representar objetos cuya interacción no es simétrica o recíproca, por otro lado en las gráficas todas las interacciones son simétricas.

Como el título de la sección lo anticipa, se trabajará con gráficas y por ello se presentarán las definiciones correspondientes. Muchas de las definiciones que se presentarán también se pueden extender para las digráficas.

Para comenzar con la formalidad, se presenta el concepto de gráfica (West, 2001).

***Definición* 3.1.** *Una gráfica G es una pareja que consiste de un conjunto finito no vacío de vértices $V(G)$, también llamados nodos, y de un conjunto de aristas $E(G) \subset V \times V$. Una arista $e = (u, v) \in E(G)$ se dice que une a los vértices $u, v \in V(G)$, también conocidos como los extremos de la arista.*

La anterior definición nos permite asociar más de una arista a un par de vértices y también permite que los extremos de una arista sean un mismo vértice (West, 2001).

***Definición* 3.2.** *Un bucle es una arista cuyos extremos son iguales. Aristas múltiples son aristas que tienen el mismo par de extremos.*

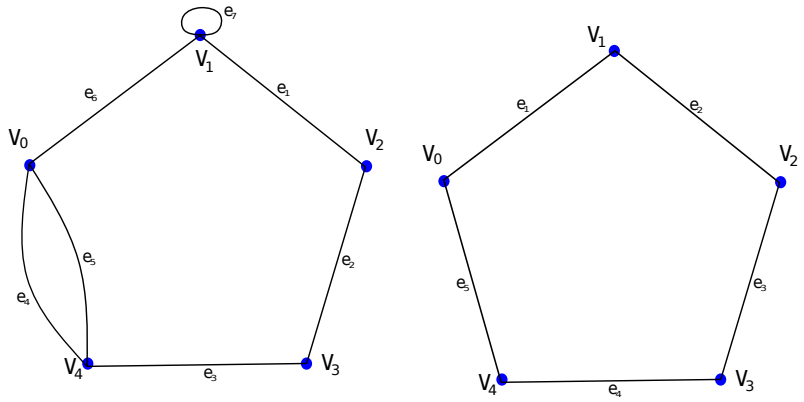
Dicho lo anterior, es posible definir una gráfica simple como se muestra a continuación (West, 2001).

***Definición* 3.3.** *Una gráfica simple es una gráfica que no tiene bucles o aristas múltiples.*

Se puede caracterizar a una gráfica simple por su conjunto de vértices V y su conjunto de aristas E, tratando al conjunto de aristas como un conjunto de pares de vértices desordenados y escribiendo $e = (u, v)$ para una arista $e \in E$ cuyos extremos son los vértices u y v. Cuando u y v son los extremos un una arista se dice que son adyacentes y se denota por $u \longleftrightarrow v$.

***Definición* 3.4.** *Dada una gráfica, su orden corresponde al número de vértices, $|V(G)|$, y su tamaño indica el número de aristas $|E(G)|$ que la componen.*

Para ilustrar la definición de una gráfica, bucle y de aristas múltiples, consideremos el conjunto de vértices $V = \{v_0, v_1, v_2, v_3, v_4\}$ junto con el conjunto de aristas $E = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7\}$ como en la Figura 3.1a. La gráfica que se forma tiene un bucle dado por e_7 y además los vértices v_0 y v_4 tienen aristas múltiples: e_4 y e_5. Para ejemplificar la definición de una gráfica simple se puede tomar en cuenta el mismo conjunto de vértices V y el conjunto de aristas $E' = \{e_1, e_2, e_3, e_4, e_5\}$ como en la Figura 3.1b, en donde se puede observar que no hay bucles y los vértices no tienen aristas múltiples.



(a) Gráfica con un bucle y aristas múltiples.

(b) Gráfica simple.

Figura 3.1: Ejemplos de gráficas.

Con el ejemplo anterior, ya es posible distinguir a una gráfica simple y por ende es conveniente definir dos conceptos que tendrán gran relevancia más adelante (West, 2001).

***Definición* 3.5.** *El complemento $\bar{G}$ de una gráfica simple G es la gráfica simple con conjunto de vértices $V(G)$ tal que $(u, v) \in E(\bar{G})$ si y sólo si $(u, v) \notin E(G)$.*

***Definición* 3.6.** *Una subgráfica de una gráfica G es una gráfica H tal que $V(H) \subseteq V(G)$ y $E(H) \subseteq E(G)$.*

En las siguientes secciones del capítulo se utilizará una particular clase de subgráficas, la cual se define a continuación (Aktas Mehmet E. y El, 2019).

***Definición* 3.7.** *Dada un gráfica $G = \{V, E\}$, se dice que G tiene un clique de tamaño n ($n-$clique) si existe una subgráfica $G' = \{V', E'\}$ de G con $|V'| = n$, y todos sus vértices conectados entre sí.*

Como ejemplo a la anterior definición, observemos la Figura 3.2 en donde se tiene una gráfica con vértices $V = \{v_0, v_1, v_2, v_3, v_4\}$, en la cual está señalada en rojo una subgráfica que es un clique de tamaño $n = 3$.

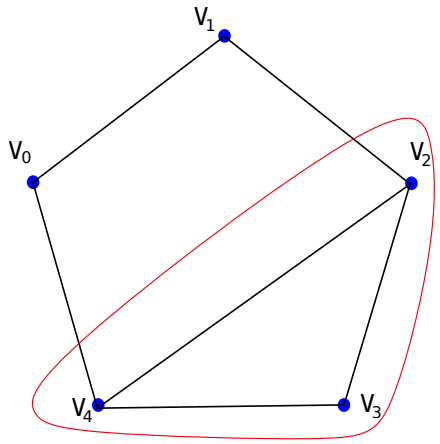


Figura 3.2: 3−clique.

Lo anterior nos introduce a definir las gráficas completas, que dicho de manera simple, son un clique.

***Definición* 3.8.** *Un gráfica simple es completa si existen aristas uniendo todos los pares posibles de vértices. Es decir, todo par de vértices (v_i, v_j), con $i \neq j$, debe tener una arista que los une. El conjunto de las gráficas completas es denominado $\mathbb{K}$, siendo $\mathbb{K}_n$ la gráfica completa de n vértices.*

Como ejemplo a la anterior definición, es posible observar la Figura 3.3, en donde se tiene una gráfica completa con cinco vértices.

Muchos de los conceptos en la teoría de gráficas se refieren a varias formas en las que uno se puede "desplazar" en una gráfica. En particular, si se piensa en los vértices como ubicaciones y en las aristas como caminos entre ciertos pares de ubicaciones,

entonces se puede considerar que la gráfica modela alguna comunidad y existen varias formas en las que uno se puede desplazar a diferentes puntos de esta.

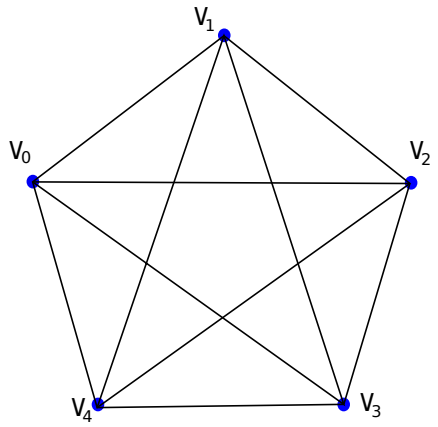


Figura 3.3: Gráfica completa $\mathbb{K}_5$.

***Definición* 3.9.** *Una caminata W de un vértice u a un vértice v en G, es una sucesión de vértices en G que comienza con u y termina en v de manera que los vértices consecutivos en la sucesión son adyacentes, es decir, podemos expresar W como*

$$W = (u = v_0, v_1, v_2, \ldots, v_k = v),$$

donde $k \geq 1$ y v_i, v_{i+1} son adyacentes para $i = 0, 1, 2, \ldots, k-1$. Cada vértice v_i $(0 \leq i \leq k)$ y cada arista $e_i = (v_i, v_{i+1})$ $(0 \leq i \leq k-1)$ se dice que se encuentra o pertenece a W.

Observe que la definición de caminata W no requiere que los vértices enumerados sean distintos, de hecho, ni siquiera u y v tienen que ser distintos. Sin embargo, los vértices consecutivos en W tienen que ser distintos, ya que son adyacentes y la gráfica es simple.

***Definición* 3.10.** *Sea W una caminata de u a v. Si $u = v$ se dice que la caminata es cerrada, si $u \neq v$ entonces la caminata es abierta. Al número de aristas en la caminata se le conoce como la longitud de la caminata.*

***Definición* 3.11.** *Sea W una caminata de u a v, si en la caminata no hay vértices repetidos, entonces W es un camino de u a v.*

Si bien es cierto que se puede listar los vértices y las aristas con sus respectivos extremos, existen otras representaciones muy útiles (Aktas Mehmet E. y El, 2019).

***Definición* 3.12.** *Sea G una gráfica sin bucles, con vértices $V(G) = \{v_1, v_2, ..., v_n\}$ y aristas $E(G) = \{e_1, e_2, ..., e_m\}$. La matriz de adyacencia de G, $A(G)$, es la matriz de tamaño $n \times n$ en donde las entradas $a_{i,j}$ son el número de aristas en G con extremos $\{v_i, v_j\}$.*

Notemos que la definición anterior no excluye a gráficas cuyos vértices tengan aristas múltiples, es por ello que los valores de las entradas de la matriz de adyacencia pueden ser mayor que uno, sin embargo es claro que si se trabaja con gráficas simples entonces las entradas son a lo más uno, los elementos de su diagonal son cero y la matriz es simétrica, es decir $a_{i,j} = a_{j,i}$.

***Definición* 3.13.** *Si el vértice v es un punto extremo de la arista e, entonces v y e inciden. El grado de un vértice v (en una gráfica sin bucles) es el número de aristas que inciden con v.*

Una manera equivalente de representar la estructura de un gráfica es mediante la matriz de incidencia.

***Definición* 3.14.** *Sea G una gráfica sin bucles, con vértices $V(G) = \{v_1, v_2, ..., v_n\}$ y aristas $E(G) = \{e_1, e_2, ..., e_m\}$. La matriz de incidencia $M(G)$ es la matriz de tamaño $n \times m$ en donde las entradas $m_{i,j}$ son 1 si v_i es un vértice extremo de e_j y 0 en otro caso.*

De las anteriores definiciones se puede notar que es posible obtener el grado de un vértice v de las matrices $A(G)$ y $M(G)$, ya que en ambas matrices el grado de v es igual a la suma de las entradas de la fila asociada a v.

Para dar un sencillo ejemplo de las matrices $A(G)$ y $M(G)$ se tomará a la gráfica con vértices $V = \{v_1, v_2, v_3, v_4\}$ y con aristas $E = \{e_1, e_2, e_3, e_4, e_5\}$ cuya asignación se puede observar en la Figura 3.4, en ella se puede observar que no se tienen bucles, asegurando que los elementos en la diagonal de la matriz de adyacencia $A(G)$ sean

cero. Sin embargo, es posible observar que el par de vértices $\{v_2, v_4\}$ tiene aristas múltiples y por ende se puede esperar que un par de elementos en la matriz de adyacencia sean mayores que 1. La información de la anterior gráfica quedaría resumida en las matrices $A(G)$ y $M(G)$ tal y como se muestra a continuación:

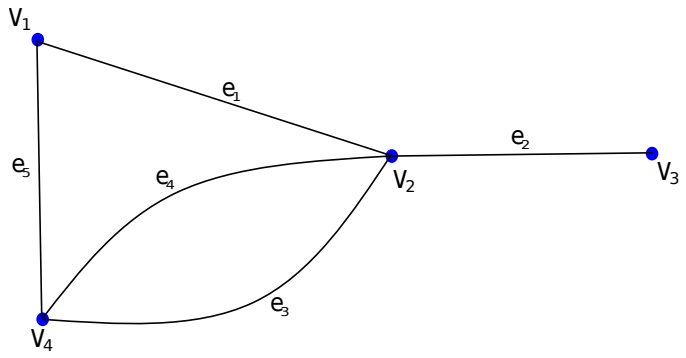


Figura 3.4: Gráfica de muestra para las matrices A y M.

$$A(G) = \begin{array}{c} \begin{matrix} v_1 & v_2 & v_3 & v_4 \end{matrix} \\ \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 2 \\ 0 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \end{pmatrix} \end{array} \begin{matrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{matrix} \qquad M(G) = \begin{array}{c} \begin{matrix} e_1 & e_2 & e_3 & e_4 & e_5 \end{matrix} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix} \end{array} \begin{matrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{matrix}$$

En muchas aplicaciones las aristas de una gráfica tienen un valor numérico asociado (por lo general no negativo), llamado peso. En las aplicaciones el peso puede ser una medida de la longitud de una ruta, la capacidad de una línea, la energía requerida para moverse entre ubicaciones a lo largo de una ruta, etc.

***Definición* 3.15.** *Sea $G = (V, E)$ una gráfica. Si a cada arista $e \in E$ se le asocia un número real $w(e)$ llamado peso, entonces a G con estos pesos en sus aristas se le conoce como gráfica ponderada.*

Para el caso de gráficas simples ponderadas, también es posible definir la respectiva matriz de adyacencia en la cual se resume por completo la información de la gráfica.

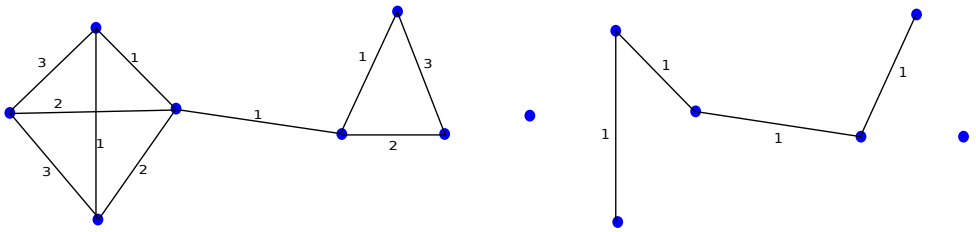
***Definición* 3.16.** *Sea $G = (V, E)$ una gráfica simple ponderada, con función de pesos w definida sobre cada arista de G. Si se denota a (v_i, v_j) como la arista cuyos extremos son $v_i, v_j \in V$, entonces los elementos de la matriz de adyacencia de la gráfica ponderada G están dados por:*

$$a_{i,j} = a_{j,i} = \begin{cases} w(v_i, v_j) & si\ (v_i, v_j) \in E, \\ 0 & si\ (v_i, v_j) \notin E. \end{cases}$$

Cabe señalar que en este trabajo se emplean mayormente gráficas simples con cierta ponderación, por lo que la matriz de adyacencia cumple un papel bastante relevante a la hora de construir las gráficas que se mostrarán a lo largo del Capítulo 4.

Llegados a este punto, es necesario mostrar dos subgráficas que permitirán definir de mejor manera algunas filtraciones que se mostrarán en la Sección 3.3.

***Definición* 3.17.** *Sea $G = (V, E)$ una gráfica ponderada no dirigida con la función de pesos $w : E \to \mathbb{R}$. Para cualquier $\epsilon \in \mathbb{R}$, se define a $G_\epsilon = (V_\epsilon, E_\epsilon) \subset G$ como la subgráfica de G donde $V_\epsilon = V$ y sus aristas $E_\epsilon \subseteq E$ solo incluye las aristas cuyos pesos son iguales o menores que ϵ, es decir, $E_\epsilon = \{e \in E; w(e) \leq \epsilon\}$.*



(a) Gráfica ponderada. **(b)** G_1.

Figura 3.5: Ejemplo de subgráfica G_ϵ.

Como ejemplo de la anterior definición, tomemos en cuenta a la gráfica ponderada que se muestra en la Figura 3.5a, en donde se puede observar que los pesos van de 1 a 3 y al extraer la subgráfica G_ϵ asociada a $\epsilon = 1$ obtenemos la gráfica que se muestra en la Figura 3.5b.

De lo anterior es posible generar una subgráfica, que en combinación con G_ϵ, permita obtener a la grafica G original.

***Definición* 3.18.** *Sea $G = (V, E)$ una gráfica ponderada no dirigida con la función de pesos $w : E \to \mathbb{R}$. Para cualquier $\epsilon \in \mathbb{R}$, se define a $\hat{G}_\epsilon = (\hat{V}_\epsilon, \hat{E}_\epsilon) \subset G$ como la subgráfica de G donde $\hat{V}_\epsilon = V$ y sus aristas $\hat{E}_\epsilon \subseteq E$ solo incluye las aristas cuyos pesos son mayores que ϵ, es decir, $\hat{E}_\epsilon = \{e \in E; w(e) > \epsilon\}$.*

La anterior definición nos permite generar una clase diferente de complemento a lo mostrado en la Definición 3.5. En la Figura 3.6 se muestra la subgráfica $\hat{G}_1$ de la gráfica G mostrada en la Figura 3.5a.

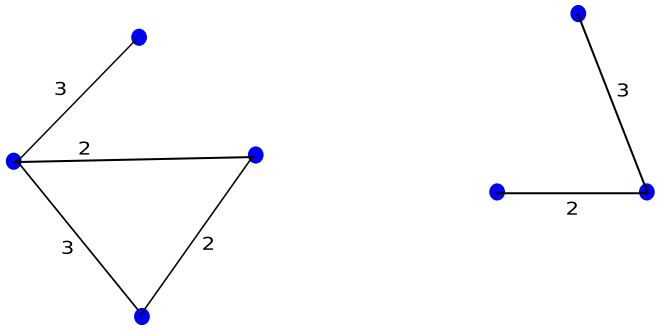


Figura 3.6: $\hat{G}_1$.

De lo anterior es posible observar que para una gráfica en particular G, se cumple que $\hat{G}_\epsilon \cup G_\epsilon$.

Es necesario resaltar que en las graficas ponderadas los vértices no tienen ningún peso asignado, sin embargo, en el análisis topológico de datos se les asigna un peso que depende de la filtración que se utilice para estudiar la evolucion de los grupos de homología.

3.2. Complejos simpliciales y gráficas

Por la manera en que se construyen los complejos simpliciales abstractos, es posible relacionarlos con las gráficas como se describe a continuación.

***Definición* 3.19.** *Un complejo simplicial abstracto cuyos simplejos tienen dimensión menor o igual que uno, se denomina gráfica abstracta.*

No obstante, se emplearán complejos simpliciales geométricos ya que cada $q-$simplejo geométrico se puede representar como un poliedro en el espacio $q-$dimensional, es decir, un $q-$simplejo geométrico se representa gráficamente como una gráfica completa con $q+1$ vértices, por lo tanto un $0-$simplejo geométrico es un punto, un $1-$simplejo geométrico es una línea, un $2-$simplejo geométrico es un triángulo y así sucesivamente (Maletić, Rajkovic, y Vasiljevic, 2008). Gracias a lo anterior, es posible identificar de mejor manera las características topológicas asociadas a una gráfica.

