Universidad de Guanajuato

Aspectos estadísticos en análisis topológico de datos y una aplicación en ecología

Tesis

que para obtener el título de

licenciado en matemáticas

presenta:

VIDAL ALÍ GONZÁLEZ CUCURACHI

director de tesis:

DR. MIGUEL NAKAMURA SAVOY

Guanajuato, Gto. 2016

UNIVERSIDAD DE GUANAJUATO

En la ciudad de Guanajuato, Gto. a las 17:00 (diecisiete horas y cero minutos) del día 01 de diciembre del año 2016, se reunieron en el (la) Salón de Actos del Departamento de Matemáticas de la División de Ciencias Naturales y Exactas del campus Guanajuato de la Universidad de Guanajuato, los señores Doctora Eloísa Díaz Francés Murguía, Doctor Rogelio Ramos Quiroga, Doctor Fermín Omar Reveles Gurrola designados para verificar el examen de titulación del señor **VIDAL ALÍ GONZÁLEZ CUCURACHI**, concedido por acuerdo de la Rectoría General a efecto de obtener el título de **LICENCIADO EN MATEMÁTICAS**.

Instalado el jurado bajo la Presidencia de la señorita Doctora Eloísa Díaz Francés Murguía y fungiendo como Secretario el señor Doctor Rogelio Ramos Quiroga, dio principio el examen e interrogaron sucesivamente los tres sinodales sobre diversas materias y aspectos del trabajo de titulación y, terminado el interrogatorio, se procedió a verificar la votación secreta, habiendo resultado el señor **VIDAL ALÍ GONZÁLEZ CUCURACHI**.

APROBADO POR UNANIMIDAD

Para constancia se levanta la presente acta por triplicado que firman los señores sinodales.

Presidente

Dra. Eloísa Díaz Francés Murguía

Secretario

Dr. Rogelio Ramos Quiroga

Vocal

Dr. Fermín Omar Reveles Gurrola

UNIVERSIDAD DE GUANAJUATO
DIRECCIÓN DE ASUNTOS ACADÉMICOS

NO ES VALIDO SI LLEVA RASPADURAS O ENMENDADURAS.

UNIVERSIDAD DE GUANAJUATO

Dirección de Asuntos Académicos

El suscrito, ***Maestro Francisco Javier Pérez Arredondo, Director de Asuntos Académicos****, hace constar que la presente acta constituye formalmente la realización del examen de obtención* ***de Título de LICENCIADO EN MATEMÁTICAS,*** *en los términos del Artículo 68 del Estatuto Académico de la Universidad de Guanajuato.*

UNIVERSIDAD DE GUANAJUATO
DIRECCIÓN DE ASUNTOS ACADÉMICOS

Agradecimientos

Agradezco primeramente a mis padres, Ana Lilia y Vidal, a quienes debo la formación de la persona que ahora soy. Gracias porque siempre han estado a mi lado, brindándome todo el apoyo en todas las etapas de mi vida. Este trabajo se los dedico, papás. Gracias a mi hermana Chuy que siempre ha estado cerca de mí para apoyarnos mutuamente. Gracias a Natalia, ¡por ser maravillosa! Y llenar toda mi vida de felicidad. Sólo tú conoces el camino recorrido detrás de esta tesis, gracias por recorrerlo conmigo.

Gracias a mi asesor, el Dr. Miguel Nakamura Savoy, por brindarme toda su paciencia y compartirme de su conocimiento y experiencias. Muchas gracias por no sólo instruirme de la mejor manera en este trabajo, sino también por el apoyo moral y los consejos que siempre me dio, tanto en lo académico como en lo personal. Agradezco también a mis sinodales, por sus valiosas aportaciones y comentarios a este escrito. En particular, agradezco a la Dra. Eloísa Díaz-Francés Murguía porque sus cursos y la experiencia como su ayudante forman parte fundamental de mi desarrollo como futuro estadístico. Agradezco al Dr. Enrique Martínez Meyer y a la M. C. Edith Calixto Pérez por facilitarnos los datos utilizados en el análisis de nichos y ambientes disponibles, así como su magnífico asesoramiento en el ámbito de ecología.

Agradezco a mis amigos, que a lo largo de seis años en Guanajuato hemos compartido un montón de experiencias. Gracias a la comunidad DEMAT-CIMAT, por acogerme en un ambiente de compañerismo y amistad, y así disfrutar de la mejor manera del aprendizaje de las matemáticas. Por último, agradezco al CIMAT y al CONACYT por los apoyos económicos que me otorgó durante la licenciatura.

Prefacio

Como lo describen en Boissonnat y Teillaud (2004), la topología es un campo de las matemáticas que se encarga de estudiar la forma, es decir, la estructura geométrica de un espacio u objeto. Está principalmente interesada en las propiedades de estos entes que se preservan bajo ciertas deformaciones, tales como expansión, traslación y contracción, pero sin cortar o pegar nuevos pedazos. Estas propiedades que se preservan se les conoce con el nombre de propiedades invariantes. La idea es que, si dos objetos poseen las mismas propiedades invariantes, entonces los objetos son equivalentes en cuanto a su estructura. En contraste, existen también características que, aunque sí varían entre los objetos, son irrelevantes, tales como su posición en el espacio, la distancia entre dos de sus puntos o la curvatura de su superficie. Son irrelevantes en la medida en que no aportan información sobre la estructura general que tienen los objetos. Esta tesis versará sobre ciertas técnicas modernas diseñadas para estudiar algunas propiedades invariantes en una nube de datos.

Para ilustrar estas propiedades invariantes, notemos que podemos comenzar con una esfera que está vacía por dentro, achatarla y convertirla en un icosaedro. Aunque aparentemente la esfera cambió de forma, hay propiedades que no cambian. El espacio vacío dentro de la esfera se mantiene intacto y el objeto sigue contando con una sola cáscara. Espacio y cáscara constituyen dos ejemplos de propiedades invariantes. Más adelante definiremos formalmente el concepto de agujero y componente conexa, propiedades invariantes que serán de interés a considerar.

Una de las ramas en las que se divide la topología es la topología computacional. Esta aborda la simplificación de los problemas topológicos complejos y el desarrollo de algoritmos eficientes para su solución, en caso de que estos problemas sean tratables. Dichos algoritmos sólo pueden trabajar con espacios y objetos que tienen una representación finita y es por eso que nos restringiremos al estudio de objetos con estructuras finitas y funciones entre estos. Más adelante se establecerá que

surge de manera natural el concepto de complejo simplicial para abordar este caso. Dentro de esta disciplina nos encontramos que la topología algebraica forma parte importante en el desarrollo de estos algoritmos, pues con ella se desarrolla gran parte del sustento teórico matemático que sirve de respaldo. Más específicamente dentro de lo que es el análisis topológico de datos (*TDA* por sus siglas en inglés), la homología es quien toma un papel primordial; de manera genérica, la homología de un objeto de dimensión n describe sus agujeros $k-$dimensionales (que definiremos más adelante), donde k va de 0 a n.

Actualmente, dentro del análisis clásico de datos multivariados se pueden encontrar técnicas para inferir propiedades de una colección datos, las cuales están basadas principalmente en proyecciones del espacio original a espacios de menor dimensión. Dentro de estas técnicas se encuentran el análisis por componentes principales y el método de regresión lineal, o para analizar el agrupamiento dentro de los datos como lo hace el análisis de conglomerados. Estas técnicas ya han sido ampliamente estudiadas y aceptadas para el análisis de datos multivariados, tal como lo describe Seber (1984). Una limitación de estas técnicas surge cuando el subconjunto que representa la nube de datos dentro del espacio $n-$dimensional está muy torcido (por así decirlo) o lleno de burbujas. Al tratar de proyectar este subconjunto en un hiperplano de dimensión menor, se tiene pérdida de información, debido a que estas peculiaridades estructurales pasan desapercibidas.

De manera general, el análisis topológico es una técnica que busca describir la estructura de los datos vistos como un objeto geométrico. Toma los datos, los representa mediante un conjunto de estructuras matemáticas llamadas complejos simpliciales, para luego estudiar de éstos su homología. Como veremos en el siguiente capítulo, la homología es un concepto útil para describir las partes o pedazos y los agujeros o huecos de un espacio. Una descripción de la homología constituye una recopilación de la información sobre la estructura geométrica del objeto, noción conocida como resumen topológico

Usualmente los datos están representados en un espacio $n-$dimensional. Su estructura geométrica global dentro de este espacio puede revelar características relacionadas directamente con el fenómeno que se estudia. La existencia de una estructura compleja en una colección de datos en algunos casos denota la presencia de características diferentes respecto a una estructura nada o poco compleja. Diversas aplicaciones de esta técnica han sido presentadas en varias ramas de la

ciencia. En Nicolau et al. (2010) se muestra la forma en que se utilizó TDA para encontrar un subgrupo de cánceres de mama que tienen un perfil de mutación única y con un nivel de supervivencia alto. En Adams y Carlsson (2014) se considera un problema de evasión de redes de sensores, dando un enfoque vía análisis topológico para su tratamiento. Por otro lado, en Chan et al. (2013) se estudia la manera de aplicar la técnica topológica para la reconstrucción del proceso evolutivo de ciertas especies, demostrando la existencia de lo que se define como evolución horizontal, contrastando la hipótesis de que sólo existe evolución vertical, es decir, transferencia genética de padres a hijos solamente. Así también, existen algunos otros campos en los que se ha aplicado y sigue estudiando actualmente, pues es un área emergente de investigación.

Nuestro objetivo, dentro del enfoque topológico, es descubrir una propiedad estructural desconocida a partir de una base de datos, lo que se traduce en un problema de inferencia. A partir de una nube de datos, buscamos describir su estructura como objeto geométrico obteniendo su resumen topológico. En este proceso nos encontramos con asuntos estadísticos a ser tomados en cuenta como lo son el ruido o aleatoriedad dentro de los datos, y la incertidumbre en la estructura obtenida por vía del resumen topológico.

Este problema plantea nociones formales de estadística matemática. Por un lado, hablaremos sobre una técnica para pruebas de hipótesis utilizando TDA para comparar datos provenientes de distintos objetos y por otro, conjuntos de confianza para diagramas de persistencia, que es el principal resumen topológico utilizado en la práctica. Con los conjuntos de confianza reconocemos posible ruido dentro del diagrama de persistencia, identificando posible información que no es significativa en cuanto a la verdadera estructura del objeto. La idea principal es definir una distancia entre diagramas de persistencia y obtener un cuantil para la distancia entre el diagrama de persistencia estimado y el verdadero. La diferencia entre los métodos se da principalmente por cómo tratar a dicha distancia y cómo podemos acotarla. Los métodos que detallaremos en el Capítulo 2 se basan en acotar dicha distancia con la distancia de Hausdorff entre la muestra y el subespacio del cual suponemos fue recabada. Esta distancia la definiremos más adelante.

Por otro lado, planteamos la aplicación de la técnica a ciertos datos originados en un contexto en ecología. Estos tienen que ver con los conceptos de distribución de especies, riqueza de especies, nicho ecológico y ambientes disponibles. Buscamos entender si existe relación entre la estructura

topológica del nicho de una especie y su tolerancia ambiental, es decir, qué tanto una especie puede subsistir al variar ambientes. También se quiere estudiar la relación entre el concepto de riqueza de especies en un espacio geográfico y la estructura topológica de su ambiente disponible.

La distribución de una especie se refiere al área geográfica en dónde se dispersan los grupos de individuos de la especie. Esto es directamente un subconjunto del área geográfica global. Por otro lado, en términos generales el nicho ecológico de una especie es el conjunto de características ambientales que le permiten subsistir. A lo largo de la historia, los ecólogos han tenido una discusión sobre los aspectos que debe englobar el término, la cual explicaremos con detalle en el Capítulo 3. El objetivo es evaluar si TDA tiene información relevante que aportar respecto a este tema.

Teniendo una especie y d variables ambientales $X_1, X_2, \ldots, X_d$ que los biólogos consideran son pertinentes, se tiene que el nicho de la especie está dado por un subconjunto N de $\mathbb{R}^d$, en el cual cada punto es un vector de características que le son favorables. Lo que nos interesa es estudiar N como subespacio topológico de $\mathbb{R}^d$ y ver si su complejidad geométrica está relacionada con la idiosincrasia de la especie en cuestión. Ahora, dada una región geográfica G podemos identificar cada punto $x \in G$ con el vector $(x_1, x_2, \ldots, x_d)$ que representa el valor de cada X_j en x. A la imagen de G bajo esta identificación E se le conoce como el ambiente disponible de la región G. Estos conceptos se ilustran en la Figura 1.

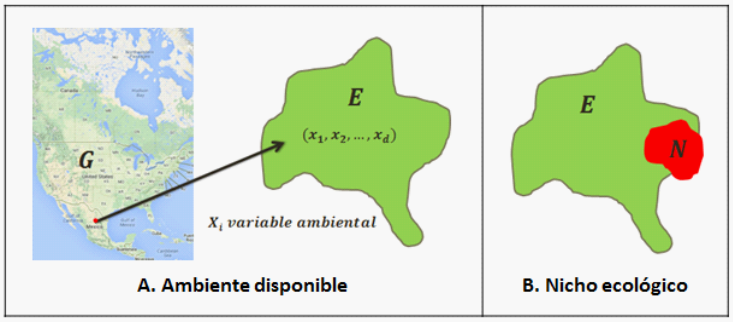

Figura 1. Ambiente disponible y nicho ecológico

Con base en los datos lo que nos interesa estudiar, ya sea un nicho o un ambiente disponible, es el subconjunto $d-$dimensional que representa. Se comienza con una nube de datos que suponemos está muestreada en el espacio original y con base en esta nube queremos inferir sobre la estructura

topológica del objeto muestreado. Aunque en el Capítulo 3 ahondaremos un poco más sobre estas definiciones y la importancia de su estudio, para mayor detalle sobre nichos y distribuciones de especies el lector puede consultar Peterson et al. (2011). En este libro además se muestra la manera tradicional de estudiar e inferir nichos de especies, combinando los puntos de vista ecológico y estadístico.

Esta tesis busca explicar de manera introductoria lo que es el análisis topológico de datos. Dentro de este enfoque surgen naturalmente aspectos estadísticos importantes de los cuales abordaremos dos de ellos. Por un lado, trabajaremos un enfoque vía TDA de pruebas de hipótesis, las cuales están principalmente diseñadas para la clasificaciones de objetos. Como ya hemos comentado, el análisis topológico busca estudiar la estructura topológica del espacio de donde están muestreados los datos. Comenzamos considerando un grupo de siluetas, imágenes o de conjuntos de puntos. Nuestro objetivo es, haciendo uso del resumen topológico que proporciona TDA, evaluar la veracidad de la hipótesis de que representan el mismo objeto.

Por otro lado, en el proceso del análisis topológico naturalmente se tiene presencia de aleatoriedad, referida en sí a la muestra que tenemos y a la estimación de los resultados. Se infiere puntualmente en la topología de los datos visto como subconjunto de un espacio topológico. Es por ello que es necesario concebir un conjunto de confianza para esta estimación.