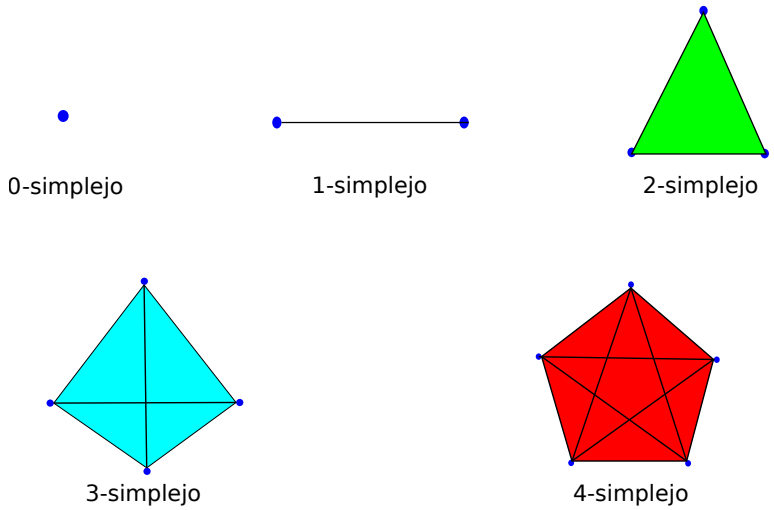


Figura 3.7: Representación gráfica de simplejos geométricos.

Si es posible representar a un simplejo geométrico como una gráfica completa, entonces las caras de dicho simplejo se pueden representar como las respectivas subgráficas completas. Para ejemplificar lo anterior, tomemos en cuenta al $3-$simplejo de la Figura 3.7 y cuyas caras se muestran en la Figura 3.8.

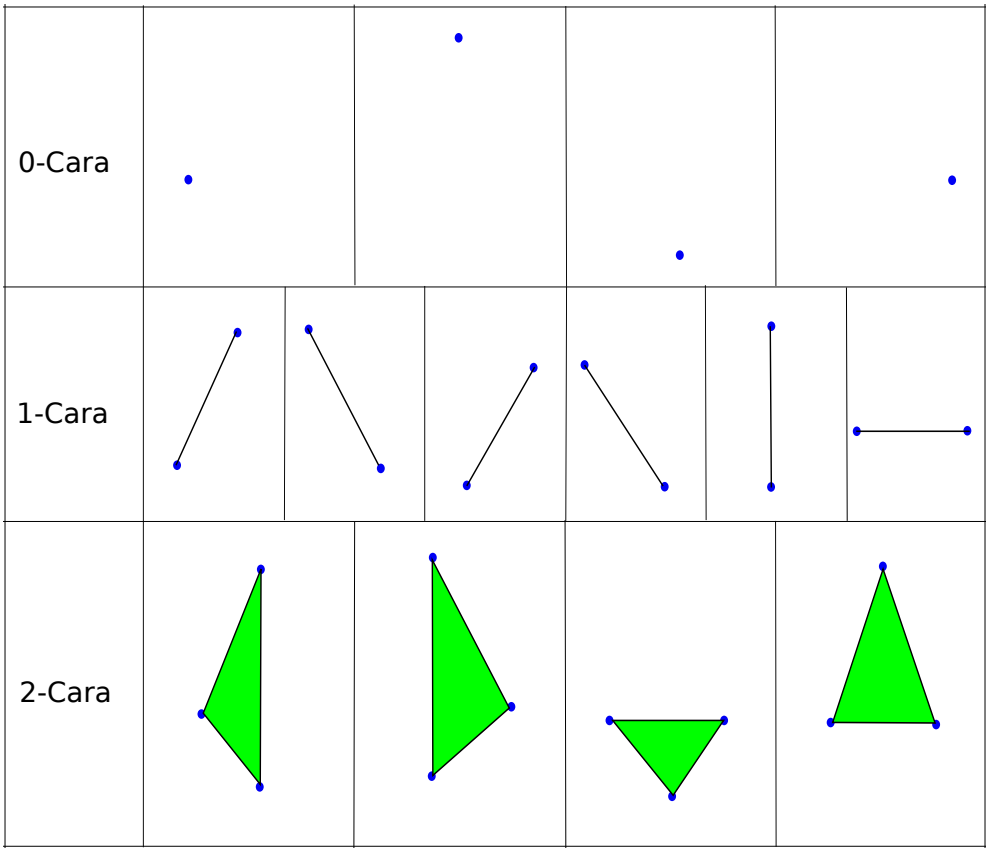
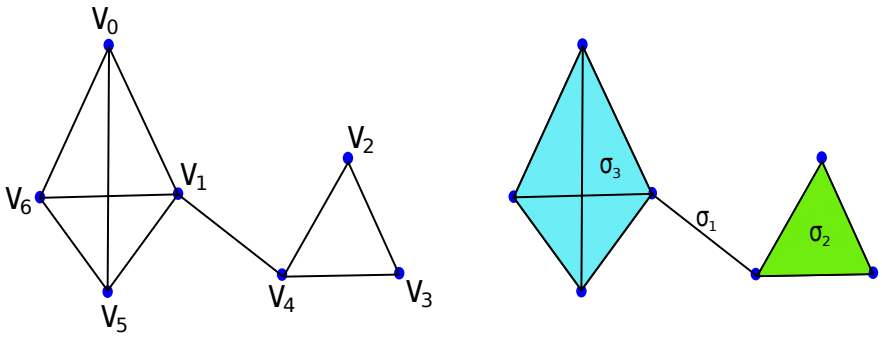


Figura 3.8: Caras de un 3−simplejo.

Existen varias maneras de utilizar la teoría de gráficas para poder crear un complejo simplicial, dentro de ellas se podrían tomar en consideración el grado de los vértices, los vecinos de los vértices, los cliques en la gráfica y más, por ello es necesario mencionar que en este trabajo sólo se toma en consideración los complejos cliques, cuya definición se muestra a continuación (Aktas Mehmet E. y El, 2019).

***Definición* 3.20.** *El complejo clique $Cl(G)$ de una gráfica simple no dirigida $G = (V, E)$, es un complejo simplicial geométrico donde los vértices de G son sus vértices y cada $k-$clique en G corresponde a un $(k-1)-$simplejo en $Cl(G)$.*

Como ejemplo de un complejo clique se tomará en cuenta la Figura 3.9a, en la cual se puede apreciar una gráfica simple con algunos cliques y en la Figura 3.9b es posible observar cómo cada clique forma un simplejo geométrico, en donde σ_q representa un simplejo de dimensión q.



(a) G. **(b)** $Cl(G)$.

Figura 3.9: Ejemplo de complejo clique.

Las definiciones mostradas en la Sección 2.3 también son válidas para los complejos simpliciales geométricos, por lo tanto es posible aplicar todas las herramientas de la homología simplicial para la construcción de los grupos de homología de un complejo que se construye a partir de una gráfica.

3.3. Filtraciones para complejos simpliciales geométricos

Ya hemos visto que existen un par de filtraciones que son clásicas en el análisis topológico de datos, sin embargo la aplicación a redes también ha dejado ciertas filtraciones que ya son clásicas y a continuación se muestran algunas de ellas (Aktas Mehmet E. y El, 2019).

***Definición* 3.21.** *Sea $G = (V, E)$ una gráfica ponderada no dirigida con función de pesos $w : E \to \mathbb{R}$ definida sobre las aristas de G. Para cualquier $\epsilon \in \mathbb{R}$, se define al complejo de Vietoris-Rips de G como el complejo clique de la subgráfica G_ϵ, es decir $Cl(G_\epsilon)$, y la filtración de Vietoris-Rips está definida como se muestra a continuación,*

$$\{Cl(G_\epsilon) \hookrightarrow Cl(G_{\epsilon'})\}_{0 \leq \epsilon \leq \epsilon'}.$$

En la anterior definición, $Cl(G_\epsilon) \hookrightarrow Cl(G_{\epsilon'})$ es la función inclusión de $Cl(G_\epsilon)$ a

$Cl(G_{\epsilon'})$. Esto nos indica que conforme ϵ incrementan de valor, los respectivos complejos cliques siempre cumplirán que $Cl(G_\epsilon) \subseteq Cl(G_{\epsilon'})$.

Notemos que la filtración de Vietoris-Rips en redes se calcula para un valor máximo $\epsilon' \in \mathbb{R}$ y los subcomplejos simpliciales geométricos están relacionados a valores de $\epsilon \leq \epsilon'$ de manera similar a lo mostrado en la Sección 2.4, sin embargo, la asignación del peso a cada $n-$simplejo geométrico es muy diferente.

En una gráfica ponderada los vértices no tienen peso, sin embargo, en la filtración de Vietoris-Rips del complejo clique debe de existir un peso para cada $n-$simplejo, con $n \geq 0$. Para que haya consistencia en las definición y en la filtración, es necesario que $w(e) \in \mathbb{R}^+$ para toda $e \in E$; además, el complejo clique $Cl(G)$ tiene una función de pesos inducida por ν para cada simplejo $\sigma \in Cl(G)$ como se muestra a continuación:

$$\nu(\sigma) = \begin{cases} \text{máx}_{e\in\sigma}\, w(e) & \text{si } dim(\sigma) \geq 1, \\ 0 & \text{si } dim(\sigma) = 0. \end{cases}$$

De modo que para todo $\epsilon \geq 0$ se cumple que

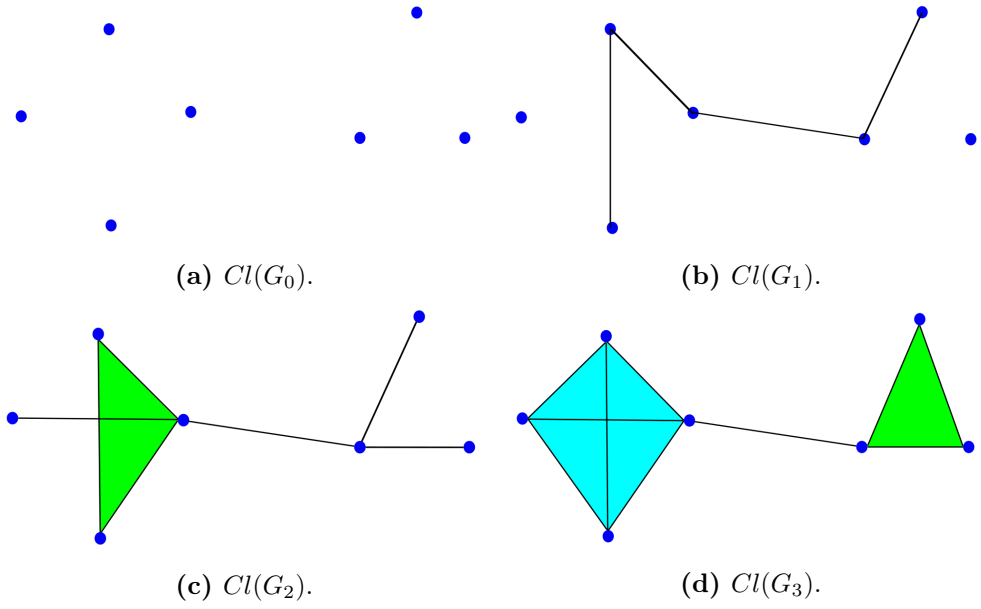
$$Cl(G_\epsilon) = Cl(G)_\epsilon,$$

donde $Cl(G)_\epsilon = \nu^{-1}((-\infty, \epsilon])$.

Es necesario mencionar que la anterior relación no se cumple si se consideran los superniveles para la filtración: $\nu^{-1}((-\infty, \epsilon])$.

De manera informal, en la filtración de Vietoris-Rips en redes aparecen primero los vértices, después se ordenan los pesos de las aristas desde $w_{\text{mín}}$ a $w_{\text{máx}}$ y se deja que el parámetro ϵ incremente de 0 a $w_{\text{máx}}$. Cada vez que ϵ incrementa se agregan las aristas correspondientes, de manera que se toma en cuenta al complejo clique de la subgráfica G_ϵ.

Para ejemplificar la filtración de Vietoris-Rips, se tomará en cuenta la gráfica de la Figura 3.5a en donde $w_{\text{máx}} = 3$, por ello el parámetro ϵ toma valores en $\{0, 1, 2, 3\}$ y crea la filtración que se muestra en la Figura 3.10.



(a) $Cl(G_0)$. **(b)** $Cl(G_1)$.

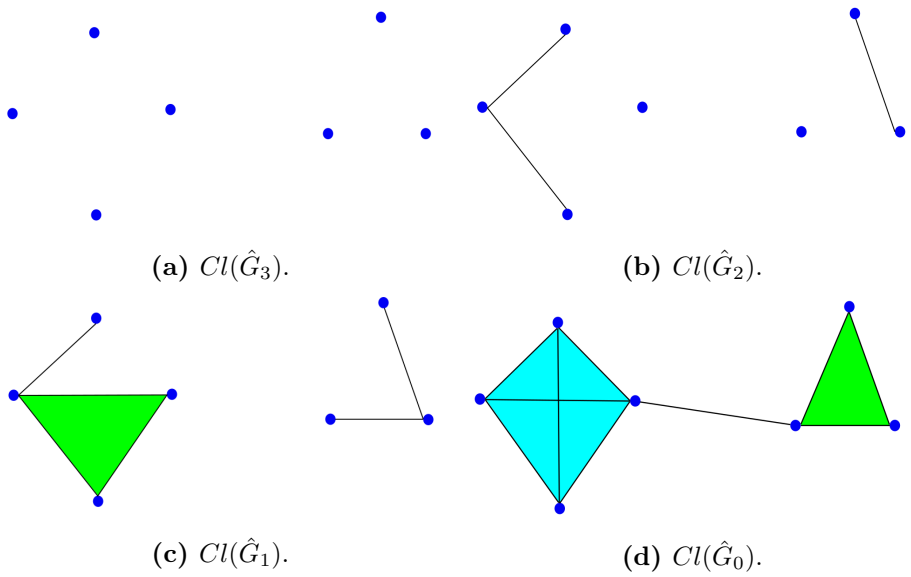
(c) $Cl(G_2)$. **(d)** $Cl(G_3)$.

Figura 3.10: Filtración de Vietoris-Rips en redes.

En muchas ocasiones es conveniente poner las aristas con mayor peso antes que aquellas aristas con pesos pequeños para enfatizar la importancia de los pesos en un gráfica. Antes de continuar, es necesario mencionar que la siguiente definición no coincide con la definición usual de análisis topológico de datos para superniveles. (Aktas Mehmet E. y El, 2019).

***Definición* 3.22.** *En la filtración de Vietoris-Rips inversa se agrega primero el conjunto de vértices, después se clasifica los pesos de las aristas de* $w_{\text{máx}}$ *a* $w_{\text{mín}}$ *y para cualquier* $\epsilon \geq 0$*, agregamos las aristas cuyo peso es mayor* ϵ*.*

Un ejemplo de la filtración de Vietoris-Rips inversa se puede observar en la Figura 3.11, en donde otra vez se toma en cuenta la gráfica de la Figura 3.5a y se utiliza la notación $\hat{G}_\epsilon$ para hacer referencia al complemento de la subgráfica G_ϵ según la Definición 3.18.



(a) $Cl(\hat{G}_3)$. **(b)** $Cl(\hat{G}_2)$.

(c) $Cl(\hat{G}_1)$. **(d)** $Cl(\hat{G}_0)$.

Figura 3.11: Filtración de Vietoris-Rips inversa en redes.

A pesar de trabajar de manera similar, es posible ver que la evolución de las características topológicas de la red obtenidas con la filtración de Vietoris-Rips son diferentes a las obtenidas con su versión inversa y para mostrar de manera precisa dichas diferencias es necesario prestar atención a las Figuras 3.10b y 3.11 b, que son las "equivalentes " en la evolución del parámetro ϵ. En la Figura 3.10b podemos observar que hay sólo tres componentes conexas y cero huecos, por otro lado en la Figura 3.11 b hay cuatro componentes conexas y cero huecos. Sucede algo similar en las Figuras 3.10c y 3.11c, ya que en la primera aparece sólo una componente conexa, mientras que en la segunda se pueden observar dos.

Es posible observar que las anteriores filtraciones tienen una estrecha relación con el complejo clique, ya que en ambas se ha utilizado el complejo clique para la construcción de los subcomplejos simpliciales geométricos. Sin embargo, existe otra filtración importante que tiene una relación más directa con el complejo clique, esta filtración se origina al poner los cliques de la gráfica original conforme la dimensión de estos va aumentando (Aktas Mehmet E. y El, 2019).

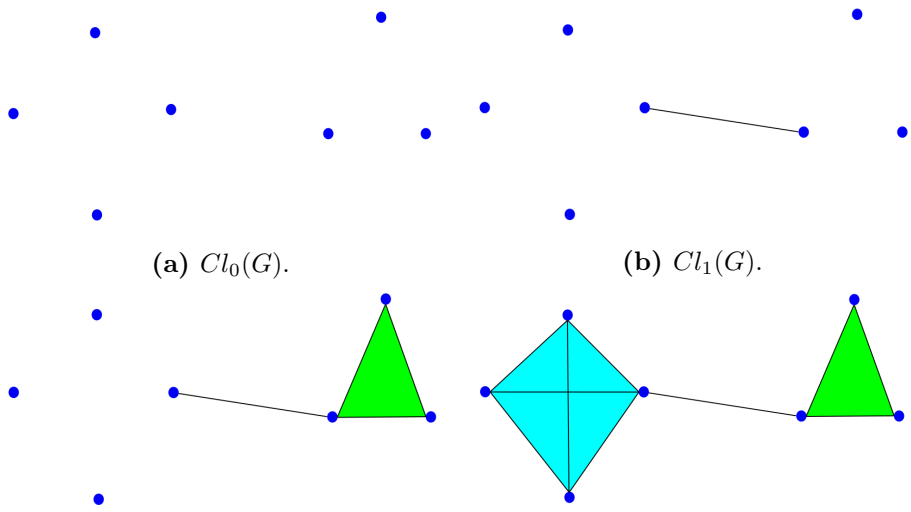
***Definición* 3.23.** *Para una gráfica G y su complejo clique $Cl(G)$ cuya dimensión es*

n, la filtración por esqueletos está definida por

$$Cl_0(G) \hookrightarrow Cl_1(G) \hookrightarrow \ldots \hookrightarrow Cl_n(G),$$

tal que $Cl_n(G) = Cl(G)$ y el $i-$ésimo complejo en la filtración está dado por $Cl_i(G) = \bigcup_{j=1}^{i} S_j$, donde S_j es el $j-$esqueleto del complejo clique original, esto es, el conjunto de simplejos geométricos con dimensión menor o igual a j.

En otras palabras, la filtración del complejo clique agrega los vértices en $\epsilon = 0$, las simplejos formados por aristas en $\epsilon = 1$, los triángulos en $\epsilon = 2$ y así sucesivamente. Como ejemplo tomemos el complejo clique de la Figura 3.9b, cuya filtración quedaría según la Figura 3.12.



(a) $Cl_0(G)$. **(b)** $Cl_1(G)$. **(c)** $Cl_2(G)$. **(d)** $Cl_3(G)$.

Figura 3.12: Filtración del complejo clique.

De la filtración por esqueletos es posible notar que no se necesitan pesos en las aristas, por lo que se puede utilizar para gráficas no ponderadas. También es posible realizar filtraciones con pesos definidos en los vértices en lugar de en las aristas, tales filtraciones no se mostrarán aquí porque no se utilizan en este trabajo, pero si el lector se ha quedado con curiosidad, puede dar un vistazo a Aktas Mehmet E. y El (2019) donde se exponen más filtraciones para complejos simpliciales geométricos.

3.4. Homología persistente en redes

El objetivo del estudio de la homología persistente en las redes es el mismo que en el caso de los complejos simpliciales abstractos, se pretende estudiar la evolución de los grupos de homología en la red de acuerdo a la filtración seleccionada.

Para mostrar el proceder en el estudio de la homología persistente en redes se utilizará un ejemplo, en donde se tomará en cuenta la gráfica ponderada de la Figura 3.13.

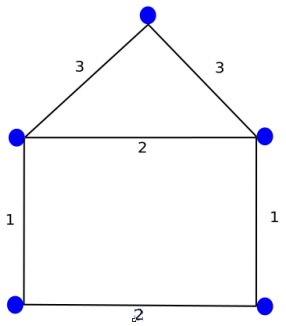
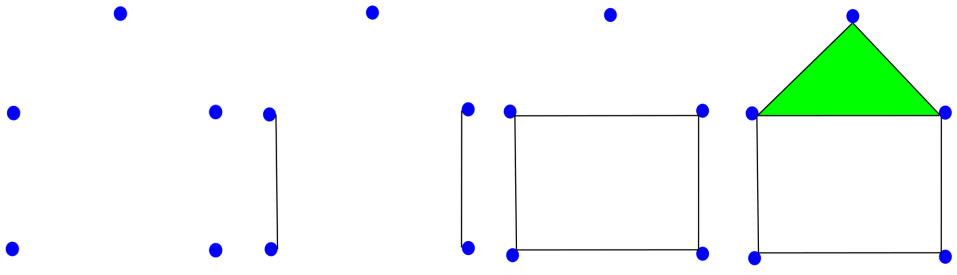


Figura 3.13: Gráfica ponderada.

Al utilizar la filtración de Vietoris-Rips según la Definición 3.21, los complejos cliques de la filtración quedan como en la Figura 3.14.



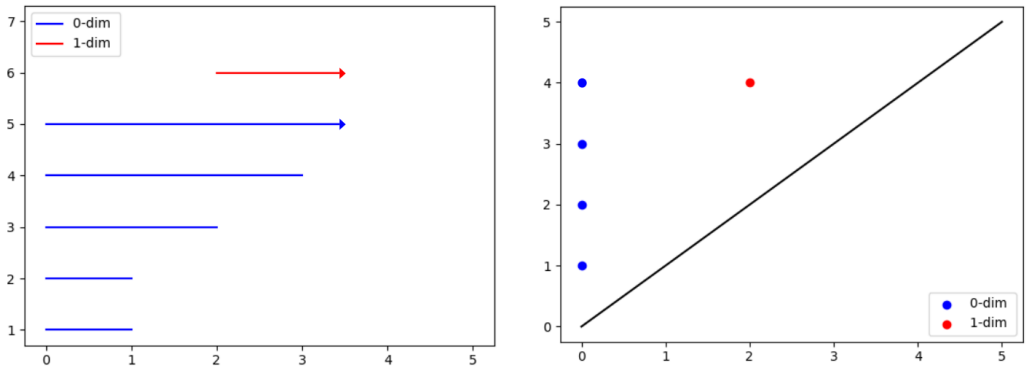
(a) $Cl(G_0)$. **(b)** $Cl(G_1)$. **(c)** $Cl(G_2)$. **(d)** $Cl(G_3)$.

Figura 3.14: Filtración de Vietoris-Rips.

Podemos observar en la filtración de la Figura 3.14 que en $Cl(G_0)$ hay cinco vértices, por lo tanto se deben de tener cinco componentes conexas. En $Cl(G_1)$ aparecen dos aristas y cada una une a un par de vértices, por ello el número de componentes

conexas disminuye a tres. En $Cl(G_2)$ aparecen otras dos aristas y en esta ocasión generan el primer hoyo de dimensión uno, además el número de componentes conexas disminuye a dos. Por último, en $Cl(G_3)$ aparece el último par de aristas, haciendo que se tenga sólo una componente conexa y además en este punto el hoyo no ha desaparecido.

Habiendo descrito lo anterior, es posible mostrar el código de barras y el diagrama de persistencia que resumen la evolución de las características topológicas de la gráfica considerada (ver Figura 3.14).



(a) Código de barras. **(b)** Diagrama de persistencia.

Figura 3.15: Evolución de grupos de homología de $Cl(G_0)$ a $Cl(G_3)$.

CAPÍTULO 4

Red de movilidad ciudadana

4.1. Datos

La invención de los celulares inteligentes trajo consigo la creciente demanda de aplicaciones móviles y las compañías que crean estas aplicaciones han sabido tomar ventaja de la información de los usuarios que se registran día con día a sus plataformas virtuales. Es bien sabido que para utilizar las aplicaciones móviles (sobre todo en las gratuitas) se deben aceptar los requisitos que las compañías exigen en sus contratos de uso y privacidad, dichos requisitos podrían variar desde pedir acceso al hardware del dispositivo móvil como el procesador, memoria, cámara, lampara, etc, hasta pedir que se comparta información de los usuarios en tiempo real como la hora y el lugar en la que se utiliza la aplicación.

El gran interés que existe hoy en día por los diversos métodos de análisis de datos se debe en parte a la información que las compañías exigen a los usuarios de aplicaciones móviles, ya que esto permite extraer información de la población más rápidamente y con costos más bajos que con otro tipo de métodos como los censos y encuestas.

Dicho lo anterior, es necesario mencionar que los datos que se han utilizado para este trabajo provienen de aplicaciones móviles, específicamente se utilizan datos como la hora y la localización satelital que el GPS de los celulares comparten cada vez que los usuarios acceden a una aplicación móvil en partícular.

Antes de comenzar a definir una posible red de movilidad, a continuación se especifica el formato de la información en la base de datos que se considerará para el estudio:

- ID: combinación de números y letras que identifican en forma única al dispositivo celular.
- Marca de tiempo: fecha y hora (UTC) de uso de la aplicación.
- Posición: par ordenado con latitud y longitud en que se localiza el dispositivo, en la fecha y hora de acceso de la aplicación.

En necesario mencionar que se tienen datos desde el 21 de septiembre del 2020 hasta el 13 de Diciembre del 2020, intervalo de tiempo en el que aún estaban vigentes la mayoría de medidas sanitarias para contrarrestar los contagios por covid19 en la ciudad de Hermosillo, Sonora y áreas aledañas..

También es importante mencionar que se va a suponer que cada ID corresponde a una persona, ya que normalmente los celulares son de uso personal, pero también puede haber eventos que no se ajusten a nuestro supuesto, como el que una persona cambie de celular (y por tanto de ID). En los análisis realizaremos el estudio por intervalos de tiempo, esto reducirá el error en el supuesto, además de permitirnos estudiar la red de movilidad en diversos intervalos temporales.

Para poder observar de mejor manera el movimiento de las personas en la ciudad de Hermosillo, quizá lo más conveniente es tomar el enfoque mostrado en Tizzoni y cols. (2014) o en Calabrese, Ferrari, y Blondel (2014), en donde se divide la superficie

total de una ciudad (o el lugar en donde se encuentra la población de estudio) en sectores más pequeños delimitados por algún criterio (como el alcance de las antenas de telefonía celular), esto permitiría estudiar los traslados de la población entre los diferentes sectores de la ciudad.

Antes de mencionar algunas metodologías que se pueden considerar para la construcción de la red de movilidad, es importante hablar de la selección y limpieza de los datos, ya que independiente de la metodología a elegir, será la información que se utilice para la construcción de la red.

4.2. Limpieza de los datos

Como se mostrará más adelante, en este trabajo se ha optado por separar a la ciudad de Hermosillo por sectores definidos por las áreas geoestadísticas básicas (AGEBs) de la ciudad. De esta manera se estudiará el desplazamiento de la población entre las AGEBs que la componen, pero antes de entrar en más detalles, es necesario mencionar el procedimiento seguido para la selección y la limpieza de los datos utilizados.

La selección y limpieza de los datos consiste en las siguientes dos etapas:

1. Selección de la muestra de datos. Como se tienen muchos IDs registrados que no aparecen de manera frecuente, y para el beneficio del trabajo de cómputo, se ha optado por separar los datos por semanas (de Lunes a Domingo) y tomar en cuenta sólo los datos de los IDs que aparecen más de cien veces en dicha semana. Nótese que los IDs que se toman en cuenta en una semana no necesariamente serán los que se tomen en cuenta en las semanas subsecuentes.

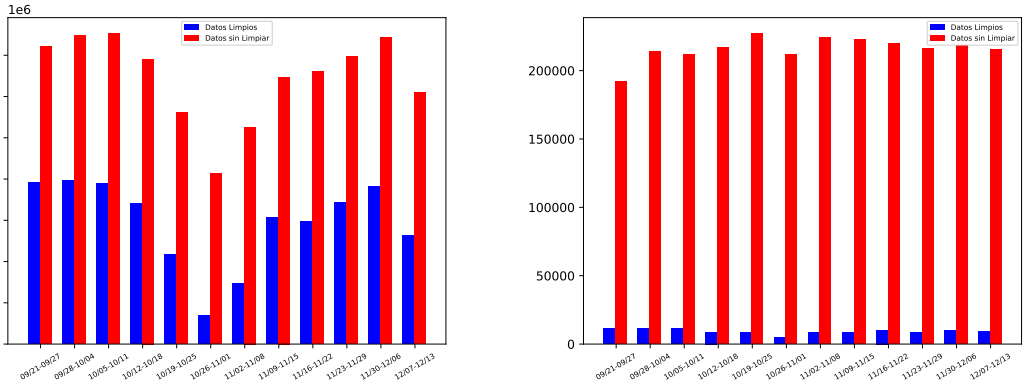
2. Limpieza de los datos. Al comparar las coordenadas de la muestra de datos anterior contra las ubicaciones de cada AGEB, es posible notar que no todas las posiciones (longitud y latitud) de la muestra caen dentro de una AGEB, por ello sólo se toman en cuenta aquellos datos cuyas posiciones coinciden con el

interior de alguna AGEB de la ciudad de Hermosillo.

Después de la anterior depuración, se consigue una tabla cuyas filas tienen la siguiente información.

- ID: combinación de números y letras.
- Marca de tiempo: fecha y hora (UTC) de uso de la aplicación.
- Posición: par ordenado con la latitud y longitud.
- AGEB: un número entre 1 y 582 que corresponde al AGEB que pertenece el dato (hay 582 AGEBs en la ciudad de Hermosillo). Cabe señalar que la numeración utilizada no es la oficial.

En la Figura 4.1 se puede observar un par de imágenes que revelan el cambio en la base de datos antes (datos sin limpiar) y después de la selección y limpieza (datos limpios).



(a) Número de activaciones. **(b)** Número de IDs.

Figura 4.1: Número de activaciones y de IDs por semana.

En la Fifura 4.1a se muestra el cambio en el número de filas que tiene cada base de datos, entendiéndose que una activación equivale a una fila en la respectiva base, además se puede notar que en muchas ocasiones se obtienen menos de la mitad de datos disponibles en la base original.

En la Figura 4.1b se muestra el número de IDs diferentes que aparecen en cada base de datos y como es notable, quedan muy pocos IDs en las bases de datos limpias de cada semana, sin embargo esa "pequeña" cantidad de IDs genera un gran volumen de activaciones.

Con lo anterior ya se puede observar que la gran mayoría de los IDs en las base original no se activan regularmente y para dar una mejor vista de esto, es posible mostrar el promedio de activaciones de los IDs en cada semana.

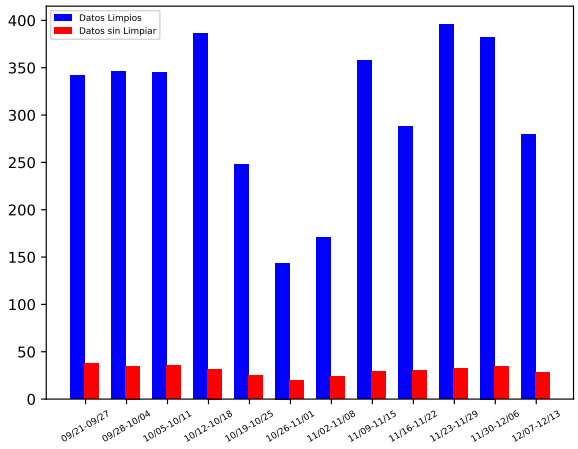


Figura 4.2: Promedio de activaciones por ID.

En la Figura 4.2 se puede observar como el promedio de activaciones por ID en la base de datos sin limpiar es menor que en la base de datos limpia. Con el proceso de selección y limpieza nos aseguramos de quitar a todos aquellos IDs que generarían ruido al momento de construir las redes de movilidad, evitando generar patrones aislados producto de algún traslado ocasional.

Para facilitar la explicación de la construcción de la red en la siguiente sección, se le llamará "Datos" a la tabla obtenida después de la selección y limpieza. Teniendo la anterior estructura en la base de datos, ya es posible mostrar la metodología usada para la construcción de las redes de movilidad en cada semana.

4.3. Red de movilidad

Como ya se mencionó anteriormente, la red, cuya construcción se muestra a continuación, estudia el desplazamiento de las personas entre las diferentes AGEBs de la ciudad de Hermosillo. Cabe señalar que para la construcción de la red no se toma en cuenta en ningún momento el número de habitantes que en teoría tendría cada AGEB.

Como los datos con los que se está trabajando tienen la hora en la que se "activan" los celulares, es tentador utilizar ventanas temporales para la construcción de la red como en Tizzoni y cols. (2014), donde se crea una red que estudia la movilidad entre el hogar y el lugar de trabajo de las personas. Sin embargo, es posible utilizar la información que se tiene para construir una red que estudie de manera más general el movimiento de la población.

Para poder utilizar todos los datos disponibles, se optó por crear una matriz de "saltos" M que muestra la manera en la que la gente se mueve entre AGEBs, la construcción de esta matriz se propone siguiendo el algoritmo en el diagrama de la Figura 4.3.

Los elementos $M_{m,n}$ de la matriz de saltos representan la cantidad de veces que las personas se desplazan "directamente" del AGEB m al AGEB n. Tomemos en cuenta que para que una persona vaya del AGEB m al AGEB n es posible que tenga que pasar por otras AGEBs intermedios, sin embargo no es posible obtener esta información de ninguna manera con los datos que se tienen. De manera que es necesario entender que cuando se dice que una persona se desplaza del AGEB m al AGEB n de manera directa, en realidad se quiere decir que después de registrarse su posición en la AGEB m se registró su posición en la AGEB n, independientemente de la ruta elegida o del tiempo que haya tardado en llegar.

Es necesario hacer notar que por construcción la matriz M no es simétrica, por lo

que no puede ser la matriz de adyacencias de una gráfica no dirigida, sin embargo es posible emplearla para la construcción de gráficas dirigidas.

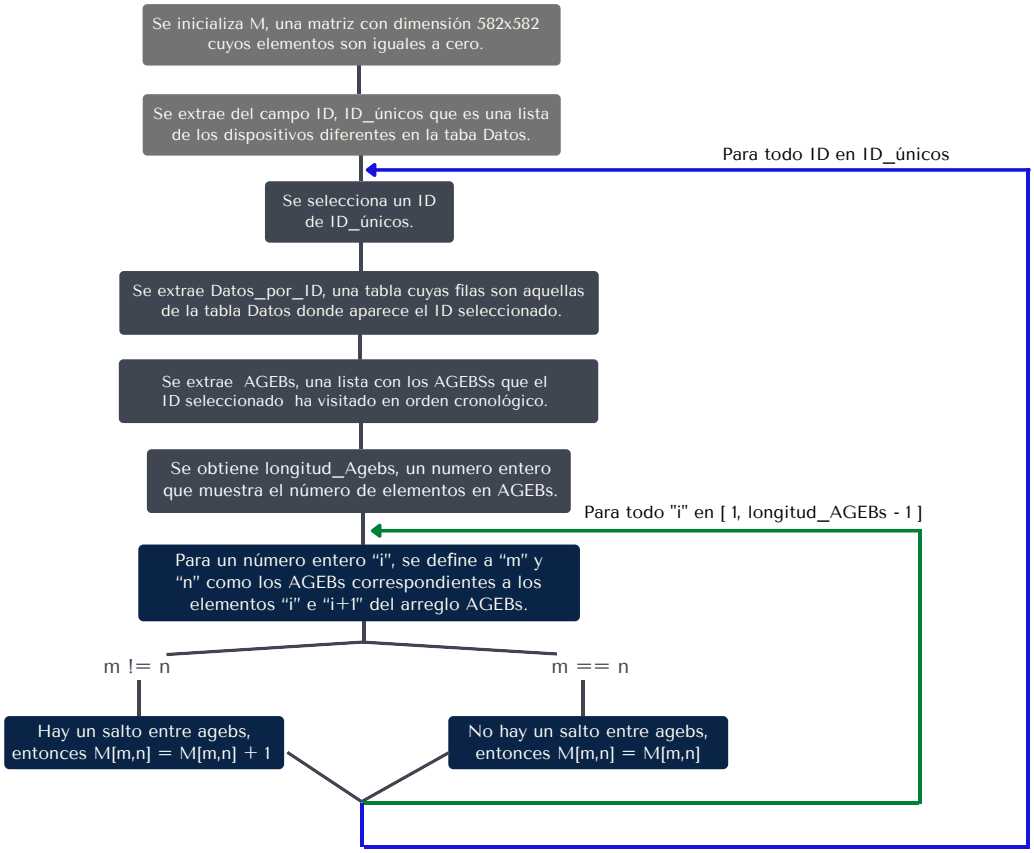


Figura 4.3: Construcción de la matriz de saltos.