Por último, se presentarán casos de estudio de cierto problema surgido en la disciplina de ecología. El problema tiene que ver con espacios de alta dimensionalidad y se busca aplicar la técnica topológica con el objetivo de ver si ésta es informativa.

La manera en que está estructurada esta tesis es la siguiente: el Capítulo 1 contiene una introducción a homología y homología persistente, siendo esta última la técnica principal del análisis topológico. Mostraremos una opción de software para implementar la técnica e ilustraremos en particular un problema estadístico que surge en el proceso, que es la necesidad de construir un conjunto de confianza para el resultado estimado: el diagrama de persistencia, de lo cual hablaremos con mayor detalle en el Capítulo 2.

En el Capítulo 2 describiremos las soluciones a este problema presentadas en Fasy et al. (2014) y su implementación en R. En este artículo se presentan cuatro métodos para generar un cuantil

para la distancia (de cuello de botella) entre el diagrama de persistencia estimado y el diagrama de persistencia verdadero. Con base en esto se construye un conjunto de confianza para la distancia, eliminando así posibles características que surgen en el resumen topológico estimado pero que en realidad no están en la estructura real. Además, detallaremos un enfoque visto en Robinson y Turner (2013), el cual propone utilizar TDA para pruebas de hipótesis principalmente destinadas a clasificación de objetos. Teniendo varios conjuntos de datos queremos probar la hipótesis de que fueron muestreados del mismo subconjunto. Haciendo uso de la estructura topológica del objeto en cuestión, se presenta una manera de hacerlo mediante pruebas de permutación en los diagramas de persistencia de cada conjunto de datos.

En el Capítulo 3 primero ejemplificaremos la importancia del estudio de nichos ecológicos y la inquietud de estudiar dos ambientes disponibles como objetos topológicos para su comparación en el sentido de abundancia y diversidad de especies. Presentaremos los datos proporcionados por el Dr. Enrique Martínez Meyer y la M. en C. Edith Calixto, el motivo por los cuales fueron elegidos y las consideraciones que fueron tomadas para el análisis. Estos comprenden los nichos de siete especies que tienen su hábitat principalmente en regiones que van desde la parte sur y sureste de México hasta la parte norte de América del Sur. También, tomamos en consideración los ambientes disponibles de las regiones geográficas de India y Australia para compararlos desde el punto de vista topológico. El principal objetivo es ver si dicho enfoque tiene algo que decir en el sentido de cierta clasificación de especies, la cual tiene que ver con la tolerancia a cambios de ambiente, y en el de riqueza y abundancia de especies en el caso de los ambientes disponibles.

Ahora bien, comenzaremos el Capítulo 1 dando una motivación a cerca de lo que es la homología para después hablar de la técnica principal de TDA que es la homología persistente. Nos introduciremos un poco en lo que son los aspectos técnicos, lo suficiente para entender los aspectos estadísticos que trataremos con detalle en el Capítulo 2.

Contenido

Capítulo 1

Análisis topológico de datos (TDA)

1.1. Descripción

Como ya hemos estipulado, dada una nube de datos, formalizada como un subconjunto de puntos de un subespacio de $\mathbb{R}^d$, o de manera más general, de un subespacio de un espacio métrico completo, el análisis topológico busca describir la estructura de dicho subespacio. Vamos a suponer que los datos que conforman la nube se encuentran muestreados con cierta distribución P y que el objetivo es describir el soporte de P. Esta distribución de probabilidad es la noción principal que permitirá tender un puente entre el punto de vista topológico y el estadístico, en términos de plantear matemáticamente un problema de inferencia. Cabe notar que nuestra meta no es precisamente inferir el soporte mismo, sino más bien inferir alguna característica de su topología. Este objetivo se plantea con la intención de entender el fenómeno bajo estudio a través de la estructura topológica de una observación empírica del mismo.

Por estructura nos referiremos directamente a lo que nuestra intuición indica: la forma de un objeto visto como cuerpo geométrico. En efecto, como se ilustró en la introducción, la complejidad topológica puede ser un factor importante. Matemáticamente hablando y con mayor precisión, lo que vamos a estudiar de la nube datos es su *homología* como espacio topológico. La homología es un procedimiento contenido en la topología algebraica que nos permite investigar los *agujeros* de un espacio. Aunque no hemos definido formalmente el término agujero en el sentido topológico, intuitivamente es lo que podríamos imaginar en el sentido coloquial. Sin embargo, sí haremos una

distinción en cuestión de la *dimensión* de los agujeros. Dado un subconjunto $X \subset \mathbb{R}^d$, los agujeros de dimensión 0 los identificaremos como las componentes conexas o pedazos de los que se conforma X; un agujero de dimensión 1 es el espacio vacío que podemos rodear con una circunferencia a lo largo de X; el de dimensión 2 es el espacio vacío que podemos cubrir con una esfera. Para dimensiones mayores $k \geq 3$, intuitivamente vemos el agujero de dimensión k como el espacio vacío que podemos envolver con S^k, la esfera de dimensión k. Como describiremos, estos agujeros constituyen las características en la topología de los datos que TDA examina.

Consideremos los siguientes ejemplos ilustrativos sencillos para explicar estas nociones de agujeros. En la Figura 2 en la parte izquierda se ilustra un objeto de dimensión 2. Tiene 2 componentes conexas o pedazos (agujeros de dimensión 0) y 3 agujeros A, B y C de dimensión 1. En cambio, en la parte derecha de la Figura 2 hay un objeto en $\mathbb{R}^3$, un "cascarón agujerado", que tiene una sola componente conexa, 2 agujeros A y B de dimensión 1 y ningún agujero de dimensión 2. Algunas de las técnicas del análisis topológico resumen esta estructura de agujeros a través de ciertos instrumentos y para ellos será importante destacar que un objeto en un espacio de dimensión d puede tener agujeros de dimensión k entre 0 y $d-1$. Aunque es difícil establecerlo de manera general, esto se visualiza en los ejemplos que acabamos en enlistar.

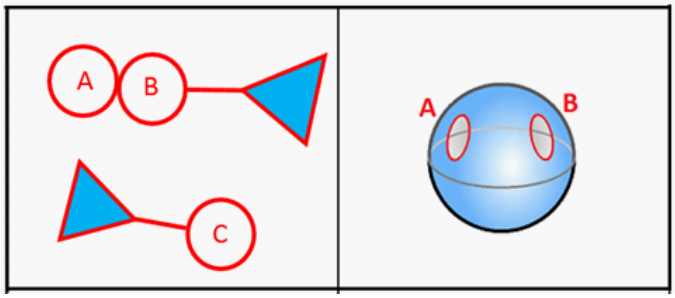

Figura 2. Ejemplos ilustrativos de agujeros de dimensión 0, 1 y 2 en el contexto de TDA.

La homología lo que nos describe es justamente los agujeros $k-$dimensionales de un subespacio $d-$dimensional, donde $k \leq d-1$. Vamos a pensar en la complejidad topológica de un subespacio como la cantidad de agujeros que tiene. Intuitivamente un subespacio que es conexo y tiene pocos agujeros tiene una estructura menos compleja que uno que está dividido en pedazos y lleno de agujeros.

Dentro de este contexto, TDA busca explicar la complejidad topológica de una nube de datos. Las ventajas de esta técnica se encuentran principalmente situadas en el hecho de que estudiamos a los datos dentro de su espacio original, sin necesidad de reducir dimensiones y siempre y cuando podamos representar nuestros datos dentro de un espacio métrico, será posible analizarlos con el procedimiento. Por otro lado, debemos tomar en cuenta que el análisis topológico es una herramienta exploratoria complementaria y no sustituye las técnicas de análisis multivariado existentes actualmente. Éste se enfoca específicamente en la topología del soporte de la distribución con la que suponemos fueron muestreados los datos y no es predictivo. También se agrega que computacionalmente en un sentido general no es sencillo de implementar. Conforme expliquemos con mayor detalle la técnica tendremos más claro estas ventajas y desventajas.

Daremos a continuación una introducción a lo que es homología simplicial y más específicamente a homología persistente principalmente expuesta de la manera en cómo se presentan dichos temas en Harer y Edelsbrunner (2010), en Hatcher (2002) y en Espinoza (2015). Se recomienda al lector consultar dichas referencias para obtener mayor detalle sobre el tema, para lo cual no es necesario tener un conocimiento profundo en topología. Además, Harer y Edelsbrunner muestran una implementación computacional de la homología persistente para complejos simpliciales. Como veremos, estas últimas estructuras serán esenciales para el estudio de la homología de un subespacio basándonos en una muestra de puntos sobre una vecindad.

1.2. Antecedentes técnicos

1.2.1. Definiciones básicas

Primero consideremos algunas definiciones para entender lo que es la homología simplicial, para después exponer cómo estudiar la homología de un objeto a partir de una muestra de puntos. Como mencionábamos antes, en TDA nos restringimos a espacios u objetos que tienen una representación finita. Es por esto que es necesario introducir los conceptos de simplejo y complejo simplicial, estructuras con las que relacionamos a los datos para estudiar su homología.

Comenzaremos dando un poco de intuición de lo que son los complejos simpliciales y por qué son importantes. Notemos que por ejemplo el toro (T), el plano proyectivo ($\mathbb{R}P^2$) y la botella de

Klein (K) pueden ser obtenidos a partir de un cuadrado identificando los lados opuestos como se muestra en la Figura 3. Si trazamos la diagonal del cuadrado justo como se muestra en las figuras, tenemos que las superficies mencionadas también se pueden producir a partir de triángulos pegados mediante sus lados. De manera similar ocurre con cualquier polígono regular: podemos trazar algunas diagonales y así obtener la superficie original mediante el pegado de varios triángulos. De hecho, si pensamos en cualquier superficie cerrada, podemos obtenerla mediante el pegado de dichas figuras de una manera similar a la que mencionamos. La idea de un complejo simplicial es generalizar este "pegado de triángulos" a dimensiones mayores que 2. Un simplejo es justamente la generalización del triángulo a cualquier dimensión. En un sentido coloquial un complejo simplicial es un conjunto de simplejos (triángulos) pegados mediante sus caras. A continuación se definirán de una manera más formal estos conceptos.

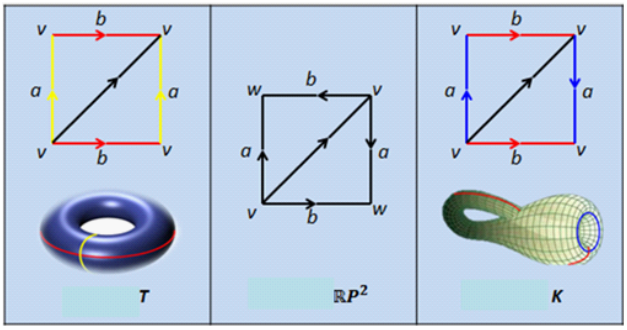

Figura 3. $T, \mathbb{R}P^2$ y K a partir de triángulos.

Simplejos

Definición 1.2.1. *Sean $u_0, u_1, ..., u_k$ puntos en $\mathbb{R}^d$. Se dice que $u_0, u_1, ..., u_k$ son **independientes** si los vectores $u_1 - u_0, u_2 - u_0, ..., u_k - u_0$ son linealmente independientes.*

Notemos que no es necesario establecer un orden en particular para que $k+1$ puntos sean independientes. Con esto, la propiedad de independencia no depende de la elección de u_0.

Definición 1.2.2. *Sea $U = \{u_0, u_1, ..., u_k\} \subset \mathbb{R}^d$. Decimos que x es **combinación convexa** de los elementos de U si existen constantes $\alpha_j \geq 0$ tales que*

a) $x = \sum_{j=0}^{k} \alpha_j u_j$ *y*

b) $\sum_{j=0}^{k} \alpha_j = 1$.

Al conjunto de todas las combinaciones convexas de U se conoce como *envolvente convexa*.

Definición 1.2.3. *Definimos el* $k-$***simplejo***, $\sigma(u_0, u_1, ..., u_k)$, *generado por* $U = \{u_0, u_1, ..., u_k\} \subset \mathbb{R}^d$ *puntos independientes, como la envolvente convexa del conjunto* U. *A los puntos* $u_0, u_1, ..., u_k$ *se les conoce como* ***vértices*** *del simplejo.*

Definimos una $m-$***cara*** *del simplejo* $\sigma(u_0, u_1, ..., u_k)$ *como la envolvente convexa de cualquier subconjunto no vacío* $\{v_0, v_1, ..., v_m\} \subseteq \{u_0, u_1, ..., u_k\}$, *que a su vez es un simplejo de dimensión* $m \leq k$. *Además, decimos que es una cara propia si el subconjunto es propio.*

Como se explica con mayor detalle en Espinoza (2015), dos simplejos que tienen la misma dimensión son *homeomorfos* (con las mismas propiedades topológicas, en algún sentido que definiremos más adelante). Es por esto que comúnmente para realizar las construcciones que se describirán a continuación, se consideran los *simplejos estándar*. El $k-$simplejo estándar es el que se obtiene considerando como vértices al cero y a la base canónica de $\mathbb{R}^k$, $\sigma(0, e_1, ..., e_k)$. Para k pequeño podemos visualizar de manera clara los simplejos de dimensión k. Para $k = 0$ se tiene un vértice, para $k = 1$ tenemos una recta, para $k = 2$ tenemos un triángulo, para $k = 3$ tenemos un tetraedro, y así sucesivamente; ver Figura 4.

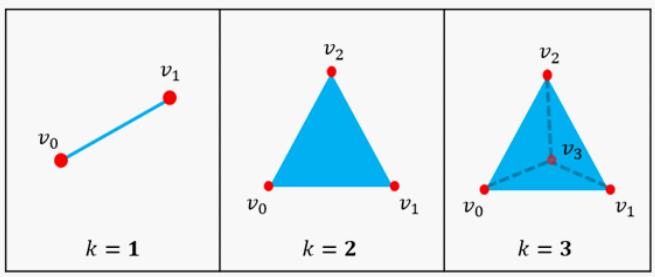

Figura 4. Simplejos de dimensión 1, 2 y 3.

Complejos simpliciales

Las estructuras que definiremos a continuación se utilizan ampliamente en topología computacional. Estas permiten asociar a una nube de puntos un espacio topológico. Además, de que dichas estructuras proporcionan un lenguaje idóneo para trabajar computacionalmente con espacios topológicos.

Definición 1.2.4. *Un complejo simplicial es un conjunto finito de simplejos K tal que:*

a) Si $\sigma \in K$ y τ es una cara de σ entonces $\tau \in K$.

b) Si $\sigma_1, \sigma_2 \in K$ entonces $\sigma_1 \cap \sigma_2$ es una cara de σ_1 y σ_2.

Definimos la *dimensión* de K como la dimensión más grande de los simplejos que contiene. Además, dado un complejo simplicial K, si un subconjunto $K' \subset K$ es a su vez un complejo por sí mismo, entonces decimos que K' es un *subcomplejo* de K. Como ya mencionamos, intuitivamente un complejo simplicial es un conjunto de simplejos que se pegan mediante sus caras. A continuación, en la Figura 5 en a y b se muestran ejemplos de complejos simpliciales, pero no es un complejo simplicial lo que se muestra en la Figura 5 c.

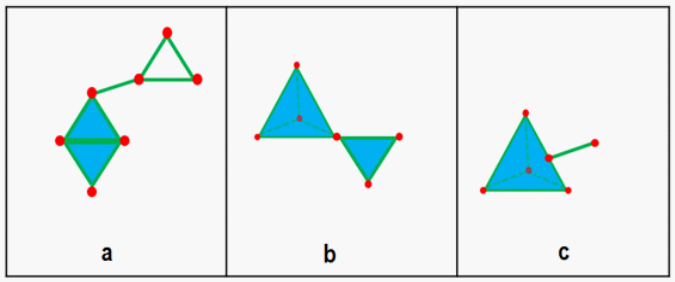

Figura 5. Ejemplos ilustrativos de complejos simpliciales (a y b) y un ejemplo de un conjunto de simplejos que no es complejo simplicial (c).

Dado un complejo simplicial K, entre los subcomplejos de K en los que centramos nuestra atención se encuentran los $j-$*esqueletos*. El $j-$esqueleto de K, $K^{(j)}$, es el conjunto de los simplejos de dimensión menor o igual a j, es decir, $K^{(j)} = \{\sigma \in K : \dim \sigma \leq j\}$. Notemos que el $0-$esqueleto de un complejo es precisamente su conjunto de vértices. Observemos que si $i \leq j$ entonces $K^{(i)} \subset K^{(j)}$; de aquí que la familia de $j-$esqueletos nos permite dar una descomposición

del complejo K a manera de que podamos *construir* el complejo a partir de su conjunto de vértices e ir agregando simplejos de dimensión cada vez mayor. Más adelante definiremos formalmente a lo que nos referimos con dicha construcción, la cual vamos a conocer como *filtración* del complejo K.

1.2.2. Homología

Cadenas simpliciales

Dentro de la topología, un cuestionamiento central es el de si dos espacios son *homeomorfos* entre sí. Dados dos espacios X y Y, decimos que X es homeomorfo a Y, y lo denotamos por $X \simeq Y$, si existe una función $f : X \to Y$ biyectiva tal que f y su inversa f^{-1}son ambas continuas; en ese caso, llamamos a f *homeomorfismo*. Para asegurar que dos espacios son homeomorfos basta con encontrar un homeomorfismo entre ellos. El poder demostrar que no lo son es una tarea un tanto más complicada. El procedimiento usual es encontrar una *propiedad topológica* (o *invariante topológico,* se define como una propiedad que se preserva bajo homeomorfismos) que se cumpla en uno, pero no en el otro. Resulta ser que los agujeros de un espacio son invariantes topológicos, y la homología nos proporciona el lenguaje matemático para describirlos.

Dentro de la teoría de homología, la homología simplicial fue un primer esfuerzo del desarrollo teórico. Como tal, presenta algunas limitaciones en comparación con algunas otras generalizaciones. Referimos al lector a Munkres (1984), en donde se da un panorama general de la teoría de homología. Por su parte, la homología simplicial posee un lenguaje propicio para las aplicaciones en topología computacional. A continuación se darán los elementos teóricos necesarios para definir formalmente la homología simplicial y más adelante la homología persistente.

Aunque ya contamos con la definición de lo que es un complejo simplicial, nos interesa trabajar con los simplejos por los que está conformado. En general un complejo está conformado por simplejos de distintas dimensiones. Es por esto que definiremos estructuras de conjuntos de estos simplejos y una operación para trabajar con ellas.

Definición 1.2.5. *Dado un complejo simplicial K de dimensión d, definimos una $k-$**cadena** de simplejos como una unión no vacía de simplejos σ_j de dimensión $k \leq d$ en K. Esto se puede*

denotar como la suma formal de simplejos

$$c = \sum \alpha_j \sigma_j,$$

donde $\alpha_j \in \{0, 1\} = \mathbb{Z}_2$.

Es necesario recalcar que en general en la teoría de homología simplicial, se definen las cadenas a manera de que estén orientadas. Esto se refleja directamente en la elección de los coeficientes, los cuales usualmente son tomados en $\mathbb{Z}$. Referimos al lector a Munkres (1984) para más detalle en la definición de homología simplicial en estos términos. Para nuestro objetivo, es preferible considerar a $\mathbb{Z}_2$, debido a que esto nos permite concebir a las $k-$cadenas como conjuntos de simplejos.

Una $k-$cadena, en lenguaje de teoría de conjuntos, es la unión de todos o algunos de los simplejos de dimensión k que están contenidos en K. Al conjunto de estas cadenas le vamos a adjudicar una estructura de grupo. Dentro de esta estructura, vamos a deducir que es natural diferenciar ciertos elementos e inducir con estas diferencias clases de equivalencia las cuales son las que van a ser de nuestro interés.

Dotamos al conjunto de $k-$cadenas $C_k(K)$ con la operación suma componente a componente como si fueran polinomios, es decir, si $c = \sum \alpha_j \sigma_j$ y $c' = \sum \alpha'_j \sigma_j$ son $k-$cadenas entonces $c + c' = \sum (\alpha_j + \alpha'_j)\sigma_j$. Es fácil observar que con esta operación suma, $(C_k, +)$ es un grupo abeliano. Notemos también que con la manera en que hemos definido la suma $c + c'$, en términos de conjuntos representa la diferencia simétrica de los conjuntos c y c', debido a que estamos trabajando con coeficientes en $\mathbb{Z}_2$. En efecto, si un simplejo σ_j está en ambas cadenas, entonces $\alpha_j + \alpha'_j = 1 + 1 = 0 \pmod 2$. Para $k < 0$ y $k > d$, C_k es el grupo trivial. Esto es consistente con la definición, pues no hay cadenas con dimensión menor que 0 o mayor que d. Con esto tenemos una sucesión de grupos de cadenas $...C_{-1} = \{0\}, C_0, C_1, ..., C_d, C_{d+1} = \{0\}, ...$ Relacionada a esta sucesión de grupos induciremos una operación entre ellos.

Definición 1.2.6. *Definimos la operación frontera* ∂_k *para un simplejo* $\tau = \sigma(u_0, u_1, ..., u_k)$ *como*

$$\partial_k \tau = \sum_{j=0}^{k} \sigma(u_0, ..., \widehat{u_j}, ..., u_k),$$

en donde el gorro significa que u_j *es omitido, es decir,* $\sigma(u_0, ..., \widehat{u_j}, ..., u_k)$ *representa el simplejo generado por* $\{u_0, u_1, ..., u_k\} \backslash \{u_j\}$.

De esta definición tenemos que la frontera de un $k-$simplejo es una $(k-1)-$cadena, la cual es la suma formal de sus $(k-1)-$caras. Así definimos $\partial_k : C_k \to C_{k-1}$ como

$$\partial_k c = \partial_k \sum \alpha_j \sigma_j = \sum \alpha_j \partial_k \sigma_j.$$

Habiendo definido la operación frontera para cadenas de simplejos, con esta vamos a diferenciar dos tipos de cadena, las que tienen frontera igual a cero y las que no. Como mencionábamos, intuitivamente lo que se conocerá como un agujero de dimensión k es el espacio vacío que podemos envolver con S^k. Situados en los complejos simpliciales, los agujeros de dimensión k serán identificados por $k-$cadenas. Esta caracterización la formalizaremos a continuación.

Ciclos y fronteras

Antes de establecer la siguiente definición, vamos a proporcionar un poco de intuición con un ejemplo, del cual esperamos se capte naturalmente el cómo es que surge la definición de *ciclo*. Observemos el complejo simplicial de dimensión 2 que se muestra en la Figura 6.

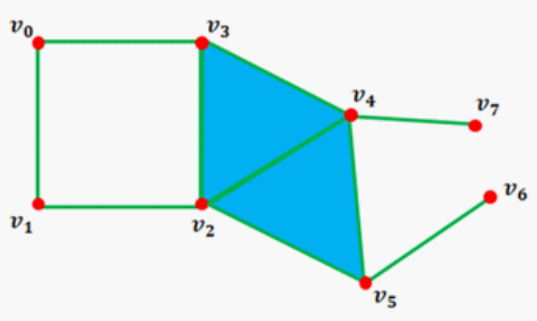

Figura 6. Ejemplo de complejo simplicial para ilustrar el concepto de ciclo.

Consideremos las cadenas $a = \sigma(v_0, v_1) + \sigma(v_1, v_2) + \sigma(v_2, v_3) + \sigma(v_3, v_0)$ y $b = \sigma(v_7, v_4) + \sigma(v_4, v_5) + \sigma(v_5, v_6)$. Por un lado, tenemos que a es un ciclo como intuitivamente lo conocemos,

mientras que b no lo es. Notemos que se cumple lo siguiente:

$$\begin{aligned}
\partial a &= (v_0 + v_1) + (v_1 + v_2) + (v_2 + v_3) + (v_3 + v_0) \\
&= (v_0 + v_0) + (v_1 + v_1) + (v_2 + v_2) + (v_3 + v_3) \\
&= 0, \text{ y} \\
\partial b &= (v_7 + v_4) + (v_4 + v_5) + (v_5 + v_6) \\
&= v_6 + v_7 \neq 0.
\end{aligned}$$

Intuitivamente, en dimensión 2 es fácil imaginar que para que una cadena sea un ciclo, en cierto sentido la suma de simplejos que representa debe iniciar y terminar en el mismo lugar, es decir, deben estar pegados en sus fronteras de manera contigua. En otras palabras, cada vértice (simplejo de dimensión 0) debe aparecer como frontera exactamente 2 veces. Dicho de esta manera podemos extender nuestra intuición a dimensiones más grandes, de lo cual surgen las siguientes definiciones.

Definición 1.2.7. *Definimos un $k-$**ciclo** c como una $k-$cadena con frontera nula, es decir $\partial c = 0$ y llamemos $Z_k = Z_k(K)$ al conjunto de $k-$ciclos.*

Definición 1.2.8. *Definimos una $k-frontera$ e como una $k-$cadena que es frontera de una $(k+1)-$cadena, $e = \partial d$ con $d \in C_{k+1}$, y llamemos $B_k = B_k(K)$ al conjunto de $k-$fronteras.*

Notemos además que $Z_k = \ker \partial_k$ y que $B_k = \operatorname{Im} \partial_{k+1}$. Con esto tenemos que los conjuntos de $k-$fronteras B_k y de $k-$ciclos Z_k son subgrupos de C_k. Estos grupos guardan entre sí una relación importante la cual enunciaremos a continuación con el *lema fundamental de homología.*

Nuestro objetivo como lo hemos venido exponiendo es el de estudiar los agujeros $k-$dimensionales de un complejo simplicial. Es natural pensar en que los ciclos que definimos pueden ser lo que formalmente sería un agujero. En cierto sentido es cierto, sólo nos falta definir cómo es que vamos a distinguir agujeros diferentes. Para ello vamos a establecer el siguiente resultado, con el cual obtenemos una relación primordial entre el grupo de $k-$ciclos Z_k y el grupo de $k-$fronteras B_k.

Lema 1.2.9. ***Lema fundamental de Homología.*** *Para todo entero k y toda $(k+1)-$cadena d se tiene que*

$$\partial_k \partial_{k+1} d = 0.$$

Este resultado nos asegura que toda $k-$frontera es también un $k-$ciclo. Así, B_k es un subgrupo de Z_k como se ilustra en la Figura 7. Con esto, sale a relucir la idea de poder diferenciar los ciclos mediante sus fronteras. A continuación se tratará esto con mayor formalidad.

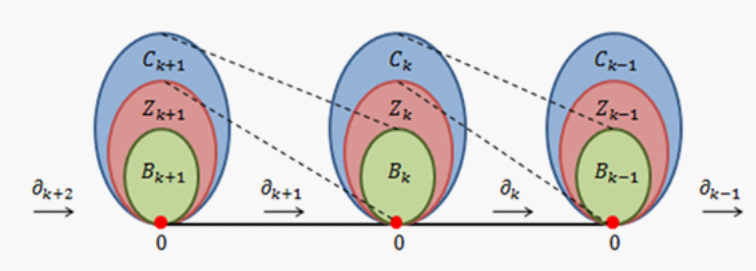

Figura 7. Relación entre los grupos de $k-$ciclos y $k-$fronteras.

Grupos de homología

Ya que los grupos de fronteras forman subgrupos de los ciclos, podemos tomar cocientes entre ellos. Así, es posible dividir el grupo de ciclos en clases de equivalencia que difieren entre sí por su frontera. Esto es lo que vamos a conocer como *grupos de homología*.

Definición 1.2.10. *Dado un complejo simplicial K definimos $H_k(K)$ el grupo de homología de dimensión k como $H_k(K) := Z_k/B_k$. Es decir, $H_k(K)$ describe los agujeros de dimensión k de K.*

A los elementos de $H_k(K)$ los identificaremos como *clases de homología*. Definimos los *números de Betti* $\beta_k(K)$ como el rango de $H_k(K)$, $\beta_k(K)$ =rank$(H_k(K))$. Es decir, $\beta_k(K)$ es el número de agujeros de dimensión k de K. Cuando es claro el contexto sólo escribiremos β_k. Estos números muchas veces tienen una interpretación evidente dentro del contexto del problema en el cual surgen los datos.

Para obtener un panorama más amplio de lo que es la homología simplicial recomendamos Espinoza (2015) la cual presenta una introducción con un enfoque directo a TDA. Por otro lado,

recomendamos Munkres (1984) para una exposición más general del tema. Habiendo establecido ya estas definiciones vamos ahora a describir la técnica principal del análisis topológico, la cual es la *homología persistente*. La idea esencial es, dado un complejo simplicial K vamos a tomar subconjuntos crecientes de K, los cuales serán subcomplejos, $K_1 \subset K_2 \subset ... \subset K$ y analizar la homología de cada K_j para estudiar la manera en que ésta va cambiando.

1.2.3. Homología persistente

Comenzaremos resumiendo algunas definiciones necesarias para el entendimiento de lo que es la homología persistente. Como hemos anticipado, lo que se busca hacer es estudiar la homología a través de una construcción del complejo. Más adelante detallaremos la relevancia del procedimiento.

Definición 1.2.11. *Sea K un complejo simplicial de dimensión d. Una filtración (finita) de K es una sucesión creciente de subcomplejos de K*

$$K_1 \subset K_2 \subset ... \subset K_p = K.$$

Es posible definir de una manera más general el concepto de filtración. Una filtración de K es una familia de subconjuntos de K, $\{K_t\}_{t\in I}$ donde I es un conjunto ordenado y cumple que

$$K_s \subset K_t \text{ si } s < t.$$

Existen muchas maneras de elegir el conjunto I para definir una filtración, pero resulta natural que I sea un intervalo. Sea $f : K \to \mathbb{R}$ una función continua, y consideremos la filtración $\{f^{-1}((-\infty, t])\}_{t\in I}$ donde $I = [a, b]$. En efecto, si $s < t$ entonces

$$f^{-1}((-\infty, s]) \subset f^{-1}((-\infty, t]).$$

Además, si tomamos una partición de $[a, b]$, $a = t_0 < t_1 < ... < t_p = b$ obtenemos una filtración finita de K

$$f^{-1}((-\infty, t_0]) \subset f^{-1}((-\infty, t_1]) \subset ... \subset f^{-1}((-\infty, t_p]).$$

En el sentido estricto, posiblemente $f^{-1}((-\infty, t_p])$ podría ser un subconjunto propio de K, pero debido a que K es compacto y f continua, se obtiene que f restringida a K es acotada. De aquí, el

hecho de poder elegir $[a, b]$ de manera conveniente de tal forma que $f^{-1}((-\infty, t_p]) = K$. A partir de esto, tenemos que podemos definir una filtración de un complejo K definiendo una función entre éste y los reales.