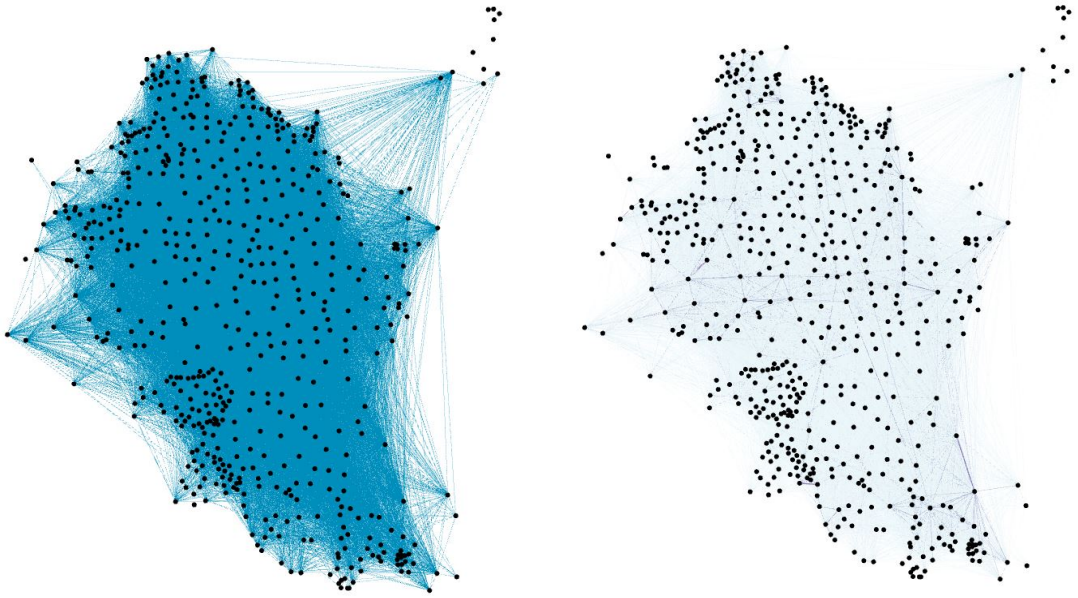
La red de movilidad se construye a partir de su matriz de adyacencias D, cuyos elementos están dados por el volumen total de movimientos entre AGEBs. Esto es,

$$D_{m,n} = M_{m,n} + M_{n,m}.$$

La matriz D por construcción es simétrica y cada uno de sus elementos nos dice la cantidad total de personas que las AGEBs m y n comparten a lo largo de la semana.

En la Figura 4.4 se pueden observar dos redes construidas a partir de la matriz D, en las cuales se representan las AGEBs de la ciudad de Hermosillo con vértices color

negro y las aristas reflejan el intercambio de personas entre las diferentes AGEBs. La intensidad de las aristas en la Figura 4.4a es la misma para todas, sin importar el peso de cada una, mostrando todas aquellas AGEBs que intercambian aunque sea una sola persona. La intensidad de las aristas en la Figura 4.4b varia conforme el peso de cada una, resaltando aquellas aristas cuyas AGEBs (extremos) intercambian mayor volumen de personas.



(a) Red sin ponderación.

(b) Red con ponderación.

Figura 4.4: Red de movilidad para la semana del 21/09/2020 al 27/09/2020.

4.4. Resultados importantes de la red de movilidad

En la anterior sección se muestra cómo es posible crear una gráfica ponderada no dirigida por medio de la matriz de adyacencias D y en cierto sentido la gráfica describe el movimiento de la población entre las diferentes AGEBs de la ciudad de Hermosillo. En esta sección se mostrarán algunas peculiaridades que se han observado en las redes de movilidad construidas.

Llegados a este punto, es necesario mostrar un poco del contexto en el que se

encontraba la ciudad de Hermosillo durante el tiempo de estudio. Para ello es necesario recordar que los datos con los que se trabajaron fueron recopilados en la época de la pandemia de covid19 y a causa de ésta, muchas ciudades del mundo habían tomado ciertas medidas sanitarias para tratar de controlar la propagación del virus entre la población. En la Figura 4.5 se puede observar el número de casos nuevos confirmados de covid19 en la ciudad de Hermosillo y en color rojo se muestran las fechas del 21 de Septiembre del 2020 y 13 de Diciembre del 2020, que es justamente el intervalo de tiempo en el que se tomaron los datos con los que se está trabajando.

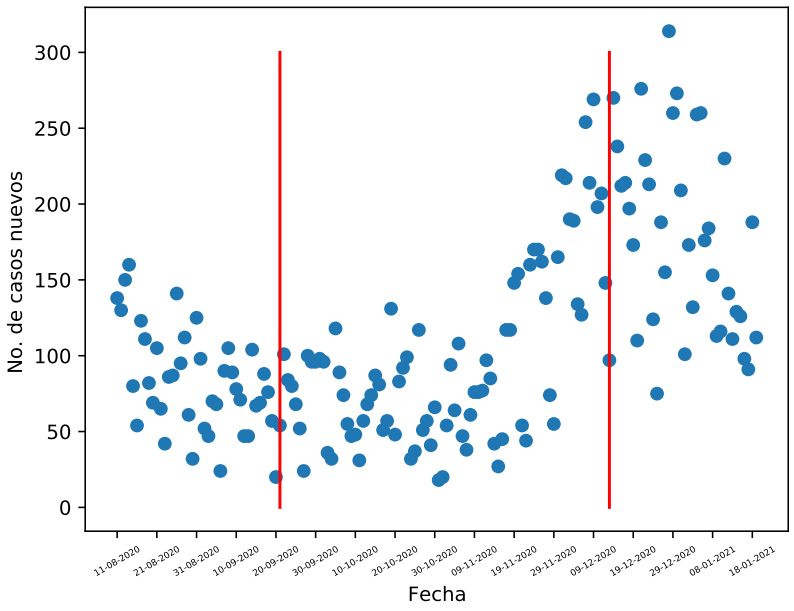
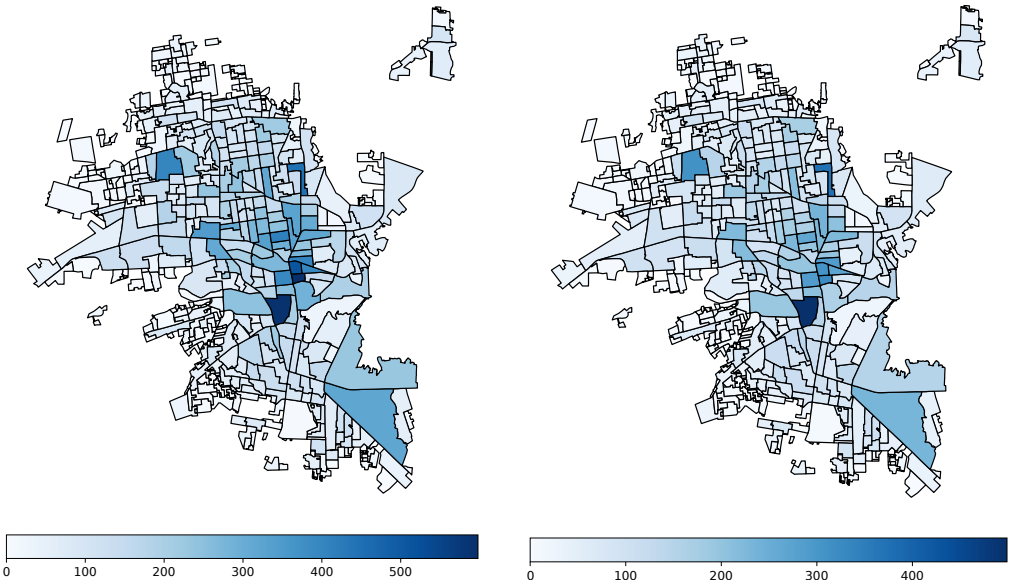


Figura 4.5: Número de casos nuevos de covid19 en Hermosillo.

Para mitigar los contagios de covid19 entre la población, en la ciudad de Hermosillo se dejaron de realizar todas aquellas actividades que no eran esenciales y en consecuencia se cerraron muchos establecimientos a los que las personas asistían de manera recurrente. Para dar más contexto acerca de la ciudad, es necesario mencionar que se utilizará "unidad económica" para referirse a un establecimiento, en el cual se realiza la producción o comercialización de bienes o servicios, asentado en un lugar de manera permanente y delimitado por construcciones e instalaciones fijas. Dicho lo anterior, en la Figura 4.6 se muestra la ciudad de Hermosillo dividida por AGEBs y

el número de unidades económicas en cada una. Se puede observar el cambio en la densidad de unidades económicas por AGEBs cuando sólo están trabajando aquellas que son esenciales en el transcurso de la pandemia.



(a) Total de número de unidades económicas.

(b) Número de unidades económicas esenciales.

Figura 4.6: Unidades económicas de la ciudad de Hermosillo.

Es importante notar que las unidades económicas son posibles atractores de personas, es decir, la necesidad de trabajo, así como la necesidad de bienes y servicios podrían generar que las personas se desplacen del AGEB donde residen hacia las AGEBs donde pueden cubrir dichas necesidades.

Para poder tener idea de cuáles AGEBs podrían ser expulsores de personas, es necesario conocer la densidad poblacional de cada AGEB, ya que la mala combinación de unidades económicas en un AGEB densamente poblado podría obligar a las personas a buscar cubrir sus necesidades en otras AGEBs. Por ello, en la Figura 4.7 se muestra la densidad poblacional de cada AGEB en Hermosillo.

Sabiendo que el elemento $M_{m,n}$ de la matriz de saltos M muestra la cantidad de personas que se desplazan directamente del AGEB m al n, es posible conocer el nú-

mero de personas que salen y entran a las diferentes AGEBs de la ciudad durante la semana, ya que si se suman todos los elementos de la $m-$ésima fila de M se obtiene el número de personas que han salido del AGEB m y de manera similar, al sumar todos los elementos de la $n-$ésima columna de M se obtiene el número de personas que han entrado al AGEB n.

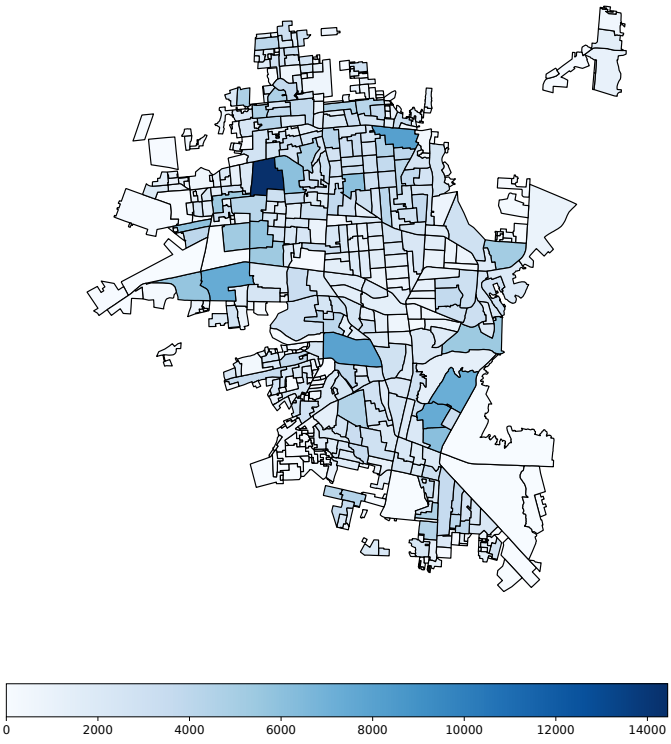
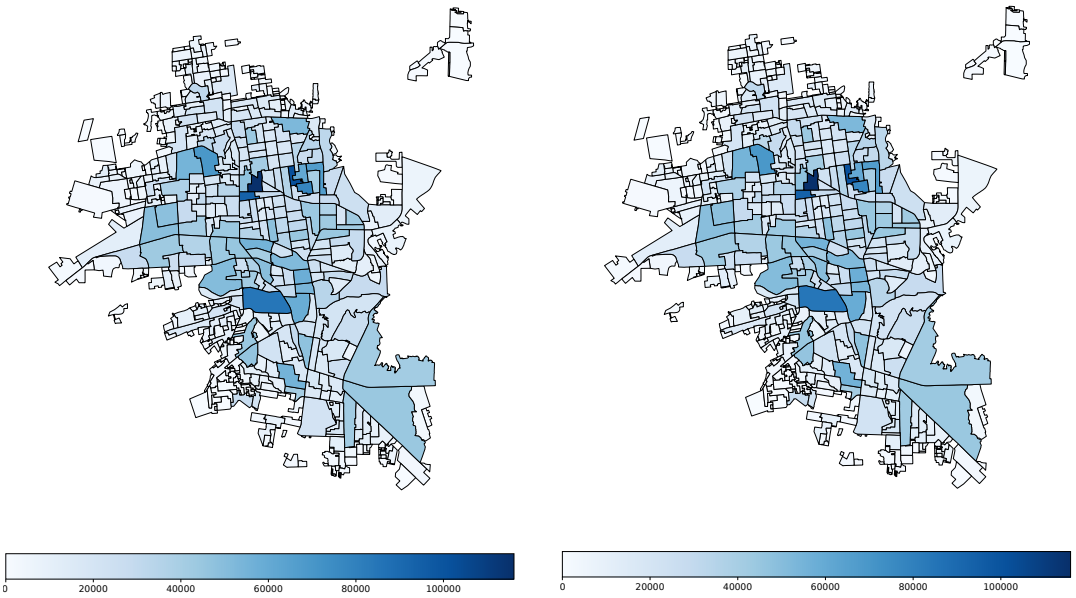


Figura 4.7: Número de habitantes.

Notemos que del 21/09/2020 al 13/12/2020 hay 12 semanas y a cada una le podemos asociar una matriz de salto M^i con $i = 1, 2, 3..., 12$. Al no traslaparse días entre las diferentes semanas, la información que se recaba en cada matriz M^i no es redundante y por ello es posible sumar las matrices de cada semana sin que se repita la información de saltos. Es decir

$$S = \sum_{i=1}^{12} M^i.$$

La matriz S tiene todos los saltos que se han dado desde el 21/09/2020 hasta el 13/12/2020, de manera que al sumar los elementos de sus filas y los elementos de sus columnas se consigue el número de personas que salieron y entraron a cada AGEB durante el periodo de estudio (ver Figura 4.8).



(a) Número de personas que entran. **(b)** Número de personas que salen.

Figura 4.8: Total de personas que entran y salen de las AGEBs desde el 21/09/2020 hasta 13/12/2020.

En la Figura 4.8a se pueden observar aquellas AGEBs que son atractores de personas, mientras que en la Figura 4.8b se muestran aquellos que son expulsores de personas. Al comparar ambas imágenes es posible notar que son casi idénticas, lo que nos sugiere que las personas eventualmente regresan a su punto de partida (como sucedería en el trayecto del día a día de las personas). También es importante observar que hay algunas AGEBs que a pesar de no tener muchas unidades económicas ni alta densidad poblacional, hay muchas personas que entran y salen de ellos. Quizá estas AGEBs sean transitorios para llegar a otros o exista algún servicio que es muy solicitado.

CAPÍTULO 5

Análisis topológico de datos en la red de movilidad ciudadana

En los anteriores capítulos se ha mostrado la teoría necesaria para poder aplicar el Análisis Topológico de Datos (ATD) en redes, así como la metodología empleada para la construcción de movilidad de la ciudad de Hermosillo. En este capítulo se exponen los resultados del estudio realizado con el ATD en la red de movilidad construida.

5.1. Análisis topológico de datos para la visualización de discrepancias en la movilidad

Como se mencionó en el Sección 2.4, es posible comparar dos diagramas de persistencia mediante la distancia de cuello de botella y la distancia de Wasserstein (Definiciones 2.41 y 2.42 respectivamente). Aunque originalmente se emplean a dichas distancias para comparar dos diagramas de persistencia que se realizaron con filtraciones diferentes del mismo complejo simplicial, también es posible utilizarlas como criterios de discrepancia entre dos complejos simpliciales que pertenecen al mismo fenómeno de estudio.

Para poder profundizar en lo anterior, es necesario recordar que en este estudio se realiza una red por cada semana (considerada como 7 días consecutivos) con la intención de estudiar la movilidad de las personas en la ciudad de Hermosillo. Si bien es cierto que la movilidad de cada persona en la ciudad dependerá específicamente del día en consideración, se esperaría encontrar ciertos patrones de movilidad poblacional independientemente de la semana que se esté considerando. En cierto sentido cada red describe el mismo fenómeno (la movilidad de las personas en la ciudad), sin embargo las redes que se construyen no son iguales, ya que pueden tener diferentes aristas e incluso diferentes pesos en aquellas aristas que tengan en común.

Si para cada semana se obtiene una red de movilidad diferente que describe el mismo fenómeno, es deseable tener algún criterio que permita conocer en cuáles semanas la movilidad puede considerarse igual y cuáles son aquellas que discrepan más. Esta información podría quedar como referencia para la planeación de actividades dentro de la ciudad y en el contexto de la pandemia, se podría observar un cambio en el flujo de la movilidad de la ciudad, el cual permitiría tomar acciones para atender posibles contagios en masa.

Dicho lo anterior, en este trabajo se considera el análisis topológico de datos para encontrar las discrepancias en la movilidad de la ciudad semana a semana, proponiendo utilizar la distancia de cuello de botella y la distancia de Wasserstein como medidas de cercanía en el comportamiento de la movilidad.

Concretamente se propone la siguiente metodología para medir las diferencias en la movilidad de cada semana.

1. Construcción de redes de movilidad (según la Sección 4.3) para cada semana.
2. Construir complejos simpliciales para cada red utilizando los mismos criterios.
3. Obtener los diagramas de persistencia para cada complejo simplicial utilizando los mismos criterios en la construcción de cada filtración.

4. Medir la cercanía entre todos los diagramas de persistencia con la distancia de cuello de botella y la distancia de Wasserstein.

Para este punto ya se ha especificado cómo se construyen las redes de movilidad, pero aún es necesario puntualizar el complejo simplicial que se le asociará a cada red y los criterios para generar la filtración a utilizar en los respectivos diagramas de persistencia.

5.2. Complejo simplicial y filtración a utilizar

Como ya se había mencionado anteriormente, la gráfica ponderada construida a partir de la matriz de adyacencias D mostrada en la Sección 4.3 representa el intercambio total de personas entre AGEBs durante una semana. Como la red construida no es dirigida, se le puede asociar un complejo clique según lo mostrado en la Definición 3.20, al cual se le pueden aplicar todas las herramientas del ATD en redes.