Para cada $0 \leq k \leq d$ con cada filtración de K, $K_1 \subset K_2 \subset ... \subset K_p = K$, más que estar interesados en los subcomplejos en sí, nos interesa la evolución topológica que se genera en ellos. Para cada $i \leq j$ tenemos una inclusión natural del espacio subyacente K_i al K_j, y con esto se induce un *homomorfismo*

$$f_k^{i,j} : H_k(K_i) \to H_k(K_j).$$

Esto da lugar a la correspondiente sucesión de grupos de homología conectados por dichos homomorfismos

$$H_k(K_1) \to H_k(K_2) \to \ldots \to H_k(K_p).$$

Esta sucesión nos permite definir el término *persistencia de una clase*. Si una clase (agujero) aparece por primera vez en $H_k(K_i)$ y deja de existir por primera vez en $H_k(K_j)$ decimos que permaneció durante $j - i$ pasos de la filtración, que *nació* en el tiempo i y *murió* en el tiempo j. Justamente, definimos la *persistencia* de la clase como $j - i$. Para entender mejor estas definiciones consideremos el ejemplo ilustrado en la Figura 8, en donde se muestra una filtración arbitraria de K.

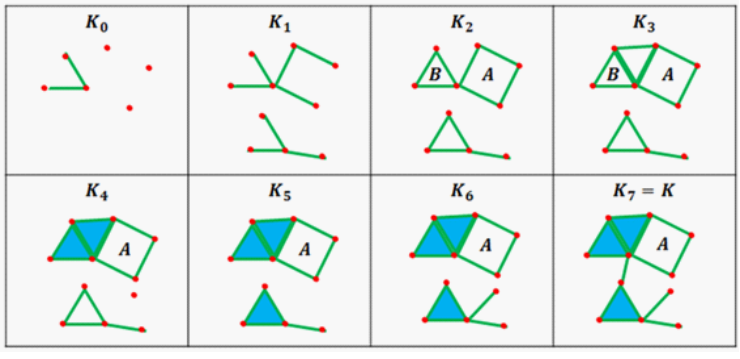

Figura 8. Filtración del complejo K.

Notemos que el ciclo A nació en el paso 2 y permaneció hasta el final de la filtración, es decir nunca murió. Entretanto, el ciclo B nació en el paso 2 y murió en el paso 4. Más adelante veremos que para filtraciones más específicas, la permanencia de una clase estará directamente relacionada

con la importancia de la misma en el subespacio K completo. Una clase que permanece por muchos pasos en la filtración será más importante que una que permanece pocos. Los términos "muchos" o "pocos" en un principio parecen muy subjetivos, pero es justamente la necesidad de objetividad la que nos lleva a definir un umbral para definir cuándo una clase es significativa y cuándo no.

Definición 1.2.12. *El $k-$ésimo grupo de homología persistente se define como la imagen de los homomorfismos inducidos por la correspondiente inclusión natural, $H_k^{i,j} = im f_k^{i,j}$ para $0 \leq i \leq j \leq p$. Los correspondientes $k-$ésimos números de Betti son los rangos de estos grupos, $\beta_k^{i,j} = rank(H_k^{i,j})$.*

Notemos que $H_k^{i,i} = H_k(K_i)$. Tenemos que los grupos de homología persistente $H_k^{i,j}$ corresponden a las clases que están presentes en $H_k(K_i)$ y que siguen vivas en $H_k(K_j)$, es decir, $H_k^{i,j} = Z_k(K_i)/(B_k(K_j) \cap Z_k(K_i))$. De aquí podemos definir más formalmente los conceptos de nacimiento y muerte de una clase. Una clase γ presente en $H_k(K_i)$ se dice que nació en K_i si $\gamma \notin H_k^{i-1,i}$. Por otra parte, si γ nació en K_i entonces esta muere en K_j si $f_k^{i,j-1}(\gamma) \notin H_k^{i-1,j-1}$ pero $f_k^{i,j}(\gamma) \in H_k^{i-1,j}$. Con esto hemos definido formalmente lo que es la homología persistente. Referimos al lector a Harer y Edelsbrunner (2010) para aumentar el panorama sobre este tema.

Hasta ahora, con lo que hemos descrito sobre homología y homología persistente no es claro aún cómo podemos aplicarlo a un conjunto de datos. Sea $S_n = \{X_1, X_2, ..., X_n\}$ un subconjunto de un espacio métrico completo (X, d). Vamos a considerar el caso en el que f es una función distancia. Recordemos que si A es compacto entonces la función $f : X \to \mathbb{R}$ definida como $f_A(x) = d(x, A) = \text{ínf}_{y \in A}\, d(x, y)$, es continua. Supongamos que S_n está muestreado con cierta distribución P cuyo soporte es una variedad compacta $d-$dimensional $\mathbb{M} \subset (X, d)$, con posible ruido en una vecindad de la misma. Nos interesa estudiar la homología de $\mathbb{M}$.

Definamos ahora para cada $\varepsilon > 0$ los conjuntos de nivel,

$$\begin{aligned}\widehat{L}_\varepsilon &= f_{S_n}^{-1}((-\infty,\varepsilon]) \\ &= \{x \in \mathbb{R}^D : d_{S_n}(x) \leq \varepsilon\} \\ &= \{x \in \mathbb{R}^D : \inf_{y \in S_n} d(x,y) \leq \varepsilon\} \\ &= \bigcup_{y \in S_n} B(y,\varepsilon),\end{aligned}$$

donde $B(y,\varepsilon)$ es la bola cerrada de radio ε y centrada en y. Si nuestra muestra S_n es lo suficientemente densa, tenemos que $H_k(\mathbb{M})$ es un subgrupo del $k-$ésimo grupo de homología de los conjuntos de nivel $\widehat{L}_\varepsilon$ para cierto intervalo de valores de ε. Es decir, podemos "hinchar" un poco nuestro espacio $\mathbb{M}$ sin perder sus propiedades en sentido de Homología, lo cual se formaliza con el *Teorema del nervio*, ver Harer y Edelsbrunner (2010). En la Figura 9, notemos que el objeto en C tiene la misma homología que la variedad $\mathbb{M}$ (en A), tiene una sola componente y dos agujeros de dimensión 1.

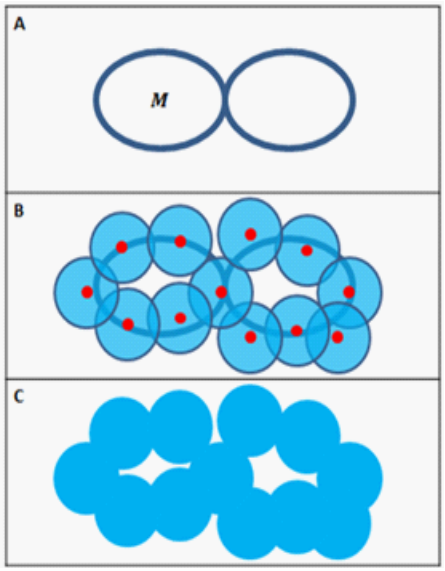

Figura 9. La variedad M en A tiene la misma homología que el conjunto de nivel $\widehat{L}_\varepsilon$ en C.

La elección del ε adecuado no es una tarea clara ni determinada; para valores muy pequeños lo que tenemos es la homología de un conjunto de n puntos, mientras que para valores muy grandes lo que obtenemos es la homología de un solo punto. Es aquí donde la importancia de la homología

persistente sale a relucir. Este procedimiento no sólo captura la información de los agujeros del complejo, sino que mediante la persistencia le da a cada clase cierta importancia. Con esto evitamos elegir un ε que no refleje de manera apropiada las características topológicas de la variedad $\mathbb{M}$.

Como hemos descrito hasta ahora, definimos los grupos de homología para complejos simpliciales. Definamos ahora los complejos de Čech y los complejos de Vietoris-Rips. Estos complejos relacionan directamente al conjunto de datos S_n con la variedad $\mathbb{M}$ de la que suponemos están muestreados.

Definición 1.2.13. ***Complejo de* Čech**. *Dado un conjunto de puntos* $S_n = \{X_1, X_2, ..., X_n\}$ *y un* $\varepsilon > 0$*, definimos el complejo* $\check{C}ech(S_n, \varepsilon)$ *como el conjunto de simplejos* $\tau = \sigma(v_1, ..., v_k)$ *tales que* $\{v_1, ..., v_k\} \subset S_n$ *y*

$$\bigcap_{j=1}^{k} B(v_j, \varepsilon) \neq \emptyset.$$

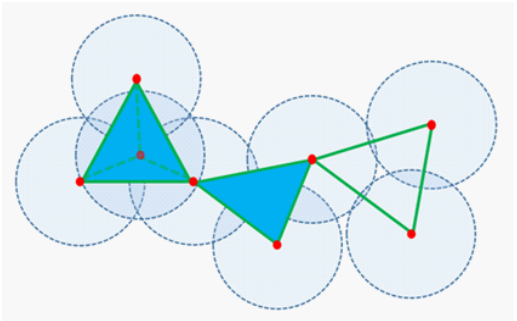

Figura 10. Complejo de Čech.

En la Figura 10 se muestra un ejemplo sencillo del complejo de Čech con base en 8 puntos. Notemos que la homología de Čech(S_n, ε) es la misma que la homología de $\widehat{L}_\varepsilon$. Así, tenemos que para cierto rango de valores de ε,

$$H_k(\widehat{L}_\varepsilon) = H_k(\check{\text{C}}\text{ech}(S_n, \varepsilon)).$$

De aquí que es factible utilizar el complejo de Čech(S_n, ε) para aproximar la homología de $\mathbb{M}$. Para una partición del intervalo $[0, \varepsilon]$, $0 < \varepsilon_1 < \varepsilon_2 < ... < \varepsilon_k = \varepsilon$ tenemos una sucesión de complejos de Čech que a su vez es una filtración para Čech(S_n, ε),

$$\check{\text{C}}\text{ech}(S_n, \varepsilon_1) \subset \check{\text{C}}\text{ech}(S_n, \varepsilon_2) \subset ... \subset \check{\text{C}}\text{ech}(S_n, \varepsilon_k).$$

Esta filtración se conoce como filtración de Čech. Con esta filtración obtenemos también sucesiones de grupos de homología

$$H_p(\check{\text{C}}\text{ech}(S_n, \varepsilon_1)) \to H_p(\check{\text{C}}\text{ech}(S_n, \varepsilon_2)) \to \ldots \to H_p(\check{\text{C}}\text{ech}(S_n, \varepsilon_k)).$$

Esta sucesión de grupos de homología es la base de la homología persistente de Čech(S_n, ε), la cual como ya mencionamos es una aproximación de la homología persistente de $\mathbb{M}$. Computacionalmente no es muy sencillo trabajar con este complejo, pero gracias al Lema de Vietoris-Rips, el cual enunciaremos a continuación, es posible aproximar al complejo de Čech con otro complejo, el cual formalizamos con la siguiente definición.

Definición 1.2.14. ***Complejos de Vietoris-Rips****. El complejo de Vietoris-Rips para un conjunto de puntos $S_n = \{X_1, X_2, ..., X_n\}$ y $\varepsilon > 0$ se define como el conjunto de simplejos $VR(S_n, \varepsilon)$ tal que,*

a) $\forall x \in S_n$, $x \in VR(S_n, \varepsilon)$

b) Para $v_1, v_2 \in S_n$, $\sigma(v_1, v_2) \in VR(S_n, \varepsilon)$ *si y sólo si* $d(v_1, v_2) \leq 2\varepsilon$.

c) Para $\{v_1, ..., v_k\} \subset S_n$, $\sigma(v_1, ..., v_k) \in VR(S_n, \varepsilon)$ *si y sólo si* $\sigma(v_i, v_j) \in VR(S_n, \varepsilon)$ $\forall v_i, v_j \in \{v_1, ..., v_k\}$.

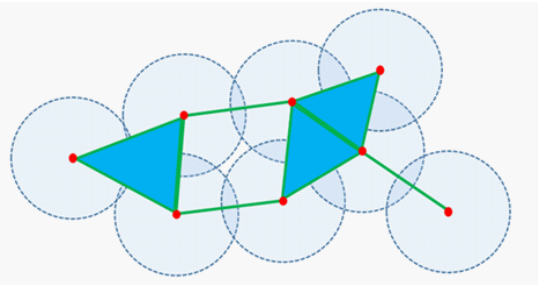

Figura 11. Complejo de Vietoris-Rips.

En la Figura 11 se muestra un ejemplo del complejo de Vietoris-Rips. Para una partición del intervalo $[0, \varepsilon]$, $0 < \varepsilon_1 < \varepsilon_2 < ... < \varepsilon_k = \varepsilon$ obtenemos la filtración de Vietoris-Rips,

$$\mathrm{VR}(S_n, \varepsilon_1) \subset \mathrm{VR}(S_n, \varepsilon_2) \subset ... \subset \mathrm{VR}(S_n, \varepsilon_k).$$

En la Figura 12, se ejemplifica esta filtración en el caso donde $\mathbb{M}$ es un anillo. Una de las principales diferencias entre el complejo de Čech y el de Vietoris-Rips es que en el primero es posible tener la frontera de un simplejo que delimite un agujero, caso contrario al complejo de Vietoris-Rips; es decir, en el complejo de Čech es posible tener triángulos (o tetraedros en dimensiones mayores) vacíos. Sin embargo, el siguiente resultado nos permite aproximar el complejo de Čech con el de Vietoris-Rips, el cual resulta muy sencillo de manejar computacionalmente.

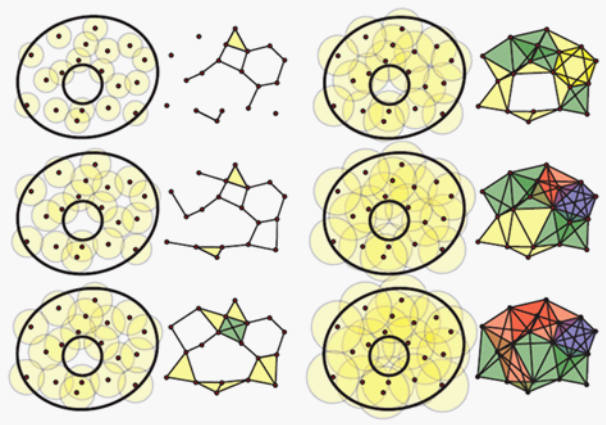

Figura 12. Tomada de Ghrist (2008). Filtración de Vietoris-Rips en un anillo (annulus).

Lema 1.2.15. ***Lema de Vietoris-Rips.*** *Para todo $\varepsilon > 0$ se tiene que*

$$\check{C}ech(S_n, \varepsilon) \subset VR(S_n, \varepsilon) \subset \check{C}ech(S_n, \sqrt{2}\varepsilon).$$

Referimos al lector a Harer y Edelsbrunner (2010) para la demostración de este lema. Es por este resultado y debido a que el complejo de Vietoris-Rips es muy sencillo de computar, que es común aproximar con éste el complejo de Čech. A partir de aquí tomaremos de manera estándar la filtración de Vietoris-Rips, aunque existen algunas otras filtraciones que por lo regular surgen directamente del contexto en que se lleva a cabo el análisis.

Código de barras y diagrama de persistencia

Habiendo establecido el procedimiento de homología persistente, ahora definiremos dos objetos que serán los que resuman los resultados de la misma. El lema fundamental de homología persistente nos permite asegurar que dichos objetos codifican correctamente toda la información de los grupos de homología persistente. A continuación vamos a definirlos y a enunciar dicho lema.

Dada una filtración del complejo K, $K_0 \subset K_1 \subset \ldots \subset K_p$, tenemos una sucesión de grupos de homología $\{H_p(K_j)\}_{j=1}^{k}$. Para $i < j$ y para toda k, definimos $\mu_k^{i,j}$ como el número de clases que nacen en $H_k(K_i)$ y murieron en $H_k(K_j)$, es decir

$$\mu_k^{i,j} = (\beta_k^{i,j-1} - \beta_k^{i,j}) - (\beta_k^{i-1,j-1} - \beta_k^{i-1,j}).$$

La primera diferencia representa las clases que nacieron en K_i o antes y que murieron en K_j, mientras que la segunda diferencia es el número de clases que nacieron en K_{i-1} o antes y que murieron en K_j. Definimos el conjunto de puntos

$$I_k = \{(i,j) : \mu_k^{i,j} > 0\},$$

es decir, I_k representa a los puntos nacimiento-muerte de todas las clases que existieron en la filtración. Cada punto (i,j) tiene multiplicidad $\mu_k^{i,j}$. Definimos el *diagrama de persistencia* de dimensión p de K, $P_k(K)$ o simplemente P_k, como el conjunto I_k con la recta identidad, es decir, $P_k(K) = I_k \cup \{(x,x) : x > 0\}$. Además, para incluir todo el resumen de la homología en un solo diagrama, con lo que vamos a trabajar es con el conjunto

$$P = P(K) = \cup_{k=0}^{D-1} P_k(K),$$

donde D es la dimensión de K. En la Figura 13 los puntos rojos representan las componentes conexas o agujeros de dimensión 0 y los triángulos verdes los agujeros de dimensión 1.

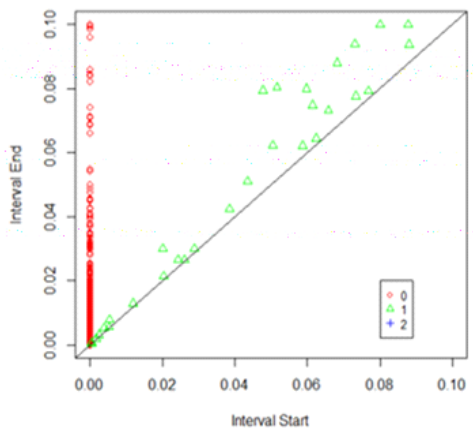

Figura 13. Diagrama de persistencia.

Lema 1.2.16. *Lema fundamental de homología persistente. Sea* $\emptyset = K_0 \subset K_1 \subset \ldots \subset K_p = K$ *una filtración de* K*. Para cada par de índices* $0 \leq t \leq s \leq p$ *y cada dimensión* k*, el* $k-$*ésimo número de Betti es*

$$\beta_k^{t,s} = \sum_{i \leq t} \sum_{j > s} \mu_k^{i,j}.$$

Como anticipábamos, este resultado nos permite asegurar que el diagrama de persistencia contiene toda la información de los grupos de homología persistente. Existen más representaciones de dicha información. Por un lado están los *panoramas de persistencia*, los cuales presentan la información como una función, lo cual permite abordarlos con teoría para funciones; referimos al lector a Bubenik (2015) para leer más sobre estos objetos, en donde se combinan herramientas de estadística y aprendizaje máquina para obtener varios resultados interesantes como una ley fuerte de grandes números y un teorema central del límite para estos objetos. Por otro lado tenemos los *códigos de barras* los cuales son una representación gráfica de los intervalos (i, j) y no vistos como puntos. Estos códigos los definiremos a continuación.

El código de barras es más bien un objeto visual más que un conjunto matemático formal. El objetivo es enlistar todas las clases que surgieron y representarlas con los intervalos (i, j), como se muestra en la Figura 14.

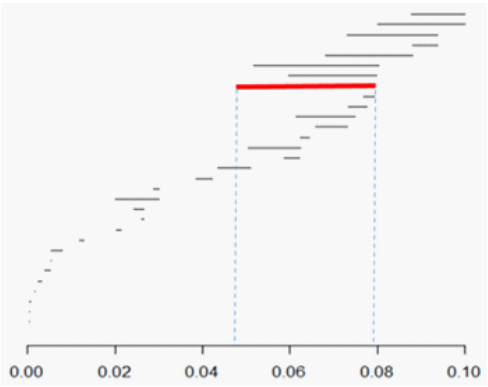

Figura 14. Código de barras.

El eje horizontal representa el momento en la filtración; así, si una barra abarca de 0.048 al 0.079 (como la que se muestra en la Figura 14) significa que existió una clase que nació en $H_k(K_{0.048})$ y persistió hasta $H_k(K_{0.079})$. Aquí podemos ver que entre más largo sea el intervalo (i, j) significa que tenemos una clase significativamente más grande que otra que tenga un intervalo de vida más corto.

A continuación describiremos el paquete pHom de R. Éste es una de las opciones disponibles que permiten obtener diagramas de persistencia y códigos de barras. Existen otras opciones como lo son el paquete TDA en R, la librería JavaPlex en Matlab y las librerías Perseus y Dionysus en C++. Estas últimas también fueron consultadas en el desarrollo de la tesis, pero no hablaremos de ellas debido a que en un principio no resultaron tener ventajas importantes respecto a los paquetes pHom y TDA, con relación a la obtención de diagramas de persistencia para los casos de estudio en ecología que trataremos en el Capítulo 3. El paquete de R TDA contiene la implementación de algunos de los métodos presentados en Fasy et al. (2014). Mi aportación a estas implementaciones es su extensión a distancias más generales y particularmente a la distancia de Mahalanobis, de lo cual hablaremos con mayor detalle en el capítulo siguiente.

1.2.4. Software (R)

Paquete *pHom*. Este fue el primer paquete que surgió para R que permitía aplicar homología persistente a un conjunto de datos. Cuenta con 3 funciones principales. La primera y más importante es "**phom**" que es la que nos permite calcular las clases de equivalencia y los intervalos de

persistencia (i, j). Se invoca mediante la línea de código

```
phom(X, dimension, max_filtration_value, mode="vr", metric="euclidian", p).
```

`X` puede ser dos cosas: por un lado, una matriz en donde cada renglón representa un dato, o bien una matriz de distancias. En el caso en el que X sea una matriz de datos, existen algunas opciones para la métrica con la cual dotamos nuestros datos como espacio métrico.

`dimension` se refiere a la dimensión máxima en la que queremos analizar la homología.

`max_filtration_value` se refiere al ε para la filtración en $(0, \varepsilon)$. En general no es claro cómo escoger este parámetro. Se puede caer en el error de escoger un ε muy pequeño. Una opción para esto es ver cuál es el diámetro del conjunto S_n, ε debe estar entre 0 y este valor.

`mode` se refiere al tipo de filtración. El paquete cuenta con 2 filtraciones: Vietoris-Rips y Lazy-Witness.

`metric` indica la métrica con la que queremos que sea tratado X. Las opciones son: "distance_matrix", "euclidean", "maximum", "manhattan", "canberra", "binary", "minkowski". distance_matrix indica que X es una matriz de distancias y no de datos, lo cual nos permite utilizar cualquier métrica que deseemos.

`p` es el parámetro para la métrica de Minkowski.

Lo que nos regresa esta función es una matriz con 3 columnas. La primera indica la dimensión de la clase, la segunda su nacimiento y la tercera su muerte. Cada renglón representa una clase que existió en algún momento de la filtración; ver Figura 15.

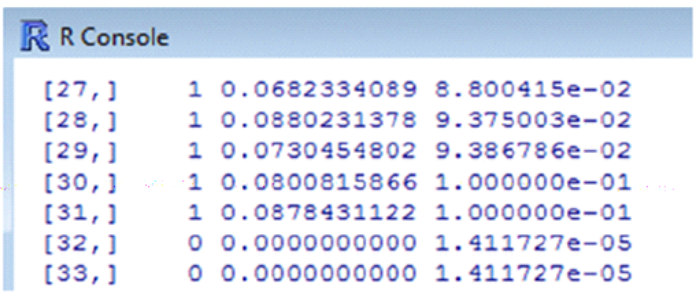

Figura 15. Resultado de la función phom en R.

"plotBarcodeDiagram" grafica el código de barras como se muestra en la Figura 14.

```
plotBarcodeDiagram(intervals, dimension, max_f, title="BarcodeDiagram")
```

`intervals` es la matriz que nos arroja la función phom.
`dimension` es la dimensión de la homología de la cual queremos graficar el código de barras.
`max_f` es el mismo valor (o uno menor) que max_filtration_value en la función "phom".
`title` es el título que llevará el gráfico.

"plotPersistenceDiagram" grafica el diagrama de persistencia como se muestra en la Figura 13. Cabe mencionar que nosotros definimos al diagrama de persistencia como un conjunto de puntos en el plano y lo definimos para cada dimensión k. Aquí se presentan las clases de dimensión menor o igual al parámetro que le indiquemos. Es decir, se presentan encimados todos los $P_k(S_n)$ para $k \leq$max_dim.

```
plotPersistenceDiagram(intervals, max_dim, max_f, title="PersistenceDiagram")
```

`intervals` es la matriz que nos arroja la función phom.
`max_dim` es la dimensión hasta la cual queremos visualizar en el diagrama de persistencia.
`max_f` es el mismo valor (o uno menor) que max_filtration_value en la función "phom".
`title` es el título que llevará el gráfico.

En la Figura 13 se muestra un ejemplo de diagrama de persistencia obtenido con la función plotPersistenceDiagram. Los puntos rojos representan las componentes, los triángulos verdes los agujeros de dimensión 1 y las cruces azules los agujeros de dimensión 2, que en este caso no hay.

Es importante mencionar que TDA es un paquete de R cuya primera versión fue publicada a mediados del 2015. Este paquete es por mucho más completo que pHom, ya que cuenta con una gama grande de funciones relacionadas al análisis topológico de datos. Dentro de las funciones que vale la pena resaltar es la de gridDiag la cual permite obtener el diagrama de persistencia con base en una filtración definida a partir de una función cualquiera f. Además como ya mencionamos, este paquete contiene implementaciones de algunos de los métodos descritos en Fasy et al. (2014) para generar conjuntos de confianza para diagramas de persistencia. Por otro lado, cuenta también con

funciones distancia entre diagramas de persistencia, herramientas necesarias para implementar la prueba descrita en Robinson y Turner (2013); de esta prueba hablaremos en el siguiente capítulo.

1.2.5. Problemas estadísticos

Como ya mencionábamos brevemente en la introducción, existen aspectos estadísticos a ser tomados en cuenta para la correcta interpretación de resultados. Al construir un diagrama de persistencia o un código de barras lo que observaremos son las clases o agujeros que contiene una estimación de la homología de $\mathbb{M}$. Sin embargo, no se trata de la homología verdadera precisamente: lo que tenemos es una estimación del diagrama de persistencia teórico. Esta estimación tendrá relacionada una noción que definiremos como ruido topológico. Debido a que en el diagrama de persistencia estamos observando la estimación con base en un conjunto finito de puntos muestreado sobre lo que supusimos un objeto continuo en $\mathbb{R}^D$, surgirán agujeros que realmente no existen en el objeto teórico $\mathbb{M}$. La variabilidad e incertidumbre de estos agujeros es lo que da pie al reconocimiento de ruido topológico, mismo que por la naturaleza discreta está relacionado con la nube de datos en sí. La presencia de ruido en la muestra es una fuente que se arrastra hasta el resultado final observado en el diagrama de persistencia. En el siguiente capítulo revisaremos el enfoque de Fasy et al. (2014) con el cual se hace frente a esta situación. La idea esencial será definir una distancia entre diagramas de persistencia y calcular un cuantil para la distancia entre el diagrama estimado y el diagrama teórico. Con base en esta distancia se construirá un método general para reconocer una propiedad verdadera oscurecida por ruido topológico.

Por otro lado, atenderemos un segundo aspecto que surge de manera menos evidente en aplicaciones de TDA cuyo estudio es interesante. Surge un problema formal de pruebas de hipótesis. En efecto, para dos siluetas o imágenes, podría ser de interés su comparación. Podría tratarse de dos muestras de un mismo objeto, sólo tal vez movido o deformado y lo que queremos investigar es si efectivamente se trata de uno solo. Para esto, la idea es estudiar la homología de ambas nubes de datos y con base en ello definir una manera de realizar una prueba de hipótesis formal. Para esto estudiaremos el trabajo realizado en Robinson y Turner (2013), el cual se basa en el uso de pruebas de permutación y la distancia entre diagramas de persistencia.

Capítulo 2

Aspectos estadísticos

2.1. Estimación del diagrama de persistencia

Una filtración de un espacio X es lo que define un diagrama de persistencia de manera primordial. Como vimos, hay muchas maneras de obtener la filtración y una manera fácil de parametrizarla es con el uso de una función real $f : X \to \mathbb{R}$. Para lo que sigue en cuanto a la cuantificación de incertidumbre, será importante poder determinar cómo es que cambia el diagrama de persistencia si cambiamos dicha función f. Si nos restringiéramos a cierta clase de funciones, sería deseable hacer uso de teoría de análisis funcional para caracterizar a través de estas funciones los diagramas de persistencia obtenidos. Sin embargo, en esta sección definiremos una distancia directamente en el espacio de diagramas de persistencia y enunciaremos un resultado—conocido como *estabilidad*—que establece que un pequeño cambio en la función que lo parametriza, produce un pequeño cambio en el diagrama.

La distancia a considerar entre diagramas de persistencia es la *distancia cuello de botella*. Para definirla y para enunciar el teorema de estabilidad es necesario hacer algunas definiciones técnicas precisas. Recordemos que la notación para un diagrama de persistencia de un subespacio X respecto a la filtración dada por la función f es

$$P(X, f).$$

Cuando es claro sobre cuál es el espacio X que estamos trabajando, sólo escribiremos $P(f)$.