Habiendo elegido el complejo simplicial que se le asociará a las redes de movilidad, lo siguiente a elegir es la filtración que se utilizará para la construcción de los diagramas de persistencia. Si el complejo simplicial a utilizar es el complejo clique, resulta fácil asociar de igual manera la filtración de esqueletos según la Definición 3.23 y de esta manera construir una familia de complejos simpliciales cuyos simplejos aparecen según su dimensión, sin embargo al utilizar esta filtración se estaría omitiendo mucha información acerca de la movilidad.

Recordemos que las aristas en la red de movilidad están ponderadas por el número de personas que intercambian dos AGEBs m y n, donde dicho peso está dado por el elemento $D_{m,n}$ de la matriz de adyacencia D. La ponderación en la red de movilidad refleja el flujo de personas que se desplazan por las diferentes zonas de la ciudad y por ello es imprescindible para el estudio aprovechar esta información. Una posible manera de aprovechar la información del flujo de personas es por medio de la filtración y para ello resulta natural trabajar con filtraciones del estilo de Vietoris-Rips

(ver definición 3.21 y 3.22).

Para realizar el análisis topológico de datos en las redes de movilidad se ha optado por utilizar dos filtraciones, ambas son una combinación de las filtraciones de Vietoris-Rips y de su versión inversa. Pero antes de explicar en qué consisten las filtraciones que se han utilizado en el estudio, es necesario mencionar los dos principales motivos por las que se han elegido.

1. Como ya se ha mencionado en la anterior sección, a pesar de que las redes de movilidad describen el mismo fenómeno, no son iguales. Sabiendo que dos gráficas ponderadas pueden tener la misma gráfica subyacente, es posible tener algunas aristas que aparescan en todas las redes, pero con una ponderación w considerablemente diferente en cada semana. Lo anterior provoca que las escalas máximas ϵ_{max} de las filtraciones del complejo simplicial (Definición 2.40) difieran mucho entre sí y como consecuencia, los diagramas de persistencia cambian mucho semana a semana. Para poder realizar una comparación más justa entre los diferentes diagramas de persistencia, se tiene que utilizar una función de peso ν que nos permita homogeneizar la escala de las filtraciones de cada diagrama.

2. Además de homogeneizar la escala de los diagramas de persistencia, se busca que la filtración priorice aquellas aristas cuyo peso es mayor que las demás, por ello se buscaría que conforme el parámetro de control de la filtración evoluciona, aparezcan primero en el complejo simplicial las aristas con mayor peso.

Habiendo dicho las necesidades que se buscan cubrir con la filtración, ya se puede hablar a detalle de las dos filtraciones que se han empleado en este trabajo.

La primer filtración utilizada en este estudio modifica la filtración de Vietoris-Rips inversa (Definición 3.22) al normalizar los pesos que tienen las aristas en la gráfica, esto se realiza al dividir las ponderaciones de cada arista entre el valor del elemento

más alto en D.

Si $G = (V, E)$ es la gráfica generada por la matriz de adyacencia D, entonces para toda arista $e_{m,n} = (m, n) \in E$ se cumple que el elemento $D_{m,n} \geq 1$ y con esto se puede generar una función de pesos $\hat{w} : E \to (0, 1]$ como se muestra a continuación:

$$\hat{w}(e_{m,n}) = \frac{D_{\text{máx}} - D_{m,n} + 1}{D_{\text{máx}}},$$

donde D_{max} corresponde al valor más alto de la matriz de adyacencias D. Con la anterior función de pesos definida sobre las aristas de la gráfica, se puede definir una función de pesos sobre los simplejos geométricos σ del complejo simplicial geométrico $Cl(G)$ como se muestra a continuación:

$$\hat{\nu}(\sigma) = \begin{cases} \text{máx}_{\{e_{m,n} \in \sigma\}} \, \hat{w}(e_{m,n}) & \text{si } \dim(\sigma) \geq 1, \\ 0 & \text{si } \dim(\sigma) = 0. \end{cases}$$

La anterior función de pesos es similar a la utilizada en la filtración de Vietoris-Rips de la Sección 3.3, además los subcomplejos simpliciales estarían dados por $Cl(G)_\epsilon = \{\sigma \in Cl(G); \hat{\nu}(\sigma) \leq \epsilon\}$ y también cumplen que $CL(G_\epsilon) = Cl(G)_\epsilon$.

Al utilizar la ponderación dada por $\hat{w}$ se puede generar una filtración al estilo de Vietoris-Rips (Definición 3.21), pero en este caso la escala ϵ siempre va de 0 a 1 sin importar cuál gráfica se utilice para generar la filtración. Lo anterior resuelve el problema de la escala entre los diagramas de persistencia, ya que todos los diagramas tendrán $\epsilon_{max} = 1$ y con esto se puede hacer una comparación más justa entre los diagramas de persistencia.

También es gracias a $\hat{w}$ que aquellas aristas $e_{m,n}$ cuyo peso original $D_{m,n}$ es mayor, aparecen primero en la filtración. Lo anterior invierte el orden con el cual aparecen las aristas en la filtración de Vietoris-Rips, haciendo que el resultado sea más similar a la versión inversa.

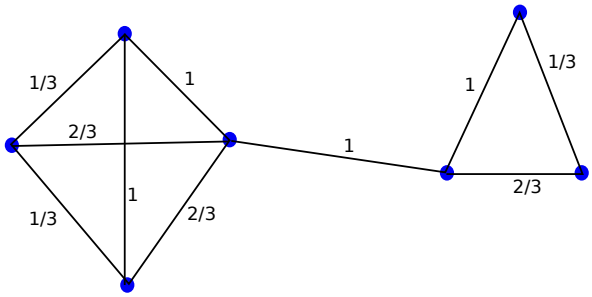
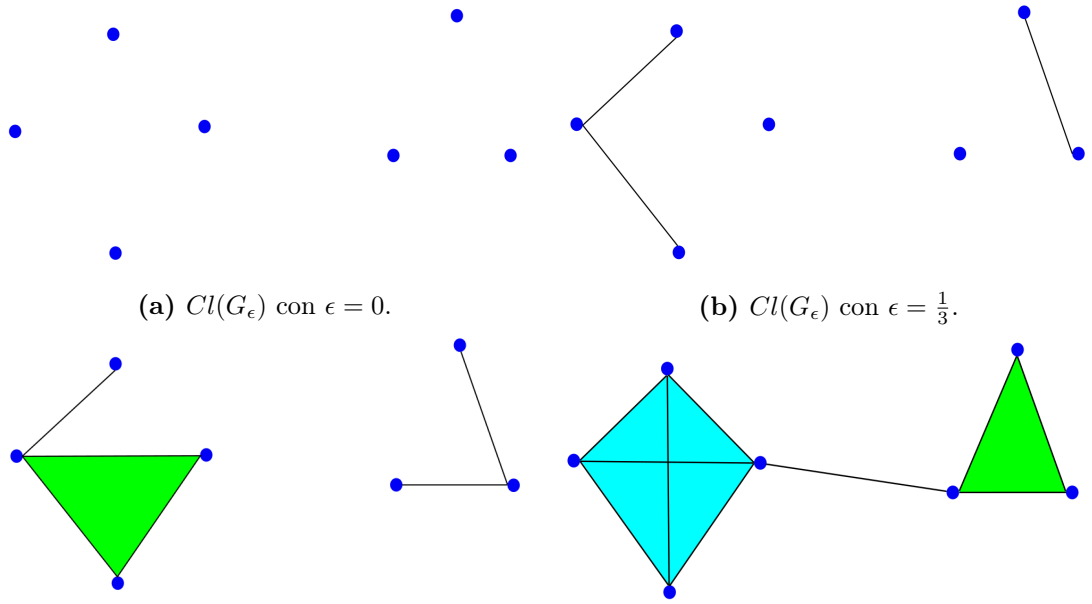


Figura 5.1: Ponderaciones transformadas según la filtración de VRIN.

Para mostrar un ejemplo de la anterior filtración, que de ahora en adelante llamaremos filtración de *Vietoris-Rips Inversa Normalizada (VRIN)*, se tomará en cuenta la gráfica ponderada de la Figura 3.5a. Notemos que en la gráfica considerada el valor del mayor peso es igual a 3, por ello $D_{max} = 3$ y por lo tanto al realizar la transformación con $\hat{w}$ se obtendrían las ponderaciones de acuerdo a la Figura 5.1. Teniendo ya transformados los pesos de las aristas, la filtración de VRIN del complejo simplicial clique asociado a la Figura 5.1 quedaría como se muestra en la Figura 5.2.



(a) $Cl(G_\epsilon)$ con $\epsilon = 0$.

(b) $Cl(G_\epsilon)$ con $\epsilon = \frac{1}{3}$.

(c) $Cl(G_\epsilon)$ con $\epsilon = \frac{2}{3}$.

(d) $Cl(G_\epsilon)$ con $\epsilon = 1$.

Figura 5.2: Filtración de VRIN para la gráfica 3.5a.

En la Figura 5.2 es posible observar cómo la filtración de VRIN es casi idéntica a la

filtración de Vietoris-Rips inversa, la diferencia recae en que en la última el parámetro ϵ evoluciona de manera descendiente y toma valores mayores a 1. Para convencerse de las semejanzas y diferencias sólo es necesario comparar las Figuras 5.2 y 3.11.

La segunda filtración utilizada en este trabajo también es una ligera variación de la filtracion de Vietoris-Rips inversa, la modificación consiste en utilizar los recíprocos de las matriz de adyacencia.

Si $G = (V, E)$ es la gráfica generada por la matriz de adyacencia D, entonces para toda arista $e_{m,n} = (m, n) \in E$ se cumple que el elemento $D_{m,n} \geq 1$ y con esto se puede generar una función de pesos $\bar{w} : E \to (0, 1]$ como se muestra a continuación:

$$\bar{w}(e_{m,n}) = \frac{1}{D_{m,n}}.$$

Con la anterior función de pesos definida sobre las aristas de la gráfica, se puede definir una función de pesos sobre los simplejos geométricos σ del complejo simplicial geométrico $Cl(G)$ como se muestra a continuación:

$$\bar{\nu}(\sigma) = \begin{cases} \text{máx}_{\{e_{m,n} \in \sigma\}} \, \bar{w}(e_{m,n}) & \text{si } \dim(\sigma) \geq 1, \\ 0 & \text{si } \dim(\sigma) = 0. \end{cases}$$

Al utilizar la ponderación dada por $\bar{\nu}$ se pueden generar los subcomplejos simpliciales de la filtración como $Cl(G)_\epsilon = \{\sigma \in Cl(G); \bar{\nu}(\sigma) \leq \epsilon\}$ y además se sigue cumpliendo que $CL(G_\epsilon) = Cl(G)_\epsilon$. Con lo anterior se consigue de nuevo una filtración al estilo de Vietoris-Rips cuya parámetro ϵ evoluciona de 0 a 1, en donde se resuelve el problema de la escala y se prioriza aquellas aristas $e_{m,n}$ con mayor peso original $D_{m,n}$.

Para mostrar un ejemplo de la anterior filtración, que de ahora en adelante llamaremos filtración de Vietoris-Rips recíproca, también se tomará en cuenta la gráfica ponderada de la Figura 3.5a y al realizar la transformación correspondiente a $\bar{w}$ se

obtiene lo mostrado en la Figura 5.3

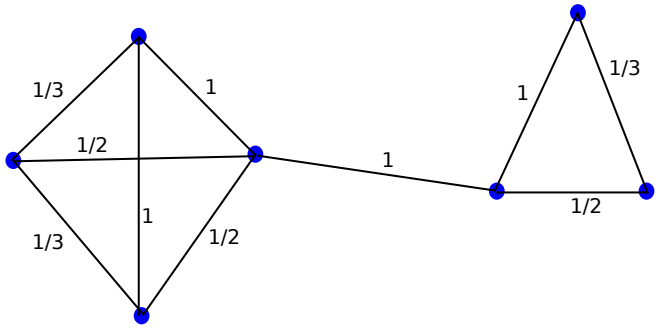
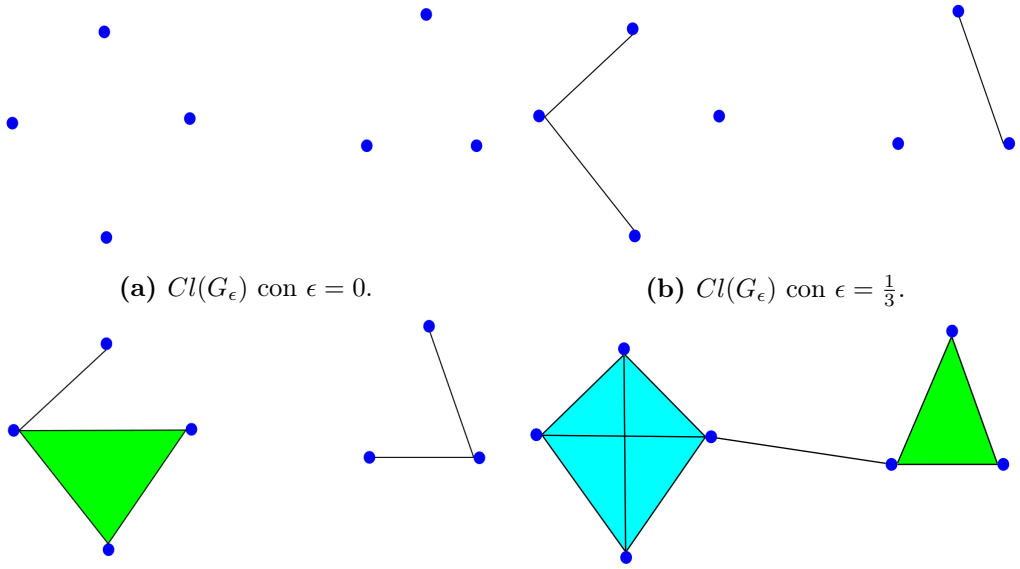


Figura 5.3: Ponderaciones transformadas según la filtración de Vietoris-Rips recíproca.

En la Figura 5.4 se muestra la evolución de los subcomplejos simpliciales en la filtración de Vietoris-Rips reciproca y de nueva cuenta es notable la semejanza con la filtración de Vietoris Rips inversa (Figura 3.11).



(a) $Cl(G_\epsilon)$ con $\epsilon = 0$.

(b) $Cl(G_\epsilon)$ con $\epsilon = \frac{1}{3}$.

(c) $Cl(G_\epsilon)$ con $\epsilon = \frac{1}{2}$.

(d) $Cl(G_\epsilon)$ con $\epsilon = 1$.

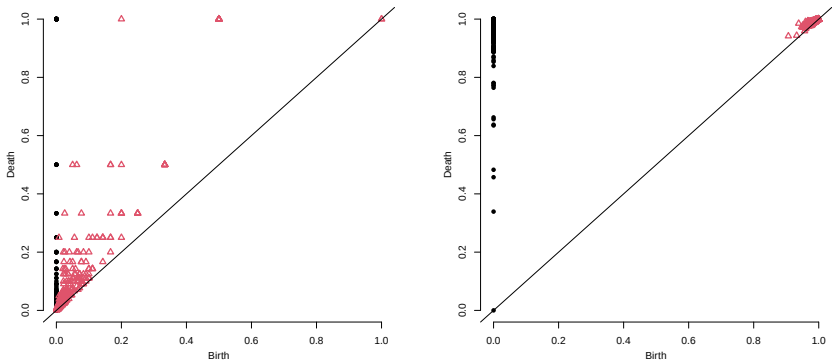
Figura 5.4: Filtración de Vietoris-Rips recíproca para la gráfica 3.5a.

A pesar de que la filtración de Vietoris-Rips inversa normalizada y la filtración de Vietoris-Rips recíproca parecieran ser la misma, en realidad no lo son. Si bien es cierto que las características topológicas irán apareciendo en el mismo orden conforme las respectivas filtraciones van evolucionando, estas persistirán de manera diferente y

dependiendo particularmente de la filtración.

Notemos que para la filtración de Vietoris-Rips inversa normalizada el parámetro de control ϵ tiene saltos de $1/D_{max}$, para ejemplificar con mayor detalle lo anterior véase la Figura 5.2c, en donde aparecen dos componentes conexas cuando $\epsilon = 2/3$ y desaparecen en $\epsilon = 1$ para hacerse una sola componente conexa. Por lo anterior, es posible decir que las dos componentes conexas persisten durante 1/3 de "tiempo", que corresponde a un paso de evolución natural del parámetro ϵ para esta filtración.

Por otro lado, en la filtración de Vietoris-Rips recíproca, el paso con el que avanza el parámetro ϵ no es constante y además el "tiempo" se va elongando conforme se acerca ϵ a uno. Ejemplo de lo anterior es que hay un infinidad de números cercanos a cero (1/10, 1/100, 1/1000,...), pero sólo se tiene un paso de 1/2 a 1 (recordemos que estamos tomando los recíprocos de números naturales), esto hace que la mayoría de las características topológicas aparezcan y desaparezcan antes de $\epsilon = 1/2$.



(a) Diagrama hecha con la filtración de Vietoris-Rips reciproca.

(b) Diagrama hecha con la filtración de Vietoris-Rips inversa normalizada.

Figura 5.5: Diagramas de persistencia para la semana del 21/09/2020 al 27/09/2020.

En la Figura 5.5 se pueden observar los diagramas de persistencia correspondientes a cada filtración utilizada en este trabajo y se puede apreciar claramente cómo los tiempos de persistencia son completamente diferentes.

5.3. Sobre las características topológicas

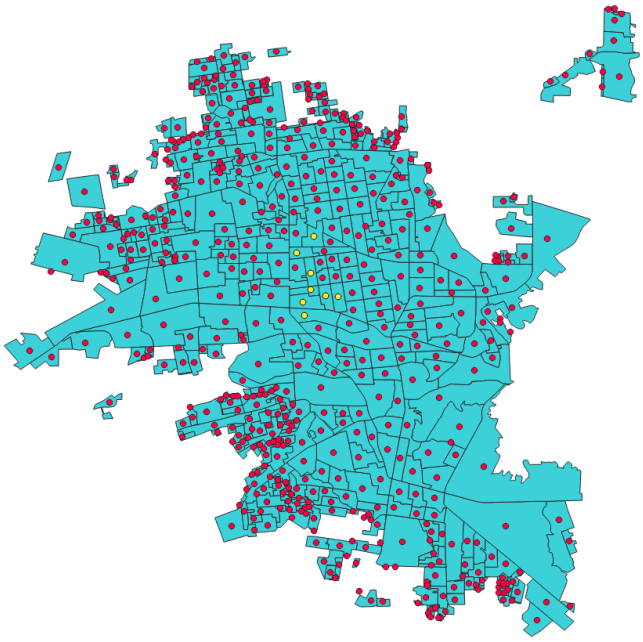
En la anterior sección se muestran las particularidades sobre el complejo simplicial y las filtraciones que se utilizarán en este trabajo, sin embargo, aún no se comenta con claridad la relevancia de estudiar las componentes conexas y los hoyos de dimensión uno del complejo simplicial asociado a cada red de movilidad.