Definición. Sean $f,\ g : X \to \mathbb{R}$. Para $a = (a_x,\ a_y)$, $b = (b_x,\ b_y) \in \mathbb{R}^2$, sea

$$d_\infty(a, b) = \text{máx}\{|a_x - b_x|,\ |a_y - b_y|\}.$$

Dada una biyección $\varphi : P(X, f) \to P(X, g)$, sea

$$d_\varphi(P(f), P(g)) = \max_{x \in P(f)} d_\infty(x, \varphi(x)).$$

Definimos la distancia *"cuello de botella"* entre $P(f)$ y $P(g)$ como

$$W_\infty(P(f), P(g)) = \inf_{\substack{\varphi: P(f) \to P(g) \\ \text{biyección}}} d_\varphi(P(f), P(g)).$$

Notemos que $P(f)$ y $P(g)$ tienen la misma cardinalidad pues ambos contienen a la recta identidad. En Turner et al. (2012) se detalla una prueba basada en el algoritmo húngaro (también conocido como algoritmo de asignación de Munkres) de que el ínfimo sobre las biyecciones siempre es alcanzable. Gracias a este resultado podemos asegurar que $W_\infty(P(f), P(g))$ está bien definida. Con esto, al conjunto de diagramas de persistencia de un subespacio X se le otorga una estructura de espacio métrico. Ahora enunciaremos un resultado que nos proporciona la estabilidad en dichos diagramas como la mencionábamos anteriormente.

Teorema 2.1.1. ***Teorema de estabilidad cuello de botella.*** *Sea X finitamente triangulable y sean $f, g : X \to \mathbb{R}$ continuas. Entonces, la distancia cuello de botella entre los correspondientes diagramas de persistencia está acotada por la distancia L_∞ entre las funciones f y g, es decir,*

$$W_\infty(P(f), P(g)) \leq ||f - g||_\infty.$$

Para la demostración de este teorema referimos al lector a Cohen-Steiner et al. (2007) y Chazal et al. (2016). Este resultado es esencial para el sustento de los métodos para generar los conjuntos de confianza. En este contexto, se busca obtener un cuantil para la distancia cuello de botella entre el diagrama de persistencia estimado y el teórico. Partiendo de la definición de la distancia cuello de botella definiremos el término ruido topológico, diferenciando con esto información verdadera en el diagrama de persistencia. Este resultado nos permite enfocarnos en $||f - g||_\infty$ en lugar de en $W_\infty(P(f), P(g))$, que para ciertas funciones f y g será más fácil de trabajar.

Se pueden considerar diferentes distancias para los diagramas de persistencia. La *distancia de Wasserstein* de grado p, $W_p(P(f), P(g))$, es un caso más general de la distancia cuello de botella, en donde la idea es minimizar la suma de distancias $\sum_{x\in P(f)} d_\infty(x, \varphi(x))^p$ en lugar de $\max_{x\in P(f)} d_\infty(x, \varphi(x))$, bajo todas las biyecciones. En la sección siguiente la definiremos formalmente. A pesar de que existe un resultado similar de estabilidad para esta distancia, es necesario asumir un número mayor de restricciones para el espacio X y para las funciones f, g. Es por esto que para nuestro objetivo utilizamos la distancia cuello de botella en lugar de la de Wasserstein, pero se recomienda al lector consultar Cohen-Steinner et al. (2010) para profundizar un poco más sobre ella: su definición formal y dos teoremas importantes de estabilidad en donde se establece una cota superior de $W_p(P(f), P(g))$ suponiendo f y g funciones Lipschitz. La importancia de la distancia de Wasserstein para diagramas de persistencia se ejemplifica de manera excelente en Mileyko et al. (2011), en donde muestran que el espacio de diagramas dotado con dicha distancia posee características que permiten la definición de medidas de probabilidad que soportan medias, varianzas, percentiles y probabilidad condicional.

Ahora bien, consideremos a f y g como funciones distancia a un conjunto compacto, generalizando las definiciones que introducimos en el capítulo anterior. La distancia $W_\infty(P(f), P(g))$ se encuentra acotada por cierta distancia entre dichos conjuntos compactos, la cual será más fácil de trabajar que $W_\infty(P(f), P(g))$. Sean A y B conjuntos compactos de $(\mathbb{R}^D, d)$. Definimos la *distancia de Hausdorff* entre A y B como

$$H(A, B) = \max\{\max_{x\in A}\min_{y\in B} d(x, y), \min_{x\in B}\max_{y\in A} d(x, y)\}.$$

Esta distancia ha sido utilizada en el ámbito estadístico en distintas aplicaciones. Cuevas y Fraiman (1998) define una distancia entre densidades basándose en la distancia de Hausdorff, definiendo la proximidad de funciones con la proximidad entre sus gráficas. Cuevas (2009) hace una exposición sobre el trabajo realizado hasta ese momento en el área de estimación de conjuntos, en donde la distancia de Hausdorff toma un rol importante para cuantificar la proximidad de la estimación con el conjunto teórico, esto para poder obtener resultados relacionados a su convergencia. En nuestro caso, la distancia de Hausdorff juega un papel importante pues resulta ser una cota para la distancia cuello de botella cuando se utiliza una función distancia para generar el diagrama de

persistencia.

Sea $\mathbb{M}$ una variedad $d-$dimensional encajada en un subespacio compacto $X \in \mathbb{R}^D$ y sea S cualquier subconjunto de $\mathbb{M}$. Sea P el diagrama obtenido con base en la función distancia $d_{\mathbb{M}}$ y sea P_S el diagrama obtenido con base en la función d_S. Notemos que P es justamente el diagrama que muestra la homología verdadera de $\mathbb{M}$, mientras que P_S es una estimación con base en un subconjunto. Por el teorema de estabilidad tenemos que

$$W_\infty(P, P_S) \leq ||d_{\mathbb{M}} - d_S||_\infty.$$

Es fácil ver que $||d_{\mathbb{M}} - d_S||_\infty = H(\mathbb{M}, S)$. De aquí obtenemos que

$$W_\infty(P, P_S) \leq H(\mathbb{M}, S).$$

La relevancia de este resultado es que si queremos acotar $W_\infty(P, P_S)$ entonces acotamos $H(\mathbb{M}, S)$, en particular para obtener un cuantil para $W_\infty(P, P_S)$ obtenemos un cuantil para $H(\mathbb{M}, S)$.

Como introducíamos en el capítulo anterior, queremos inferir la homología de $\mathbb{M}$. No observamos directamente a $\mathbb{M}$ sino una muestra $S_n = \{X_1, ..., X_n\}$ con distribución $\mathbb{P}$ que tiene concentración sobre $\mathbb{M} \subset \mathbb{R}^D$ o en una vecindad de ella. La homología del conjunto S_n en general no es la misma que la de $\mathbb{M}$. Si nuestra muestra es representativa, el conjunto $\widehat{L}_\varepsilon = \{x \in \mathbb{R}^D : d_{S_n}(x) \leq \varepsilon\}$ captura la homología de $\mathbb{M}$ para cierto intervalo de valores de ε. Sea P el diagrama de persistencia generado con base en la función distancia $d_{\mathbb{M}}$ y sea $\widehat{P}$ el generado con base en d_{S_n}, teniendo así que $\widehat{P}$ es una estimación de P respecto a la muestra S_n. Nuestro objetivo ahora es obtener un cuantil para la distancia $W_\infty(P, \widehat{P})$ y con este valor definir un umbral para diferenciar en $\widehat{P}$ una característica real (agujero) de una característica artificial, lo cual es lo que denominaremos ruido topológico.

Sea $\alpha \in (0, 1)$. Queremos encontrar $c_n = c_n(X_1, ..., X_n)$ tal que

$$\lim_{n\to\infty} \sup \mathbb{P}(W_\infty(P,\ \widehat{P}) > c_n) \leq \alpha.$$

Tenemos que dicho valor c_n es tal que $C_n = [0, c_n]$ es un conjunto de confianza asintótico para $W_\infty(P,\ \widehat{P})$ con probabilidad de cobertura $1 - \alpha$. Esto es,

$$\lim_{n\to\infty} \inf \mathbb{P}(W_\infty(P,\ \widehat{P}) \in C_n) \geq 1 - \alpha.$$

De aquí se obtiene que el conjunto de confianza C_n para $\widehat{P}$ es el conjunto de diagramas de persistencia $\widetilde{P}$ cuya distancia a $\widehat{P}$ es a lo más c_n:

$$C_n = \{\widetilde{P} : W_\infty(\widetilde{P},\ \widehat{P}) \leq c_n\}.$$

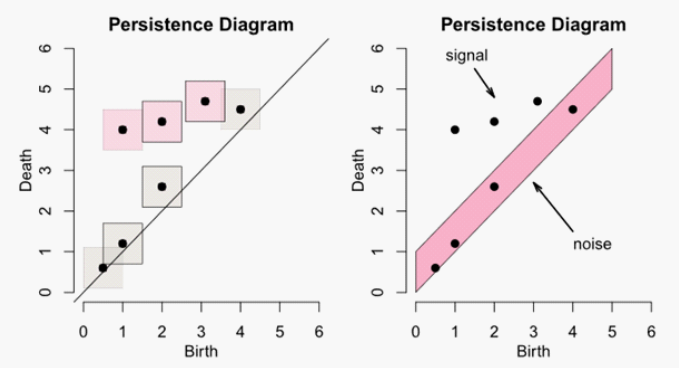

Figura 16. Tomada de Fasy et al. (2014). Detección de ruido y características verdaderas.

A partir de este conjunto de confianza podemos definir formalmente lo que es ruido topológico. Tenemos que la vecindad de radio c_n centrada en un punto $x \in \mathbb{R}^2$ respecto a la métrica d_∞, resulta ser un cuadrado de lado $2c_n$ centrado en x; ver Figura 16, tomada de Fasy et al. (2014). Si un diagrama de persistencia P contiene un punto con una vecindad de radio c_n que toca la diagonal, entonces hay un diagrama de persistencia $\widetilde{P}$ que no contiene dicho punto, y está a distancia de P menor o igual a c_n. Este hecho se debe a que podemos hacer una biyección entre P y $\widetilde{P}$ relacionando el punto en cuestión con un punto en la diagonal de $\widetilde{P}$. De aquí que si c_n es el cuantil para $W_\infty(P,\ \widehat{P})$ entonces a todo punto en $\widehat{P}$ que tenga una vecindad de radio c_n que no toque la diagonal será considerado como una señal verdadera. En efecto, dichos puntos pertenecen a todo diagrama contenido en C_n, en particular al diagrama de persistencia teórico P con alta probabilidad. Siguiendo este razonamiento, definiremos como ruido topológico a los puntos restantes (los que tienen una vecindad de radio c_n que sí toca la diagonal), pues es probable que éstos no figuren en el diagrama teórico. Vamos a señalar en el diagrama de persistencia estimado $\widehat{P}$ una banda que nos permita distinguir señales verdaderas directamente. Si trazamos una banda de ancho $\sqrt{2}c_n$ arriba de la diagonal (medida perpendicularmente), como se muestra en la Figura 20, tenemos que los puntos que están fuera de la banda representan agujeros reales, mientras que los que se encuentran dentro conforman el ruido topológico.

La cuestión ahora es cómo obtenemos el cuantil c_n. El teorema de estabilidad nos permite encontrar un cuantil para $||d_{\mathbb{M}} - d_{S_n}||_\infty$ y este mismo será un cuantil para $W_\infty(P, \widehat{P})$. Como mencionábamos, $||d_{\mathbb{M}} - d_{S_n}||_\infty$ es la distancia de Hausdorff entre $\mathbb{M}$ y S_n. Los primeros tres métodos que describiremos versan en la construcción de un cuantil para $H(\mathbb{M}, S_n)$. En efecto, del teorema de estabilidad y del hecho de que $||d_{\mathbb{M}} - d_{S_n}||_\infty = H(\mathbb{M}, S_n)$ se tiene que

$$W_\infty(P, \widehat{P}) \leq H(\mathbb{M}, S_n).$$

Por ello es suficiente encontrar c_n tal que

$$\lim_{n\to\infty} \sup \mathbb{P}(H(\mathbb{M}, S_n) > c_n) \leq \alpha.$$

Dentro de las ventajas de enfocarnos en la distancia de Hausdorff se encuentran que computacionalmente es muy sencilla de manejar y geométricamente su definición es muy intuitiva en el sentido de medir la cercanía de dos conjuntos. Además, como se puede analizar, la distancia de Hausdorff se enfoca en la geometría completa del objeto, permitiendo que el cuantil que obtengamos para dicha distancia sea utilizable para el diagrama de persistencia completo y no sólo de alguna dimensión de la homología.

A continuación describiremos los métodos descritos por Fasy et al. (2014) para generar el valor c_n, pero antes será necesario establecer algunas restricciones. Estas se extienden sobre el espacio $\mathbb{M}$ y sobre la medida de probabilidad con la que suponemos están muestreados los datos. De manera intuitiva, estas restricciones están situadas en la suavidad de la medida de probabilidad y en una propiedad del espacio $\mathbb{M}$ que le permita "hincharse un poco" sin perder sus características en sentido de homología. Estos supuestos son necesarios para que los métodos expuestos en Fasy et al. (2014) tengan validez, ya que los requieren directamente los teoremas que los sustentan.

Sean $X_1, X_2, ..., X_n \overset{\text{i.i.d}}{\sim} \mathbb{P}$ donde $X_j \in \mathbb{R}^D$. Sea $\mathbb{M}$ el soporte $d-$dimensional de $\mathbb{P}$ y definamos las cantidades

$$\rho(x, t) = \frac{\mathbb{P}(B(x, t/2))}{t^d} \text{ y}$$

$$\rho(t) = \inf_{x \in \mathbb{M}} \rho(x, t).$$

Supongamos que $\rho(x, t)$ es una función continua de t, definamos

$$\rho(x, \downarrow 0) = \lim_{t \to 0} \rho(x, t) \text{ y}$$

$$\rho = \lim_{t \to 0} \rho(t).$$

Supongamos lo siguiente:

1. $\mathbb{M}$ es una variedad compacta $d-$dimensional encajada en $\mathbb{R}^D$ y tal que reach$(\mathbb{M}) > 0$.

2. Para cada $x \in \mathbb{M}$, $\rho(x, t)$ es una función continua y acotada de t, diferenciable para $t \in (0, t_0)$ y diferenciable por la derecha en 0. Más aún, $\partial\rho(x, t)/\partial t$ existe y está acotada en una vecindad del cero. También, para algún $t_0 > 0$ y para algunas constantes C_1 y C_2 tenemos que

$$\sup_x \left[\sup_{0 \leq t \leq t_0} \left| \frac{\partial \rho(x, t)}{\partial t} \right| \right] \leq C_1 < \infty, \text{ y}$$

$$\sup_{0 \leq t \leq t_0} |\rho'(t)| \leq C_2 < \infty.$$

Los métodos a considerar son el método por submuestreo, el método por concentración de medida, el método por capas y el método por estimación de densidad. Para los primeros tres métodos vamos a suponer que la dimensión de $\mathbb{M}$ es conocida y que el soporte de la distribución es precisamente $\mathbb{M}$, es decir, vamos a suponer que no hay ruido en la muestra. Estas suposiciones serán innecesarias para el cuarto método. Específicamente se supone la existencia de datos atípicos fuera de $\mathbb{M}$ y para calcular la homología se hace uso de conjuntos de nivel, pero de la función densidad. Esto hará que dicho método sea más robusto que los precedentes, pero como expondremos, el método por estimación de densidad tiene un enfoque distinto al que se utiliza en los primeros tres.