Como anteriormente se ha comentado, en un diagrama de persistencia todas las componentes conexas "nacen" en cero, esto es porque lo primero que aparece en las filtraciones son los 0-simplejos o los vértices del complejo simplicial. A partir de que los vértices aparecen y en medida de que el parámetro de control de cada filtración avanza, van apareciendo los simplejos de dimensiones mayores a 0.

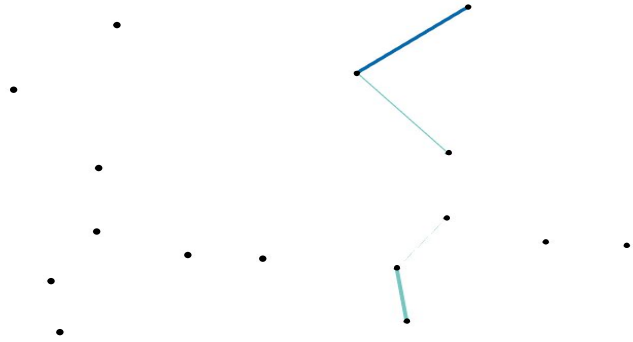
Recordemos que las filtraciones que se están utilizando en este trabajo, en realidad consisten en tomar los complejos cliques de ciertas subgráficas G_ϵ, por lo tanto en nuestro caso, conforme avanza el parámetro ϵ de cero a uno, se agregan a G_ϵ las aristas cuyo peso dado por $\hat{D}_{m.n}$ o por $\bar{D}_{m.n}$ es menor o igual a ϵ.

En la Figura 5.6 se muestra un sencillo ejemplo de la evolucion de una subgráfica G_ϵ, en este caso no se toman en cuenta todas las AGEBs de la ciudad, sólo se consideran aquellos que se muestran de color amarillo en la Figura 5.6a. Como se había mencionado anteriormente, en la evolución del parámetro ϵ de cero a uno lo primero en aparecer en la subgráfica G_ϵ son los vértices, tal y como se muestra en la Figura 5.6b. Eventualmente se van agregando las aristas cuyo peso es mayor, así como se muestran en las Figuras 5.6c-f, particularmente la Figura 5.6f es de gran utilidad para observar la ponderación de las aristas, ya que aquellas con mayor peso son más visibles y es posible notar de mejor manera la movilidad de las personas.

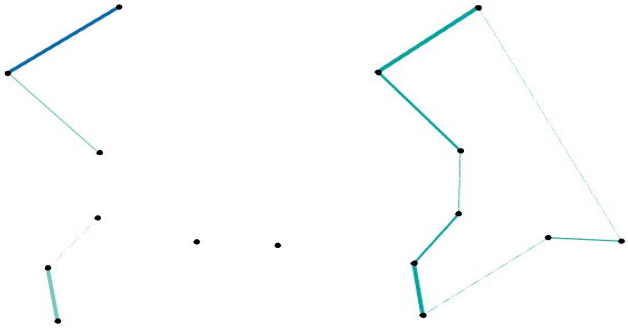
Al tomar en cuenta los complejos cliques de las Figuras 5.6b-f, es posible notar como van cambiando las componentes conexas. Concretamente, en la Figura 5.6b "nacen" ocho componentes conexas que corresponden a cada vértice. En la Figura 5.6c se



(a) AGEBs con sus respectivos centroides como nodos.

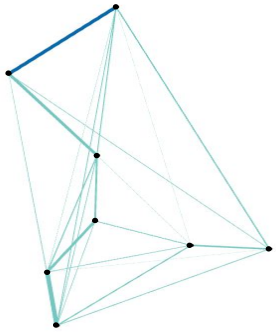


(b) Nodos a considerar.

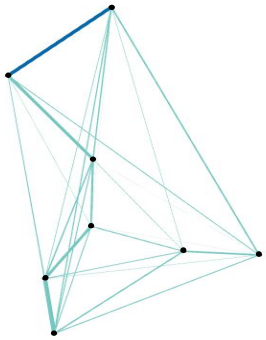


(c) Subgráfica con aristas cuyos pesos son mayores a 600.

(d) Subgráfica con aristas cuyos pesos son mayores a 300.



(e) Subgráfica con aristas cuyos pesos son mayores a 100.



(f) Subgráfica con aristas cuyos pesos son mayores a 0.

Figura 5.6: Evolución de las subgraficas G_ϵ.

unen algunas aristas por medio de aristas, haciendo que algunas componentes conexas "mueran" y en este punto se tienen cuatro componentes conexas correspondientes dos vértices y a dos grupos de tres vértices. En la Figura 5.6d ya solo queda una componente conexa correspondiente a un grupo de ocho vértices.

Con el anterior ejemplo es posible notar la importancia de estudiar las componentes conexas para obtener información de la mobilidad en la ciudad de Hermosillo, ya que en cierto sentido las componentes conexas nos muestran agrupamientos de AGEBs que dependen de un número mínimo de intercambio de personas. De lo anterior podemos esperar que aquellas AGEBs que están en el mismo grupo estén estrechamente relacionados y por lo tanto esperaríamos que los eventos que ocurran en una AGEB de un grupo en particular repercutan más rápidamente en el resto de las AGEBs del mismo grupo que en aquellas AGEBs que se encuentran en grupos distintos.

Conforme se van agregando las aristas con menor ponderación, es posible que todas las AGEBs se conecten entre sí formando una clique, haciendo más evidente que estas AGEBs pertenecen a grupo. De lo anterior nos podemos hacer la siguiente pregunta ¿qué niveles de agrupamiento son más relevantes? Quizá la respuesta obvia sería: .ªquellas en los que la ponderación es mayor", pero lo anterior no define aún un nivel crítico que además puede ser variable según las relaciones o dinámicas bajo estudio.

Para dar una respuesta un poco más elaborada se tomará en cuenta la Figura 5.6d, en donde todas las aristas forman una única componente conexa y además forman un hoyo de dimensión uno. En este punto se han agregado las aristas cuyo peso es al menos de 300 y a partir de aquí se agregan más aristas cuya ponderación en menor, pero ya no se agregan más vertices.

Con este ejemplo podemos darnos cuenta que los hoyos de dimensión uno son componentes conexas especiales, revelando una movilidad cíclica entre las AGEBs y en

cierto sentido muestran qué niveles de agrupamiento son más importantes, es decir, a partir de la Figura 5.6d ya se sabe que las ocho AGEBs forman un grupo, de manera que esperaríamos que estén estrechamente relacionados independientemente de como las aristas subsecuentes los conecten entre sí.

Con los ejemplos anteriores ya se puede tener más noción de lo que las componentes conexas y hoyos de dimensión uno nos dicen acerca de la movilidad y de la ciudad. Por otro lado, se puede apreciar la información que se puede extraer de las características topológicas de una red no direccionada y ponderada, con el fin de estudiar las diferencias entre ellas y en particular la información y diferencias entre la movilidad. Aunque dos redes de movilidad tengan los mismos arcos, se puede capturar la diferencia de sus relevancias a través de sus pesos, originando diferentes diagramas de persistencia. La elaboración de esta comparación se muestra en la siguiente sección.

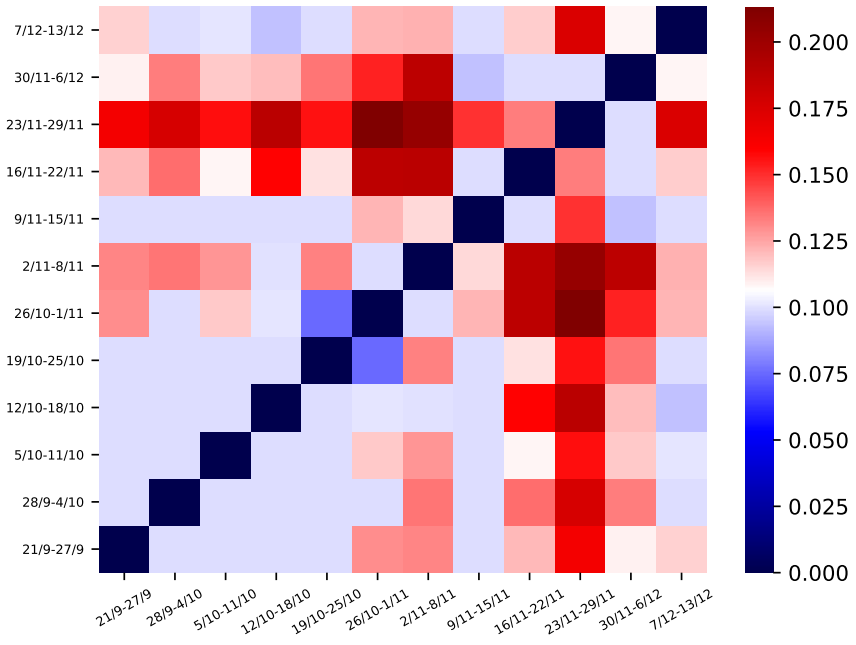
5.4. Comparación de la movilidad para cada semana

Ya se ha mencionado cómo será la construcción de los complejos simpliciales para cada red de movilidad y se ha dicho cuáles son las filtraciones que se utilizaron para obtener el diagrama de persistencia de cada semana, por lo tanto ya es posible proceder a comparar los diagramas de cada semana de acuerdo a la distancia de cuello de botella y a la distancia de Wasserstein.

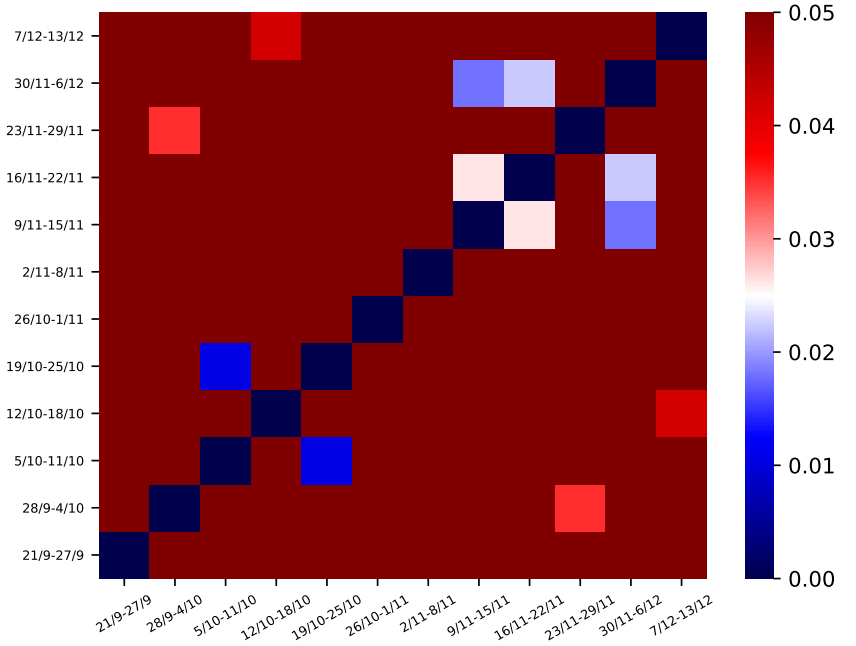
Antes de mostrar los resultados de las comparaciones, es necesario recordar que la distancia de Wasserstein depende de un parámetro p (ver la Definición 2.39) y para este estudio se ha optado por utilizar $p = 1$, de manera que es necesario tomar esto cuenta en caso de que se quieran replicar los resultados obtenidos en este trabajo.

A continuación se muestran los resultados obtenidos al comparar los diagramas de persistencia de cada semana.

En la Figura 5.7 se muestran los resultados de aplicar la distancia de cuello de



(a) Distancia para componentes conexas.



(b) Distancia para hoyos de dimensión 1.

Figura 5.7: Distancias de cuello de botella con Vietoris-Rips inversa normalizada.

botella a los diagramas de persistencia obtenidos utilizando la filtración de VRIN. Se tiene que aclarar que a pesar de que las características topológicas (componentes conexas y hoyos de dimensión 1) típicamente se grafican en el mismo diagrama de persistencia, al comparar dos diagramas con la distancia de cuello de botella sólo se toma en cuenta un tipo de característica topológica a la vez y es por ello que se muestran dos imágenes, una para la comparación de los diagramas mediante las componentes conexas y la otra para la comparación de los diagramas mediante los hoyos de dimensión 1.

Es necesario mencionar que se están graficando distancias que son simétricas, esto provoca que las imágenes sean simétricas respecto a la línea identidad y por ello se pueden leer de manera horizontal o vertical. También es necesario hacer una ligera aclaración acerca de la notación utilizada en los valores de los ejes de cada figura, ya que por motivos de espacio no se ha alcanzado expresar de manera detallada lo que significa cada valor, por ello cuando se lee en alguno de los ejes la notación UU/VV – XX/YY en realidad se quiere decir "la semana que comienza en el día UU del mes VV del año 2020 y termina el día XX del mes YY del año 2020".

En la Figura 5.7a se observan las comparaciones en las persistencias de las componentes conexas de cada diagrama de persistencia y lo primero a resaltar es que en la diagonal se encuentra la comparación de cada semana con sigo misma, por ello solo hay valores de 0. Lo segundo a resaltar es que la diferencia máxima en el comportamiento de las persistencias entre dos semanas es ligeramente mayor 0.2, y corresponde a la comparación realizada entre las semanas 26/10 - 1/11 y 23/11 - 29/11. Es posible observar que el comportamiento de las persistencias entre las semanas 21/9-27/9, 28/9-4/10, 5/10-11/10, 12/10-18/10, 19/10-25/10 y 9/11-15/11 distan de manera similar, con una distancia de cuello de botella ligeramente mayor a 0.1. También es posible notar que las persistencias del par de semanas 26/10-1/11 y 2/11-8/11 tienen un comportamiento ligeramente diferente al resto de las semanas, acentuándose al compararlas con las semanas 16/11-22/11, 23/11-29/11 y 30/11-6/12. Por último es posible notar como las persistencias de las semanas 16/11-22/11, 23/11-29/11 y

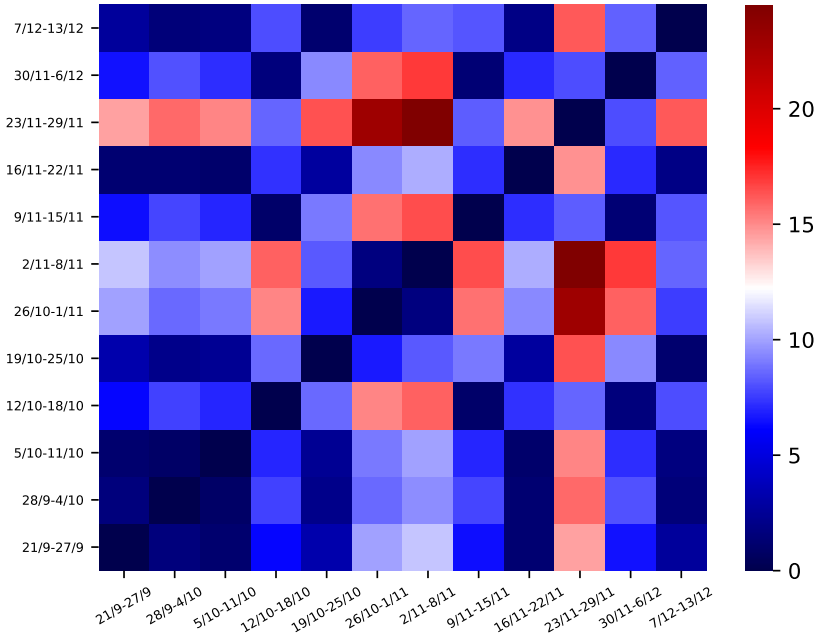
30/11-6/12 se comportan notablemente distinto al resto de las semanas, principalmente a aquellas previas a la semana 9/11-15/11.

En la Figura 5.7b se observan las comparaciones en las persistencias de los hoyos de dimensión 1 de cada diagrama de persistencia. En esta ocasión, casi todas las persistencias correspondientes a cada semana distan de igual manera con distancia de cuello de botella igual a 0.05. Cabe señalar que es notable la diferencia entre las Figuras 5.7a y 5.7b, quedando claro que al utilizar la distancia de cuello de botella sobre los diagramas de persistencia obtenidos de las respectivas filtraciones de VRIN, es más sencillo observar comportamientos interesantes al enfocarse en las componentes conexas.

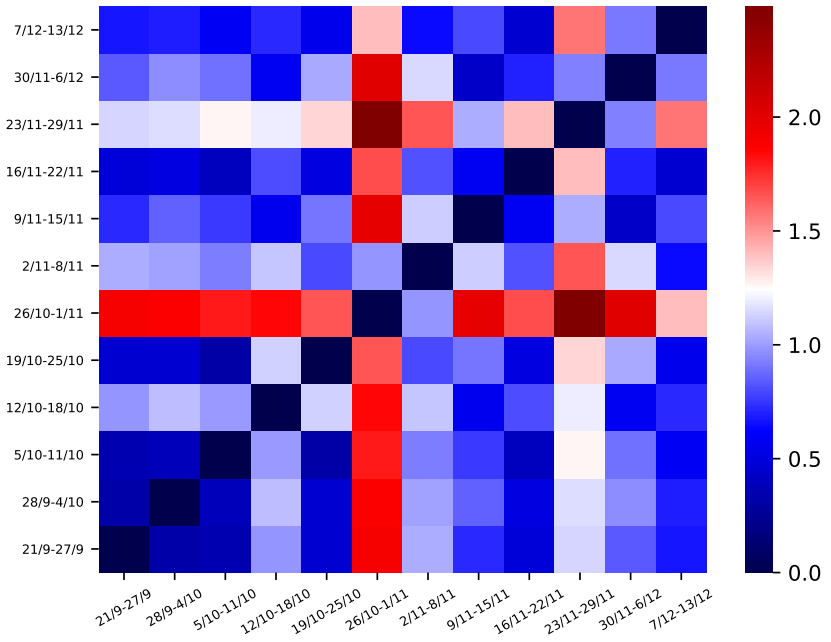
En la Figura 5.8 se muestran los resultados de aplicar la distancia de Wasserstein a los diagramas de persistencia obtenidos utilizando la filtración de VRIN. Es importante resaltar que con la distancia de Wasserstein, la diferencia máxima en el comportamiento de las persistencias de cada semana es mayor que con la distancia de cuello de botella, es decir, mientras que en la Figura 5.7a la diferencia máxima entre las persistencias de dos semanas es ligeramente mayor a 0.2, en la Figura 5.8a la diferencia máxima entre las persistencias de dos semanas es mayor a 20 y es posible notar que sucede algo similar con las distancias máximas de las Figuras 5.7b y 5.8b.