2.1.1. Método por submuestreo

El primer método se basa en el uso de submuestreo. Politis et al. (1999) presenta el método de submuestreo tradicional, a donde referimos al lector para estudiar a más detalle la técnica, sus aplicaciones y la discusión que hacen los autores sobre su comparación con el método de Bootstrap. Básicamente la diferencia fundamental consiste en que el submuestreo considera muestreo sin remplazo mientras que el bootstrap lo hace con reemplazo. Originalmente para el uso de este método se adopta la suposición de que se tiene un estimador $\widehat{\theta}$ de un parámetro θ tal que $n^{\xi}(\widehat{\theta}-\theta)$ converge en distribución a una distribución fija F para algún $\xi > 0$. En nuestro caso no podemos suponer esto, pues necesitamos demostrar la convergencia de $n^{\xi}(\widehat{P}-P)$ para algún $\xi > 0$, pero hasta el momento no hemos definido medidas de probabilidad en el espacio de diagramas de persistencia. Sin meternos en más detalles al respecto, es posible utilizar el submuestreo sabiendo que lo que obtendremos es un intervalo de confianza conservativo para $W_{\infty}(P,\, \widehat{P})$.

Sea $b = b(n)$ tal que $b(n) = o\left[n/\log(n)\right]$ y $b(n) \underset{n}{\to} \infty$. Obtenemos las N submuestras $S^1_{b(n)}, S^2_{b(n)}, ..., S^N_{b(n)}$ de tamaño b de la muestra original S_n, donde $N = \binom{n}{b}$. Sea

$$T_j = H(S^j_{b(n)}, S_n) = \text{máx}\{ \underset{x\in S^j_{b(n)}}{\text{máx}} \underset{y\in S_n}{\text{mín}}\, d(x,y), \underset{x\in S_n}{\text{mín}} \underset{y\in S^j_{b(n)}}{\text{máx}}\, d(x,y)\},$$

para $j = 1, 2, ..., n$ y definamos

$$L_b(t) = \frac{1}{N}\sum_{j=1}^{N} 1_{\{T_j > t\}}.$$

Sea $c_b = 2L_b^{-1}(\alpha)$. El siguiente teorema nos proporciona el sustento teórico para utilizar c_b como una aproximación al cuantil que queremos.

Teorema 2.1.2. *Supongamos que $\rho > 0$. Entonces para n grande tenemos que*

$$\mathbb{P}(W_{\infty}(P,\ \widehat{P}) > c_b) \leq \mathbb{P}(H(\mathbb{M}, S_n) > c_b) \leq \alpha + O\left(\frac{b}{n}\right)^{1/4}.$$

Para la demostración del teorema referimos al lector a Fasy et al. (2014), la cual no es complicada en cuestión de teoría pero bastante detallada. Hace uso de resultados obtenidos en Romano y Shaikh (2012), en donde se establecen condiciones bajo las cuales es posible utilizar submuestreo

y/o bootstrap para la construcción de cuantiles. Cabe mencionar que en la práctica, incluso en el uso tradicional de submuestreo, no se obtienen las N submuestras sino un número suficientemente grande de submuestras de tamaño b. Esto, debido a que podría resultar computacionalmente muy costoso obtener las N submuestras.

El paquete TDA de R cuenta con la función *hausdInterval* que regresa justamente el valor de c_b para el caso en donde los datos están en el espacio euclidiano. Dicho paquete apareció por primera vez a mediados del 2015. Sin embargo, de cualquier forma fue necesaria la implementación de este método para datos con distancia diferente a la euclidiana. La distancia de Mahalanobis es una métrica sumamente utilizada en el ámbito estadístico. En el siguiente capítulo ahondaremos en la importancia de esta distancia y en la necesidad de utilizarla, en particular para los datos de la aplicación en ecología. Se implementó en R el método de submuestreo para esta distancia y se utilizó para generar una banda en los diagramas de persistencia obtenidos en el capítulo siguiente.

2.1.2. Método por concentración de medida

El enfoque de este método surge de resultados similares al del siguiente lema. En Devroye y Wise (1980) se tratan dos problemas que involucran estimación no paramétrica del soporte de un vector aleatorio, y se demuestra un resultado parecido para obtener la convergencia de una sucesión de medidas a cero. Por otro lado, en Smale and Weinberger (2008) tratan el problema de estudiar la homología de una subvariedad con un nivel alto de confianza, en donde se dan cotas para la complejidad de aprendizaje en el sentido computacional. El siguiente lema nos da una cota superior para $\mathbb{P}\left[W_\infty(\widehat{P}, P) > t\right]$, que en conjunto con los dos resultados que le siguen nos permitirá obtener el cuantil c_n resolviendo una ecuación.

Lema 2.1.3. *Para toda* $t > 0$

$$\mathbb{P}\left[W_\infty(\widehat{P}, P) > t\right] \leq \mathbb{P}\left[H(S_n, \mathbb{M}) > t\right] \leq \frac{2^d}{\rho(t/2)t^d} \exp\left[-n\rho(t)t^d\right].$$

Si además, $t < \min\{\rho/(2C_2), t_0\}$*, entonces*

$$\mathbb{P}\left[H(S_n, \mathbb{M}) > t\right] \leq \frac{2^{d+1}}{\rho t^d} \exp\left[-n\frac{\rho t^d}{2}\right].$$

Así, si $t_n(\alpha) < \min\{\rho/(2C_2), t_0\}$ *es la solución a la ecuación*

$$\frac{2^{d+1}}{\rho t_n^d} \exp\left[-n\frac{\rho t_n^d}{2}\right] = \alpha,$$

entonces

$$\mathbb{P}\left[W_\infty(\widehat{P}, P) > t_n(\alpha)\right] \leq \mathbb{P}\left[H(S_n, \mathbb{M}) > t_n(\alpha)\right] \leq \alpha.$$

Para usar este lema es necesario estimar ρ pues nos interesa encontrar la solución t_n de la ecuación $2^{d+1}/\rho t_n^d \exp\left[-n\rho t_n^d/2\right] = \alpha$. Para ello, sea P_n la medida empírica inducida por la muestra S_n, es decir, para todo conjunto boreliano $A \subset \mathbb{R}^D$,

$$P_n(A) = \frac{1}{n}\sum_{j=1}^{n} I_A(X_j),$$

donde I_A es la función indicadora en el conjunto A. Sea $r_n > 0$ y consideremos el estimador

$$\widehat{\rho}_n = \min_j \frac{P_n\left[B(X_j, r_n/2)\right]}{r_n^d}.$$

El siguiente resultado nos asegura que con una tasa adecuada de convergencia de $r_n \to 0$, el estimador $\widehat{\rho}_n$ es consistente para ρ.

Teorema 2.1.4. *Sea* $r_n \asymp [\log(n)/n]^{1/(d+2)}$, *entonces*

$$\widehat{\rho}_n - \rho = O_P(r_n),$$

en donde $a_n \asymp b_n$ *significa que existen* c_1 *y* c_2 *tal que* $a_n \leq c_1 b_n$ *y* $b_n \leq c_2 a_n$ *para* n *suficientemente grande.*

Habiendo estimado ρ, necesitamos $\widehat{\rho}_n$ para estimar $t_n(\alpha)$ como sigue. Supongamos que n es par en lo siguiente,

1. Dividamos la muestra en dos mitades, $S_n = S_{1,n} \sqcup S_{2,n}$ (unión disjunta).

2. Sea $\widehat{\rho}_{1,n}$ el estimador de ρ basado en $S_{1,n}$, y sea $\widehat{t}_{1,n}$ la solución a la ecuación

$$\frac{2^{d+1}}{t_n^d \widehat{\rho}_{1,n}} \exp\left(-n\frac{\widehat{\rho}_{1,n} t_n^d}{2}\right) = \alpha.$$

El siguiente teorema nos asegura que $\widehat{t}_{1,n}$ es un estimador para c_n con una tasa de convergencia $O\left[(\log n/n)^{1/(2+d)}\right]$.

Teorema 2.1.5. *Sea $\widehat{P}_2$ el diagrama de persistencia respecto a la función distancia a $S_{2,n}$. Entonces*

$$\begin{aligned}\mathbb{P}\left[W_\infty(\widehat{P}_2, P) > \widehat{t}_{1,n}\right] &\leq \mathbb{P}\left[H(S_{2,n}, \mathbb{M}) > \widehat{t}_{1,n}\right] \\ &\leq \alpha + O\left[\left(\frac{\log n}{n}\right)^{1/(2+d)}\right],\end{aligned}$$

donde $\mathbb{P}$ es la probabilidad conjunta respecto a la muestra completa y la aleatoriedad inducida por la división de la muestra.

En el artículo Fasy et al. (2014) demuestran el teorema anterior; se menciona además que en la práctica se encontró, pero sin demostrarse formalmente, que resolver la ecuación sin necesidad de dividir la muestra en realidad también funciona bien. Dentro del trabajo de esta tesis se implementó este método y se utilizaron las bandas para los diagramas de persistencia que se presentan en el capítulo siguiente. Respecto al método anterior, computacionalmente hablando, el método por concentración de medida es más rápido ya que como es sabido en el enfoque por submuestreo resulta frecuentemente muy complejo la obtención del cuantil. Al finalizar de exponer los cuatro métodos que se muestran en Fasy et al. (2014), veremos algunos ejemplos didácticos con estas implementaciones.

Por otro lado, se puede definir $\widehat{t}_n$ de la siguiente manera. Aunque es una forma muy sencilla, obtendremos un intervalo más conservador:

$$\widehat{t}_n = \left[\frac{2}{n\widehat{\rho}_n}\log\left(\frac{n}{\alpha}\right)\right]^{1/d}.$$

Así se tiene que $\widehat{t}_n = u_n\left[1 + O(\widehat{\rho}_n - \rho)\right]$, donde $u_n = \left[\rho\log(n/\alpha)/(n\widehat{\rho}_n)\right]^{1/d}$. Con esto,

$$\begin{aligned}\mathbb{P}\left[H(S_n, \mathbb{M}) > \widehat{t}_n\right] &= \mathbb{P}\left[H(S_n, \mathbb{M}) > u_n\right] + O\left[\left(\frac{\log n}{n}\right)^{1/(2+d)}\right] \\ &\leq \alpha + O\left[\left(\frac{\log n}{n}\right)^{1/(2+d)}\right].\end{aligned}$$

Hemos descrito el segundo método para calcular el cuantil c_n. La dependencia que tiene este del parámetro ρ lo convierte en un método muy frágil. Si la densidad tiene un valor cercano a cero,

incluso en alguna región pequeña, nuestra estimación no será buena. El siguiente método trabaja con esta problemática y obtiene una cota para $\mathbb{P}\left[H(S_n, \mathbb{M}) > t\right]$ más fina que la que se obtiene en este método.

2.1.3. Método por capas

A continuación vamos a exponer el tercer método desarrollado en Fasy et al. (2014). La idea principal es similar al del método anterior: se da una cota para $\mathbb{P}\left[W_\infty(\widehat{P}, P) > t\right]$ que depende de t (y otros parámetros) para así resolver una ecuación para t y encontrar el cuantil deseado. Recordemos la definición de

$$\rho(x, \downarrow 0) = \lim_{t \to 0} \rho(x, t) = \lim_{t \to 0} \frac{\mathbb{P}(B(x, t/2))}{t^d}.$$

Sea $G(y) = P\left[\rho(x, \downarrow 0) \leq y\right]$ la función de distribución de $\rho(x, \downarrow 0)$ y sea $g(y) = G'(y)$ su densidad.

Teorema 2.1.6. *Supongamos que g es acotada y tiene derivada uniformemente acotada y continua. Entonces para todo $t \leq \rho/(2C_1)$,*

$$\begin{aligned} \mathbb{P}\left[W_\infty(\widehat{P}, P) > t\right] &\leq \mathbb{P}\left[H(S_n, \mathbb{M}) > t\right] \\ &\leq \frac{2^{d+1}}{t^d} \int_\rho^\infty \frac{g(y)}{y} \exp\left(\frac{-nyt^d}{2}\right) dy. \end{aligned}$$

Para hacer uso de la expresión en la cota superior de este teorema, es necesario estimar g y ρ y así poder resolver una ecuación de la forma

$$\frac{2^{d+1}}{t^d} \int_\rho^\infty \frac{g(y)}{y} \exp\left(\frac{-nyt^d}{2}\right) dy = \alpha.$$

Sea K un kernel suave y simétrico que cumple las condiciones mencionadas en Giné y Guillou (2002), las cuales no mencionaremos por su naturaleza muy técnica pero que las cumplen los kernels comúnmente utilizados en la práctica. Sea

$$\widehat{g}(y) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h} K\left(\frac{y - Y_j}{h}\right),$$

donde $h > 0$, $Y_j = \widehat{\rho}(X_j, \downarrow r_n)$ y

$$\widehat{\rho}(x, \downarrow r_n) = \frac{P_n\left[B(x, r_n/2)\right]}{r_n^d}.$$

Teorema 2.1.7. *Sea* $r_n \asymp [\log(n)/n]^{1/(d+2)}$.

(1) Tenemos que

$$\sup_y |\widehat{g}(y) - g(y)| = O_P\left(h^2 + \sqrt{\frac{\log n}{nh}} + \frac{r_n}{h^2}\right).$$

Así, si tomamos $h \equiv h_n \asymp r_n^{1/4}$, *entonces*

$$\sup_y |\widehat{g}(y) - g(y)| = O_P\left[\left(\frac{\log n}{n}\right)^{1/(2(d+2))}\right].$$

(2) Supongamos que n *es par y que* $\rho > 0$. *Supongamos además que* g *es estrictamente positiva en su soporte* $[\rho, B]$. *Aleatoriamente separamos los datos en dos mitades* $S_n = S_{1,n} \sqcup S_{2,n}$. *Sean* $\widehat{g}_{1,n}$ *y* $\widehat{\rho}_{1,n}$ *estimadores de* g *y* ρ *respecto a* $S_{1,n}$ *y definamos* $\widehat{t}_{1,n}$ *la solución a la ecuación*

$$\frac{2^{d+1}}{\widehat{t}_{1,n}^d} \int_{\widehat{\rho}_{1,n}}^{\infty} \frac{\widehat{g}_{1,n}(y)}{y} \exp\left(\frac{-ny\widehat{t}_{1,n}^d}{2}\right) dy = \alpha.$$

Entonces

$$\mathbb{P}\left[W_\infty(\widehat{P}_2, P) > \widehat{t}_{1,n}\right] \leq \mathbb{P}\left[H(S_{2,n}, \mathbb{M}) > \widehat{t}_{1,n}\right] \leq \alpha + O(r_n).$$

En esta tesis no se implementó este método, debido a su complejidad computacional. En el sentido de convergencia, este método es igual que el anterior pues converge con una tasa de $O([\log(n)/n]^{1/(d+2)})$. Para generar los cuantiles para los diagramas obtenidos en la aplicación en nichos ecológicos y ambientes disponibles se utilizó principalmente el método por submuestreo y en algunos casos el de concentración de medida. Si suponemos que la densidad de la medida con la que se muestrean los datos no es cercana a cero, y es plausible si lo que tenemos es una sola componente conexa, entonces los primeros dos métodos son adecuados para generar el cuantil deseado.