En la Figura 5.8a se observan las comparaciones en las persistencias de las componentes conexas de cada semana. En esta imagen es posible observar que el comportamiento de las persistencias de las semanas 21/9-27/9, 28/9-4/10, 5/10-11/10, 19/10-25/10, 16/11-22/11 y 7/12-13/12 distan de manera similar, con distancia de Wasserstein menor a 5. Se puede observar de nueva cuenta que las persistencias de las semanas 26/10-1/11 y 2/11-8/11 se comportan notablemente diferente al resto de las semanas pero muy similar entre sí. Por último, la semana 23/11-29/11 es la que mayor diferencias muestra en comparación al resto de las semanas.

En la Figura 5.8b se muestran las comparaciones en las persistencias de los hoyos



(a) Distancia para componentes conexas.



(b) Distancia para hoyos de dimensión 1.

Figura 5.8: Distancias de Wasserstein con Vietoris-Rips inversa normalizada.

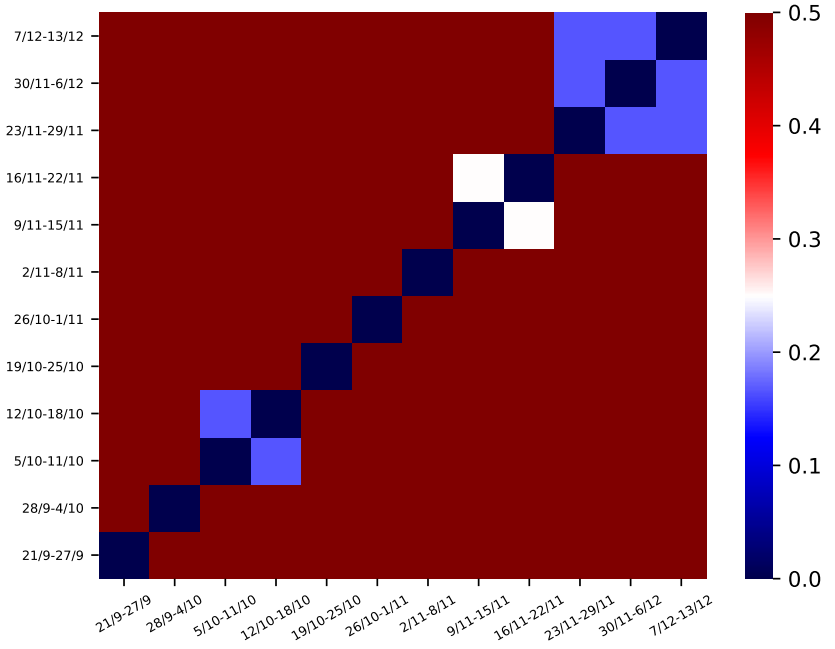
de dimensión 1 de cada semana. Lo más destacable en esta imagen es que la semana 26/10-1/11 muestra las mayores diferencias en el comportamiento de las persistencias y la semana 23/11-29/11 también muestra diferencias importantes en el comportamiento de las persistencias.

Por lo anterior, pareciera que la distancia de Wasserstein es consistente con los resultados obtenidos, independientemente de si se toman en cuenta las componentes conexas o los hoyos de dimensión 1 en el cálculo de las distancias, cosa que no sucede con la distancia de cuello de botella.

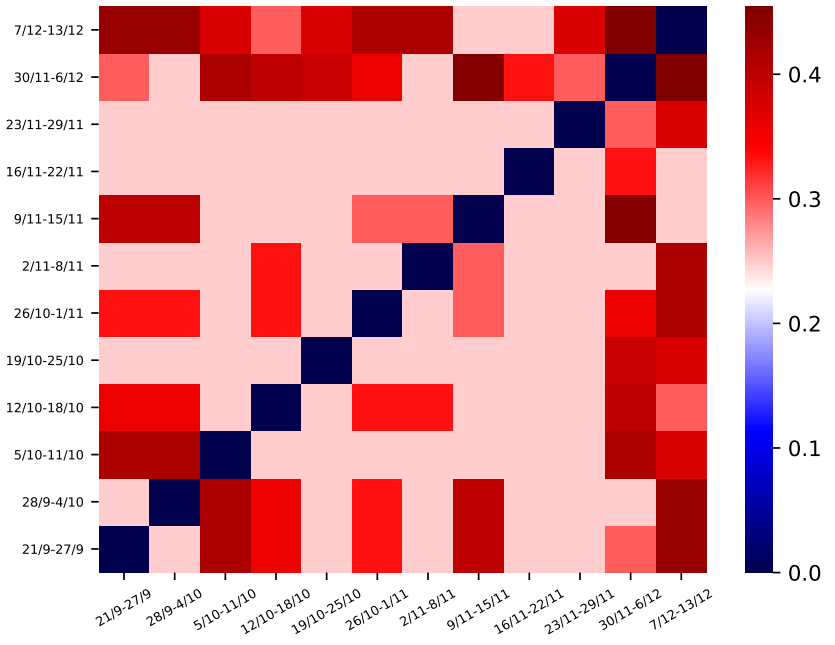
Es posible seguir el anterior proceder y comparar la distancias entre los diagramas de persistencia obtenidos con la filtración de Vietorie-Rips recíproca, con lo cual podremos ver algunos patrones de comportamiento similares a lo obtenido con la filtración de Vietorie-Rips inversa normalizada.

En la Figura 5.9 se muestran los resultados de aplicar la distancia de cuello de botella a los diagramas de persistencia obtenidos utilizando la filtración de Vietoris-Rips recíproca. Al igual que en la Figura 5.7b donde se aplica la distancia de cuello de botella en los hoyos de dimensión uno, la Figura 5.9a revela poca información acerca de lo distante que están unas semana de las otras, sólo que en esta ocasión la distancia de cuello de botella se utiliza sobre las componentes conexas. Lo más sobresaliente de la Figura 5.9a, es que aparecen dos grupos de semanas que tienen un comportamiento notablemente similar, el primero formado por las semanas 5/10-11/10 y 12/10-18/10, mientras que el segundo está formado por las semanas 23/11-29/11, 30/11-6/12 y 7/12-13/12.

En la Figura 5.9b se puede observar de mejor manera como algunas semanas distan ligeramente más de otras. Es necesario notar que en la Figura 5.9a las semanas 23/11-29/11, 30/11-6/12 y 7/12-13/12 muestran un comportamiento semejante, sin embargo, en la Figura 5.9b estás mismas semanas muestran un comportamiento



(a) Distancia para componentes conexas.



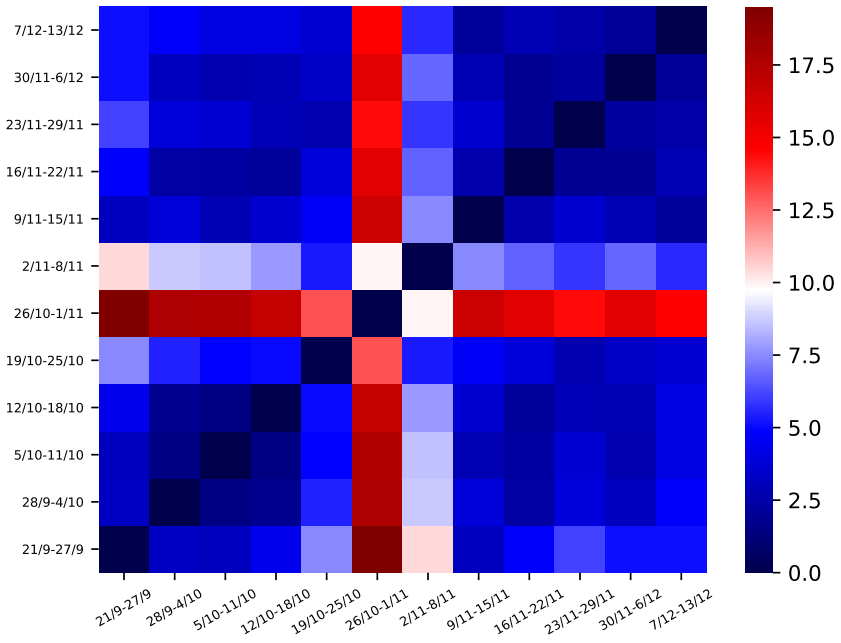
(b) Distancia para hoyos de dimensión 1.

Figura 5.9: Distancias de cuello de botella con Vietoris-Rips recíproca.

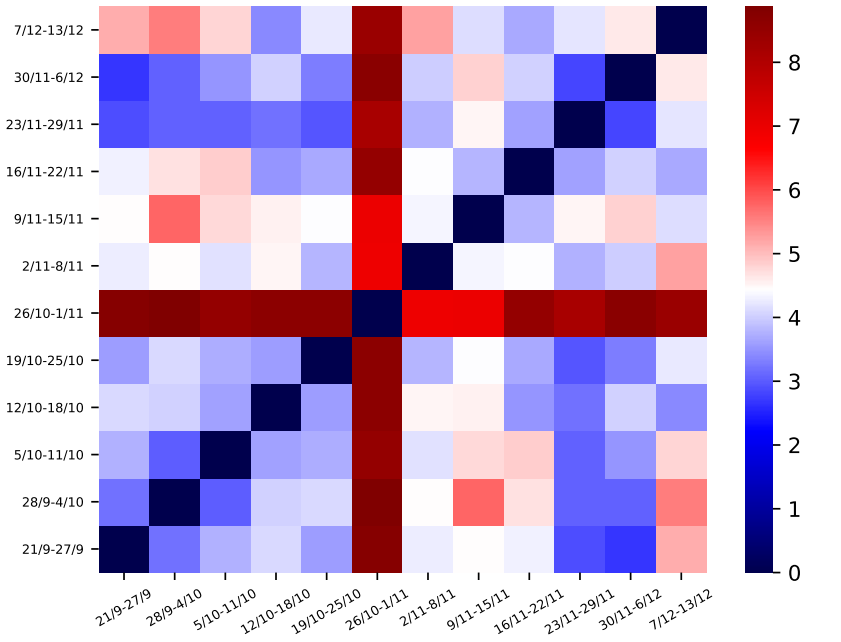
particularmente diferente entre sí. De manera similar, en la Figura 5.9a las semanas 5/10-11/10 y 12/10-18/10 muestran un comportamiento semejante, pero en la Figura 5.9b se muestra lo contrario, de manera que lo obtenido con las componentes conexas y lo obtenido con los hoyos de dimensión 1 reflejan comportamientos contradictorios.

En la Figura 5.10 se muestran los resultados de aplicar la distancia de Wasserstein a los diagramas de persistencia obtenidos utilizando la filtración de Vietoris-Rips recíproca. De manera inmediata es posible ver que con la distancia de Wasserstein se pueden diferenciar mejor los comportamientos entre las persistencias de cada semana, similar a lo que se mostró en la Figura 5.8. En la Figura 5.10a se observa que las persistencias de las semanas 26/10-1/11 y 2/11-8/11 muestran un comportamiento notablemente diferente al resto de las semanas y ligeramente similar entre sí, cosa que se ha observado anteriormente en la Figura 5.8a. En la Figura 5.10b el comportamiento de las persistencias de la semana 26/10-1/11 muestran un comportamiento notablemente diferente al resto de las semanas, coincidiendo con lo que se muestra en la Figura 5.10a y en la Figura 5.8b.

Llegados a este punto, es necesario mencionar que según lo mostrado en las Figuras 5.7, 5.8, 5.9 y 5.10, la distancia de Wasserstein es más sensible que la distancia de cuello de botella, haciendo posible observar mayores diferencias entre las semanas. También es necesario resaltar que por lo obtenido con la filtración de VRIN y con la filtración de Vietoris-Rips reciproca, es evidente que la semana 26/10-1/11 tiene un comportamiento bastante diferente al resto de las semanas. En la filtración de VRIN se muestra que la semana 23/11-29/11 se comporta notablemente diferente del resto, pero esto no se refleja en lo obtenido con la filtración Vietoris-Rips reciproca.



(a) Distancia para componentes conexas.



(b) Distancia para hoyos de dimensión 1.

Figura 5.10: Distancias de Wasserstein con Vietoris-Rips recíproca.

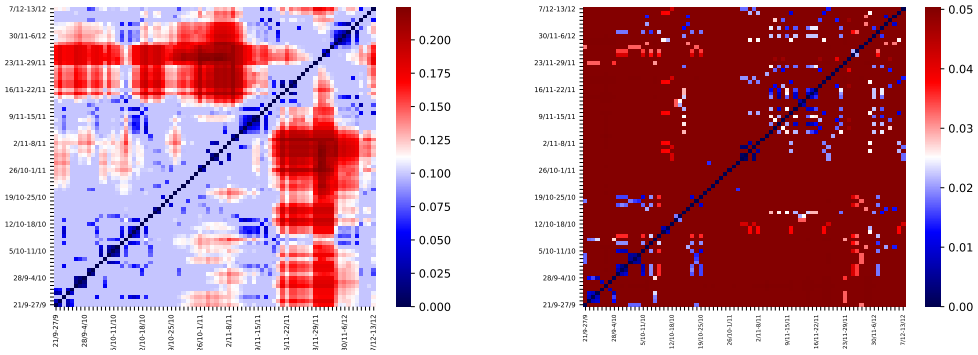
5.5. Comparación de la movilidad para semanas con desfase

En vista de lo obtenido en la anterior sección, aparece de manera natural la siguiente pregunta ¿Es posible ver las diferencias entre los diagrama de persistencia de cada red de movilidad en intervalos de tiempo más pequeños? La respuesta a la anterior pregunta es sí, pero a un alto costo.

En teoría no hay nada que nos impida realizar una red para cada día, asociar a cada una de ellas un complejo simplicial y eventualmente obtener los diagramas de persistencias para poder compararlas con las distancias antes mencionadas. Lamentablemente al realizar lo anterior nos podríamos encontrar con un problema imposible de ignorar, la incapacidad de elegir muestras que sean significativas para brindar certeza al estudio.

Una de las principales razones por las que se toman intervalos de una semana para la construcción de la red es porque se puede obtener suficiente información con la cual construir una red que refleje, de manera más cercana, la verdadera movilidad de la población en Hermosillo. Considerando ésto, es posible crear redes de movilidad semanales que se vayan desfasando por un día hacia adelante, de esta manera se podría tomar en cuenta la evolución diaria en la movilidad.

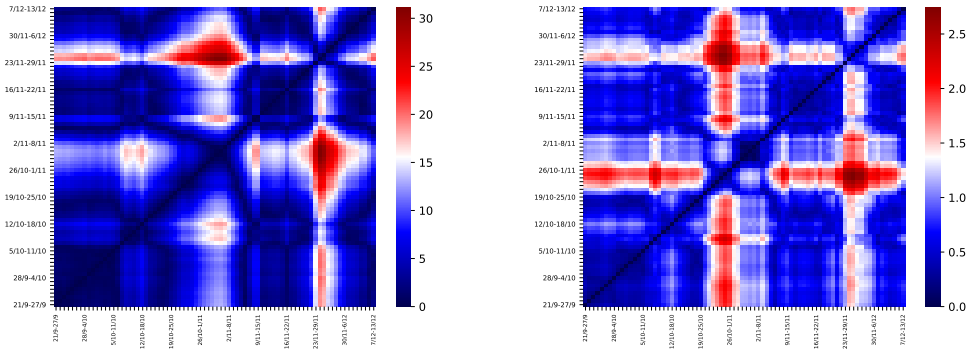
Antes de mostrar cómo cambian los diagramas de persistencia en las semanas desfasadas, es necesario aclarar que la construcción de las respectivas redes de movilidad se realiza conforme lo expuesto en la Sección 4.3, a las cuales se les asocia un complejo clique, además se utiliza la filtración de VRIN y la filtración de Vietoris-Rips recíproca para la construcción de los correspondientes diagramas de persistencia. A continuación se presentan sus correspondientes resultados, empleando las distancias de cuello de botella y de Wasserstein.



(a) Distancia para componentes conexas. **(b)** Distancia para hoyos de dimensión 1.

Figura 5.11: Distancias de cuello de botella para semanas desfasadas con Vietoris-Rips inversa normalizada.

Al comparar la Figura 5.11 con la Figura 5.7, resulta casi imposible no observar las similitudes entre las correspondientes imágenes, sin embargo es posible notar que a diferencia de la Figura 5.7a, en la Figura 5.11a se alcanzan a distinguir con mayor resolución las diferencias en los diagramas de persistencia de cada semana, aún y cuando las semanas tienen días en común.

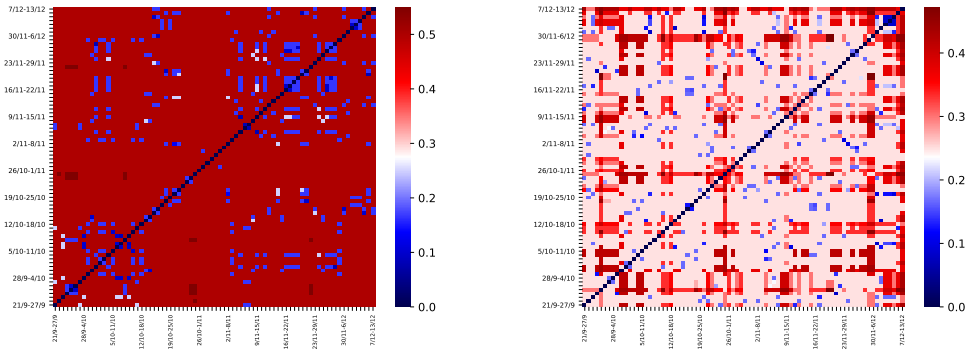


(a) Distancia para componentes conexas. **(b)** Distancia para hoyos de dimensión 1.

Figura 5.12: Distancias de Wasserstein para semanas desfasadas con Vietoris-Rips inversa normalizada.

Al comparar la Figura 5.12 con la Figura 5.8, podemos observar de nueva cuenta los mismos patrones en las correspondientes imágenes y hasta el momento sería válido pensar que desfasar las semanas por un día sólo hace que ganemos resolución entre las distancias los diagramas cuando se obtienen con la filtración de Vietoris-Rips inversa

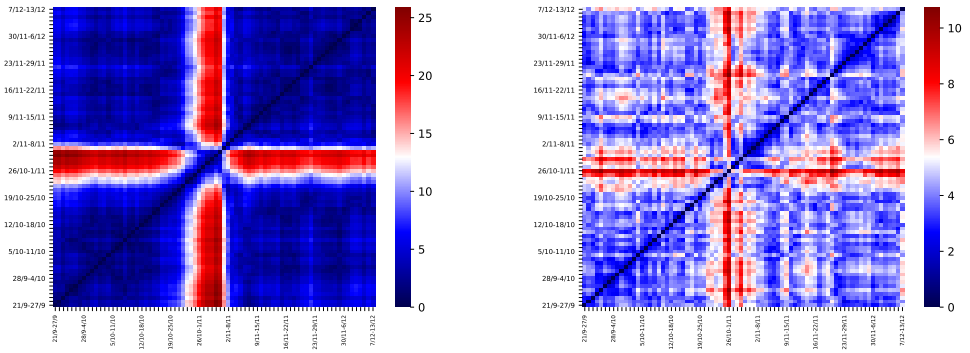
normalizada.



(a) Distancia para componentes conexas. **(b)** Distancia para hoyos de dimensión 1.

Figura 5.13: Distancias de cuello de botella para semanas desfasadas con Vietoris-Rips recíproca.

Al observar la Figura 5.13b podemos observar un incremento en los patrones en comparación a la Figura 5.9b, haciendo que resalten las diferencias entre los diagramas de algunas semanas.



(a) Distancia para componentes conexas. **(b)** Distancia para hoyos de dimensión 1.

Figura 5.14: Distancias de Wasserstein para semanas desfasadas con Vietoris-Rips recíproca.

Al comparar la Figura 5.14 con la Figura 5.11, podemos notar que la resolución incrementa de nueva cuenta, haciendo que las diferencias entre ciertas semanas en específico se vean con mayor detalle.

5.6. Interpretación y aplicación de la comparación de la movilidad entre semanas

En la Sección 5.3 se aborda la información que las características topológicas nos ofrecían acerca de la movilidad en la ciudad y de la relación entre las diferentes AGEBs de la ciudad. En este sentido, en el diagrama de persistencia se resume la información antes mencionada y por lo tanto es natural pensar que dos diagramas de persistencia diferentes muestren dos comportamientos diferentes, a pesar de que estos estudien el mismo fenómeno.

Es posible utilizar lo anterior para poder estudiar el comportamiento de la movilidad de la ciudad a través del tiempo, con lo cual es posible ir haciendo registros de la movilidad que ayuden al estudio, prevención y predicción de diversos fenómenos que estén asociados a la movilidad. En otras palabras, si se observan ciertos fenómenos sociales relacionados a la movilidad en una semana, entonces se esperaría observar los mismos fenómenos sociales aparezcan en semanas posteriores con el mismo comportamiento y con ello se podría tener tiempo anticipado para actuar y prevenir catástrofes.

Teniendo en mente lo anterior, es posible utilizar las comparaciones de la movilidad entre semanas para poder prever los nuevos casos de covid que se obtendrían en un futuro y con ello poder preparar el equipo y personal necesarios para atender las diversas olas de contingencias sanitarias.

También se podría utilizar como herramienta de control, es decir, si con un comportamiento en particular de la movilidad la taza de contagios entre la población es baja, lo que se desearía sería replicar esta movilidad para mantener bajos los contagios en un futuro, de manera que cuando se comience a ver que la movilidad dista del comportamiento deseado, se puede actuar rápidamente para volver a retomar el control de la población y regresarla de nuevo al comportamiento de movilidad deseado.

En general, este tipo de análisis sería de gran ayuda al momento de estudiar fenómenos que estén relacionados a la movilidad y cuyos efectos se perciban tiempo después, como cualquier fenómeno que se pueda modelar con una red de contacto.

CAPÍTULO 6

Conclusiones y trabajo futuro

Con el fin de estudiar algunas de las características de movilidad urbana que luego pueden utilizarse para formalizar pruebas de diferencias de patrones de movilidad y translado, en este trabajo se presentan los elementos más importantes para realizar el análisis topológico de datos organizados en forma de una red.

En la Sección 4.2 se explica el proceso que se le aplica a la base de datos original para obtener las bases de datos con las que se crean las redes de movilidad. En dicho proceder se trata de eliminar el ruido que los IDs con bajo número de activaciones pueden agregar a la movilidad de la ciudad (selección) y además se descartan las activaciones cuyas posiciones no se encuentren dentro de alguna AGEB de la ciudad (limpieza), con ello se consigue suficiente información para conocer aquellas AGEB cuyo intercambio de personas es normalmente mayor, haciendo posible estudiar el flujo cotidiano de personas dentro de la ciudad.

Los supuestos con los que se contruye la red buscan hacer factible su construcción utilizando únicamente la base con la que se cuenta. Éstos son: las personas cuyos

celulares se encuentran en la base son (al menos aproximadamente) aleatoriamente seleccionadas de la población y los registros de geolocalización no está fuertemente influenciado por variables asociados a la localización misma.

Una característica que se observa en la base de datos es que la composición de los individuos según su nivel de actividad para registrar sus geolocalizaciones, es altamente heterogénea. En general sería bastante bueno poder tener más activaciones por cada ID, con ello se podría rastrear el traslado de las personas con mucha precisión, y que las personas tuviesen aproximadamente el mismo número de activaciones. Al haber pocas personas con un alto número de activaciones, estas personas podrían aportar mayor peso a las aristas que unen las AGEB por donde se desplazan, de esta manera ciertas aristas pueden llegar a tener un peso muy grande debido al traslado excesivo de estas pocas personas y no porque se traslade mucha gente entre esas AGEB. Es por ello que se podría mejorar el criterio de selección mostrado en la Sección 4.2 al restringir el número de activaciones cuando sea muy grande y los desplazamientos sean entre pocas AGEB.

En la Sección 4.3 se muestran los pasos para la construcción de las redes de cada semana y se presenta la red construida para la semana que va del 21/09/2020 al 27/09/2020 (Figura 4.4), en la cual es posible observar las aristas con mayor peso o flujo total de personas (Figura 4.4b). A pesar de que en la Figura 4.4b ya se pueden observar ciertas relaciones importantes entre AGEB, solo se puede validar en algunos casos para lo que se recurre a información particular de la ciudad como servicios y distribución poblacional.

En la Sección 4.4 se muestran algunas imágenes que ayudan a entender un poco la realidad que vive la ciudad de Hermosillo. Por una parte se muestran mapas que permiten ubicar a las AGEB que tienen la mayor cantidad de unidades económicas (Figura 4.6), los cuales se esperaría que sean los que potencialmente atraigan a la población por servicios o trabajo. Por otro lado se muestra el número de personas que

reside en cada AGEB (Figura 4.7), con lo cual es posible ubicar aquellas AGEB con mayor cantidad de pobladores y se esperaría que estos sean los que potencialmente aporten más personas en la dinámica poblacional de la ciudad. Por último se muestra la cantidad de persona que entran y salen de cada AGEB (Figura 4.8) según las redes construidas de acuerdo a la Sección 4.3, en donde se observa que las AGEB que más sobresalen son congruentes con lo esperado por la cantidad de pobladores y la cantidad de unidades económicas. Por ello es posible asegurar que las redes de movilidad construidas según lo mostrado en la Sección 4.3 muestran, razonablemente bien, la movilidad de la población en la ciudad de Hermosillo.

Con la información que se obtiene del algoritmo mostrado en la Sección 4.3 (Figura 4.3), se pueden construir gráficas dirigidas a las cuales también se les podría aplicar el ATD, con la diferencia de que se tendrían que construir los complejos simpliciales de cada red con diferente criterio al utilizado para el complejo clique utilizado en las gráficas no dirigidas y también se tendrían que utilizar diferentes filtraciones para el estudio de homología persistente, ya que las filtraciones de VRIN y la de Vietoris-Rips recíproca no toman en cuenta la posibilidad de que las aristas tengan dirección.

Independientemente de si se toma a las gráficas dirigidas o a las no dirigidas, es importante resaltar que se pueden estudiar muchos fenómenos interesantes propios de la teoría de gráficas con las redes de movilidad construidas, por ello quizá también hay que complementar estudios como éste con los análisis provenientes de teoría de gráficas.

De las Figuras 4.1 mostrada en el Sección 4.2 se puede observar como el número de activaciones cae notablemente para la semanas 19/10-25/10, 26/10-1/11 y 2/11-8/11 impactando directamente el peso de las aristas de las redes correspondientes a estas semanas, ya que al haber menor número de activaciones se esperaría que se registren menos saltos y por lo tanto parecería que la población es menos activa aunque no necesariamente sea así. La diferencia en el número de activaciones (Figura 4.1a) es

la principal culpable de que se tengan que transformar los pesos cuando se utilizan las filtraciones de VRIN y de Vietoris-Rips recíproca, ya que como se mencionó en la Sección 5.2, al utilizar estas filtraciones para generar los diagramas de persistencia se busca hacer una comparación justa entre los diagramas de cada semana y por lo tanto, contrarrestar la diferencia en las activaciones de cada semana. No obstante, es necesario mencionar que sólo la filtración de VRIN normaliza los pesos de cada arista en el proceso de transformación y por ello se esperaría que los efectos de la reducción en el volumen de activaciones sea mucho menor que en la filtración de Vietoris-Rips recíproca.

Ya se había mencionado en la Sección 5.3 que las componentes conexas de una red de movilidad muestran como las AGEB se agrupan de acuerdo al valor de los pesos en las aristas que los conectan. Es importante resaltar que conocer esta información en el contexto de la pandemia ayudaría a planificar cuarentenas dentro de la ciudad pero sin necesidad de suspender toda actividad en ella. Lo anterior se podría hacer al aislar alguna AGEB con alto riesgo biológico y a las AGEB con mayor intercambio de personas con éste (ya que también serían potencialmente un riesgo), pero manteniendo la actividad (con algunas medidas restrictivas) en el resto de las AGEB. En este sentido, todas las AGEB que forman los hoyos de dimensión uno serían fuertes candidatos a entrar en cuarentena, sí alguno de ellos tiene alto riesgo biológico.

Por otro lado, a medida que se agregan las aristas con menor flujo de personas a la red, las componentes conexas y los hoyos de dimensión 1 son menos estables, apareciendo y desapareciendo, lo que hace difícil la tarea de identificar aquellos grupos de AGEB que están estrechamente relacionados entre sí. Es por ello que sería bueno explorar alguna metodología que permita agregar bandas de confianza a los diagramas de persistencia (Brittany Terese Fasy, 2014; Bubenik, 2015; Chazal, Cohen-Steiner, y Mérigot, 2011) y así poder determinar que características topológicas realmente persisten por estar generadas a partir de AGEB estrechamente relacionados por el alto volumen de personas que comparten entre sí.

De las comparaciones realizadas en las Secciones 5.4 y 5.5, se puede concluir que tomar en cuenta semanas desplazadas por un día es buena idea para observar mejor las diferencias en la movilidad de la ciudad, otorgando información más inmediata que podría utilizarse para tomar acciones de control a conveniencia la movilidad de la ciudad.

De las comparaciones realizadas al utilizar la filtración de VRIN en las Secciones 5.4 y 5.5, se puede observar que la distancia de cuello de botella nos otorga una comparación valiosa al tomar en cuenta únicamente las componentes conexas, pero no sucede lo mismo con la comparación realizada al tomar en cuenta unicamente los hoyos de dimensión 1. Con la distancia de Wasserstein podemos ver que para ambas características toplógicas se puede identificar cómo ciertas semanas difieren entre sí de manera clara.

Para las comparaciones realizadas con la filtración de Vietoris-Rips recíproca, podemos observar que la distancia de cuello de botella utilizada en las persistencias de ambas características topológicas no otorga mucha información, ya que salvo algunas semanas como excepción, las diferencias parecieran ser iguales sin importar que semanas se comparen. Aquí de nueva cuenta se puede observar que la distancia de Wasserstein otorga diferencias claras y bien marcadas entre ciertas semanas, no obstante podemos ver que las diferencias se acentúan sólo en las semanas donde se ha tenido bajo número de activaciones, lo cual nos podría sugerir que la filtración de Vietoris-Rips recíproca no es lo suficientemente justa como para utilizarla en la comparación de diagramas de persistencia, sin embargo aún podría otorgar información valiosa sobre aquellas características topológicas que en verdad persisten en el tiempo.

Sin duda alguna las comparaciones realizadas entre los diagramas de persistencia obtenidos con la filtración de VRIN, muestran cierta evolución en la actividad de la movilidad de la ciudad, acentuándose las diferencias entre cierto conjunto de semanas

y pareciendo casi nula en otras. Uno de los trabajos futuros que este trabajo puede apuntar a realizar, es el de ligar la movilidad con fenómenos como el infeccioso y utilizar las diferencias entre la movilidad semanal para poder predecir el aumento de casos de Covid-19 en la ciudad.

Referencias

Aktas Mehmet E., A. E., y El, F. A. (2019). Persistence homology of networks: methods and applications. *Applied Network Science*, *4*(61).

Baer, R. (1966). *Linear algebra and projeceive geomeery.* ACADEMIC PRESS INC. (LONDON) LTD.

Blumberg, R. R. . A. J. (2020). *Topological data analysis for genomics and evolution.* Cambridge University Press.

Brittany Terese Fasy, A. R. L. W. S. B. A. S., Fabrizio Lecci. (2014, 03). Confidence sets for persistence diagrams. *The Annals of Statistics*, *42*, 2301-2339.

Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.*, *16*(1), 77–102.

Calabrese, F., Ferrari, L., y Blondel, V. (2014, 01). Urban sensing using mobile phone network data: A survey of research. *ACM Computing Surveys*, *47*, 1-20. doi: 10.1145/2655691

Chazal, F., Cohen-Steiner, D., y Mérigot, Q. (2011). Geometric inference for probability measures. *Foundations of Computational Mathematics*, *11*(6), 733–751.

Clough, J., Byrne, N., Oksuz, I., Zimmer, V., Schnabel, J., y King, A. (2020, 09). A topological loss function for deep-learning based image segmentation using persistent homology. *IEEE transactions on pattern analysis and machine inte-*

lligence., *PP*.

Deo, S. (2018). *Algebraic topology a primer*. Springer.

Dey, T. K., y Mandal, S. (2018). Protein Classification with Improved Topological Data Analysis. En L. Parida y E. Ukkonen (Eds.), *18th international workshop on algorithms in bioinformatics (wabi 2018)* (Vol. 113, pp. 6:1–6:13). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. Descargado de http://drops.dagstuhl.de/opus/volltexte/2018/9308 doi: 10.4230/LIPIcs.WABI.2018.6

Dongjin, L., Bresten, C., Youm, K., Seo, K.-W., y Jung, J.-H. (2020, 06). Model discrepancy of earth polar motion using topological data analysis and convolutional neural network analysis. *International Journal of Modern Physics C*, *31*.

Espinoza, J., Hernández-Amador, R., Hernandez, H. A., y Ramonetti-Valencia, B. (2019, 12). A numerical approach for the filtered generalized Čech complex. *Algorithms*, *13*, 11. doi: 10.3390/a13010011

Gidea, M., y Katz, Y. (2018). Topological data analysis of financial time series: Landscapes of crashes. *Physica A: Statistical Mechanics and its Applications*, *491*, 820-834. Descargado de https://www.sciencedirect.com/science/article/pii/S0378437117309202 doi: https://doi.org/10.1016/j.physa.2017.09.028

Goel, A., Pasricha, P., y Mehra., A. (2020). Topological data analysis in investment decisions. *Expert Systems with Applications*, *147*, 113222. Descargado de https://www.sciencedirect.com/science/article/pii/S0957417420300488 doi: https://doi.org/10.1016/j.eswa.2020.113222

Jean-Daniel Boissonnat, F. C., y Yvinec, M. (2018). *Geometric and topological inference*. Cambridge University Press.

Maletić, S., Rajkovic, M., y Vasiljevic, D. (2008, 06). Simplicial complexes of networks and their statistical properties. En (p. 568-575). doi: 10.1007/978-3-540-69387-1_65

Munkres, J. R. (1984). *Elements of algebraic topology*. Addison-Wesley Publishing

Company, INC.

Muszynski, G., Kashinath, K., Kurlin, V., y Wehner, M. (2019, 02). Topological data analysis and machine learning for recognizing atmospheric river patterns in large climate datasets. *Geoscientific Model Development*, *12*, 613-628.

Nicolau, M., Levine, A. J., y Carlssona, G. (2011, 04). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(17), 7265–7270. doi: 10.1073/pnas.1102826108

Nielson, J., Paquette, J., Liu, A., Guandique, C., Tovar, C., Inoue, T., ... Ferguson, A. (2015, 10). Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nature Communications*, *6*, 8581.

Ofori-Boateng, D., Lee, H., Gorski, K., Garay, M., y Gel, Y. (2021, 06). Application of topological data analysis to multi-resolution matching of aerosol optical depth maps. *Frontiers in Environmental Science*, *9*, 684716.

Oyama, A., Hiraoka, Y., Obayashi, I., Saikawa, Y., Furui, S., Shiraishi, K., ... Kotoku, J. (2019, 06). Hepatic tumor classification using texture and topology analysis of non-contrast-enhanced three-dimensional t1-weighted mr images with a radiomics approach. *Scientific Reports.*, *9*.

Rudin, W. (1980). *Principios de análisis matemático*. McGRAW-HILL.

Smith, A. D., Dłotko, P., y Zavala, V. M. (2021). Topological data analysis: Concepts, computation, and applications in chemical engineering. *Computers & Chemical Engineering*, *146*, 107202. Descargado de https://www.sciencedirect.com/science/article/pii/S009813542031245X doi: https://doi.org/10.1016/j.compchemeng.2020.107202

Sørensen, S. S., Biscio, C. A. N., Bauchy, M., Fajstrup, L., y Smedskjaer, M. M. (2020.). Revealing hidden medium-range order in amorphous materials using topological data analysis. *Science Advances*, *6*(37), eabc2320.

Tizzoni, M., Bajardi, P., Decuyper, A., King, G. K. K., Schneider, C. M., Blondel, V., ... Colizza, V. (2014). On the use of human mobility proxies for modeling epidemics. *PLoS Comput Biol*, *10*(7), e1003716.

Wang, Y., Ombao, H., y Chung, M. K. (2019). Statistical persistent homology of brain signals. En *Icassp 2019 - 2019 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 1125-1129). doi: 10.1109/ICASSP.2019.8682978

West, D. B. (2001). *Introduction to graph theory.* Pearson Education, Inc.

Zomorodian, A. (2010). Fast construction of the vietoris-rips complex. *Comput. Graph.*, *34*, 263-271.