2.1.4. Método por estimación de densidad

En este método vamos a tomar un enfoque totalmente diferente. Se construye un estimador suave para la densidad con la que suponemos están muestreados los datos y con base en esto generamos el diagrama de persistencia definido por los conjuntos de nivel de dicho estimador (Ver Figura 17). Recordemos que el término suavizar en estadística es utilizado para referirnos a la metodología de crear una función de proximidad sobre los datos con el objetivo de capturar

características importantes en ellos, restando importancia a ruido y estructuras a escala muy pequeña. En Bendich et al. (2011) se estudia un enfoque de suavizado basado en distancias de difusión. La idea principal en su trabajo es que, mediante el uso de caminatas aleatorias en la nube de datos, se define una métrica que mejora las propiedades topológicas e incrementa la capacidad de la homología persistente de capturarlas. Como veremos a continuación, con este enfoque lograremos capturar la homología de nuestra nube de datos de una manera más robusta que con los métodos anteriores.

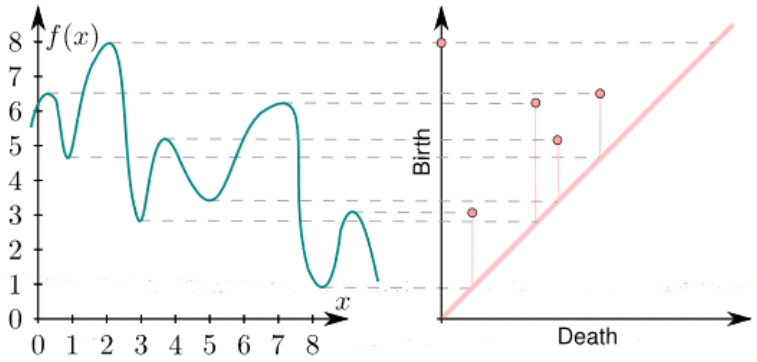

Figura 17. Tomada de Fasy et al. (2014). Diagrama de persistencia basado en los conjuntos de nivel de la densidad f.

Recordando, tenemos $X_1, X_2, ..., X_n$, una muestra con distribución P. Se define

$$p_h(x) = \int_{\mathbb{M}} \frac{1}{h^D} K\left(\frac{||x-u||_2}{h}\right) dP(u).$$

Notemos que p_h es la densidad de la medida de probabilidad $P_h = P * \mathbb{K}_h$, donde $\mathbb{K}_h(A) = h^{-D}\mathbb{K}(h^{-1}A)$ y $\mathbb{K}(A) = \int_A K(t)dt$. Esto es, p_h es una versión suavizada de P. El objetivo ahora es obtener el diagrama de persistencia $\mathcal{P}_h$ obtenido con base en los conjuntos de nivel de p_h. Bajo condiciones apropiadas, los conjuntos de nivel de p_h son capaces de capturar información topológica importante a cerca de nuestra variedad de interés $\mathbb{M}$. Para ver esto, supongamos que $\mathbb{M}$ es una variedad compacta $d-$dimensional y supongamos que el soporte de P es $\mathbb{M}$. Sea p la densidad de P respecto a la medida de Hausdorff en $\mathbb{M}$. En el caso particular en el que P es la distribución uniforme, para $t > 0$ suficientemente pequeño los conjuntos de nivel $L_p(t) = \{x : p(x) > t\}$ de p son exactamente $\mathbb{M}$. Entonces, si $\mathcal{P}$ es el diagrama de persistencia respecto a los conjuntos de

nivel $L_p(t)$ y $\mathcal{Q}$ el diagrama de persistencia generado con base en la función distancia $d_{\mathbb{M}}$, entonces los puntos de $\mathcal{P}$ están en biyección con los generadores de $H(\mathbb{M})$ al igual que los puntos con persistencia alta en $\mathcal{Q}$. Por ejemplo, supongamos que $\mathbb{M}$ es una circunferencia en el plano con radio r, es decir,

$$\mathbb{M} = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = r^2\}.$$

Entonces $\mathcal{Q}$ tiene dos puntos, uno en $(0, \infty)$ que representa la componente conexa y otro que representa el ciclo en $(0, r)$. Por otro lado, $\mathcal{P}$ tiene también dos puntos en $(0, 1/(2\pi r))$, pues $1/(2\pi r)$ es el valor máximo que toma la densidad. Si P no es la distribución uniforme pero tiene una densidad suave y acotada en $(0, \infty)$ entonces existe un intervalo $[a, A]$ tal que $L_p(t) \simeq \mathbb{M}$ (Homotópicamente equivalentes) para $a \leq t \leq A$. Se pueden encontrar ejemplos en los cuales ningún conjunto de nivel es igual a $\mathbb{M}$, pero en dichos casos ninguno de los métodos resulta bueno para capturar su homología.

Por otro lado, supongamos que tenemos ruido en la muestra: observamos $Y_1, Y_2, ..., Y_n$, donde $Y_j = X_j + \sigma\varepsilon_j$ y $\varepsilon_j \sim \Phi$. Supongamos que $X_1, X_2, ..., X_n \sim Q$ donde Q tiene soporte en $\mathbb{M}$. La distribución P de Y_j tiene densidad

$$p(y) = \int_{\mathbb{M}} \phi_\sigma(y - u) dQ(u),$$

donde ϕ es la densidad de ε_j y $\phi_\sigma(z) = \sigma^{-D}\phi(z/\sigma)$. En este caso no habrá conjuntos de nivel tales que $L_p(t)$ sea igual a $\mathbb{M}$. Sin embargo, conforme ϕ sea suave y σ pequeño habrá un intervalo de valores $a \leq t \leq A$ tal que $L_p(t) \simeq \mathbb{M}$. Además, el método por estimación de densidad es muy insensible a ruido en la muestra. Sea $P = \pi U + (1 - \pi)Q$, donde Q tiene soporte en $\mathbb{M}$, $\pi > 0$ es una constante (pequeña) y U es una función de distribución suave en $\mathbb{R}^D$. Tenemos que la distancia cuello de botella entre $\mathcal{P}_P$ y $\mathcal{P}_Q$ es a lo más π, pues por el teorema de estabilidad

$$\begin{aligned} W_\infty(\mathcal{P}_P, \mathcal{P}_Q) &\leq ||P - Q||_\infty \\ &= \pi||U - Q||_\infty \\ &\leq \pi. \end{aligned}$$

De aquí que el diagrama de persistencia es afectado de manera insignificante por ruido en los datos.

Ahora bien, el estimador estándar de para p_h es

$$\widehat{p}_h(x) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{h^D}K\left(\frac{||x-X_j||_2}{h}\right).$$

Es fácil ver que $\mathbb{E}\left[\widehat{p}_h(x)\right] = p_h(x)$, y de aquí que $\widehat{p}_h(x)$ es un estimador insesgado de $p_h(x)$. También, si p es continua entonces el estimador $\widehat{p}_h(x)$ es consistente para p, siempre y cuando el ancho de banda cumpla que $h = h_n \to 0$ y $nh_n \to \infty$ cuando $n \to \infty$. De cualquier manera, para estudiar homología, no es necesario hacer tender h a cero. En efecto, se pueden hacer inferencias topológicas de manera precisa si consideramos un ancho de banda positivo. Considerar $h > 0$ tiene un efecto suavizador en la densidad, pero los conjuntos de nivel siguen capturando la información topológica correcta.

Es necesario también puntualizar que los diagramas $\mathcal{P}_h$ y $\widehat{\mathcal{P}}_h$, el diagrama obtenido con base en $\widehat{p}_h$, son más robustos y se comportan estadísticamente mejor que el complejo de Čech construido mediante los datos. En el lenguaje de topología computacional, $\mathcal{P}_h$ es considerado como una simplificación topológica de $\mathcal{P}$. Aunque $\mathcal{P}_h$ puede omitir características que estén presentes en $\mathcal{P}$, es mucho más estable. Es por esto que es más acertado trabajar $\mathcal{P}_h$ para estudiar la homología de $\mathbb{M}$ si consideramos que nuestra muestra está contaminada por ruido.

Recordemos del teorema de estabilidad que

$$W_\infty(\widehat{\mathcal{P}}_h, \mathcal{P}_h) \leq ||\widehat{p}_h - p_h||_\infty.$$

Con esto, es suficiente encontrar c_n tal que

$$\lim_{n\to\infty}\sup \mathbb{P}(||\widehat{p}_h - p_h||_\infty > c_n) \leq \alpha.$$

Banda de muestra finita. Supongamos que el soporte de P está contenido en $\mathcal{X} = [-C, C]^D$. Sea p la densidad de P, sea K un kernel y escojamos un ancho de banda h. Sea

$$\widehat{p}_h(x) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{h^D}K\left(\frac{||x-X_j||_2}{h}\right)$$

el estimador kernel de la densidad y sea

$$p_h(x) = \frac{1}{h^D}\int_X K\left(\frac{||x-u||_2}{h}\right)dP(u)$$

la media de $\widehat{p}_h$.

Lema 2.1.8. *Supongamos que* $\sup_x K(x) = K(0)$ *y que* K *es L-Lipschitz, es decir que* $|K(x) - K(y)| \leq L \|x - y\|_2 \ \forall x, y$. *Entonces*

$$\mathbb{P}(||\widehat{p}_h - p_h||_\infty > \delta) \leq 2 \left(\frac{4CL\sqrt{D}}{\delta h^{D+1}} \right)^D \exp\left(-\frac{n\delta^2 h^{2D}}{2K^2(0)} \right).$$

La demostración de este lema está detallada en Fasy et al. (2014), la cual está basada principalmente en la desigualdad de Hoeffding. Esta desigualdad nos dice que si $X_1, X_2, ..., X_n$ son variables aleatorias independientes tal que $X_j \in [a_j, b_j]$ casi seguramente, entonces

$$\mathbb{P}\left[|\overline{X}_n - E(\overline{X}_n)| \geq t \right] \leq 2 \exp\left[-\frac{2n^2 t^2}{\sum_{j=1}^n (b_j - a_j)^2} \right],$$

donde $\overline{X}_n = \frac{1}{n} \sum_{j=1}^n X_j$. Por otro lado, es posible obtener una cota más fina haciendo uso de la desigualdad de Bernstein la cual establece que

$$\mathbb{P}\left[|\overline{X}_n - E(\overline{X}_n)| \geq t \right] \leq 2 \exp\left[-\frac{nt^2}{2(V + tM/3)} \right],$$

donde $V = \text{Var}(\overline{X}_n)$ y M es tal que $|X_j| \leq M$ casi seguramente. El problema de utilizar esta última desigualdad radica en que es necesario hacer la estimación de V. Gracias a este lema y al siguiente corolario es que podemos asegurar que δ_n es un cuantil para $W_\infty(\widehat{\mathcal{P}}_h, \mathcal{P}_h)$ de probabilidad α.

Corolario 2.1.9. *Sea* δ_n *la solución de la ecuación*

$$2 \left(\frac{4CL\sqrt{D}}{\delta_n h^{D+1}} \right)^D \exp\left(-\frac{n\delta_n^2 h^{2D}}{2K^2(0)} \right) = \alpha.$$

Entonces

$$\sup_{P \in Q} \mathbb{P}(W_\infty(\widehat{\mathcal{P}}_h, \mathcal{P}_h) > \delta_n) \leq \sup_{P \in Q} \mathbb{P}(||\widehat{p}_h - p_h||_\infty > \delta_n) \leq \alpha,$$

donde Q *es el conjunto de todas las medidas de probabilidad con soporte en* $\mathcal{X}$.

Consideremos ahora una banda diferente. Computacionalmente, la homología persistente de los conjuntos de nivel de $\widehat{p}_h$ está basada en una aproximación lineal a pedazos. Sea $R \subset \mathbb{R}^D$ una rejilla y hagamos una triangulación sobre ella. Definimos $\widehat{p}_h^*$ de la siguiente manera: para $x \in R$, $\widehat{p}_h^*(x) = \widehat{p}_h(x)$, para x fuera de la rejilla definimos $\widehat{p}_h^*(x)$ con una interpolación lineal a través de la

triangulación. Sea $p_h^*(x) = \mathbb{E}\left[\widehat{p}_h^*(x)\right]$, lo que nos interesa es el diagrama de persistencia $\mathcal{P}_h^*$ generado por los conjuntos de nivel de $p_h^*(x)$. Lo que obtenemos es una aproximación de este $\widehat{\mathcal{P}}_h^*$ mediante $\widehat{p}_h^*(x)$. Como ya hemos mencionado,

$$W_\infty(\widehat{\mathcal{P}}_h^*, \mathcal{P}_h^*) \leq ||\widehat{p}_h^* - p_h^*||_\infty.$$

Sin embargo, debido a la naturaleza lineal a pedazos de estas funciones se tiene que

$$||\widehat{p}_h^* - p_h^*||_\infty \leq \max_{x \in R} |\widehat{p}_h^*(x) - p_h^*(x)|.$$

Lema 2.1.10. *Sea $N = |R|$ el tamaño de la rejilla. Entonces*

$$\mathbb{P}(||\widehat{p}_h^* - p_h^*||_\infty > \delta) \leq 2N \exp\left(-\frac{2n\delta^2 h^{2D}}{2K^2(0)}\right).$$

Con esto, si

$$\delta_n = \left(\frac{K(0)}{h}\right)^D \sqrt{\frac{1}{2n} \log\left(\frac{2N}{\alpha}\right)},$$

entonces

$$\mathbb{P}(||\widehat{p}_h^* - p_h^*||_\infty > \delta) \leq \alpha.$$

Esta banda puede ser sustancialmente más estrecha que la descrita primero, siempre y cuando no utilicemos una rejilla demasiado fina, es decir, con un N muy grande. En cierto sentido, nos favorece saber que nuestras inferencias topológicas toman lugar en una escala finita. Tomando esto en cuenta, si consideramos una rejilla adecuada podemos obtener una banda más angosta que la anterior.

Banda asintótica. Se puede obtener una banda más angosta haciendo uso de teoría asintótica. El procedimiento más sencillo para nuestro objetivo es el *bootstrap*. Este procedimiento fue introducido en Efron (1979) en donde se da una técnica alternativa al ya conocido en ese entonces método Jackknife. Además muestran que el método Jackknife puede ser visto como una expansión lineal para aproximar el método bootstrap. A continuación explicaremos brevemente en lo que consiste el método bootstrap.

Sea $x_1, x_2, ..., x_n$ una muestra aleatoria de $X_1, X_2, ..., X_n$ que son variables aleatorias i.i.d. con distribución común $X_1 \sim F$. Supongamos que nos interesa conocer un cuantil de la distribución

de la estadística $T = T(X_1, X_2, ..., X_n)$. Para esto lo que haremos es obtener nuevas muestras $x_1^b, x_2^b, ..., x_n^b$ con la distribución empírica $\widehat{F}_n$, la cual es una aproximación de la distribución F. Así,

$$x_1^b, x_2^b, ..., x_n^b \sim \widehat{F}_n \approx F.$$

Recordemos que

$$\widehat{F}_n(t) = \frac{1}{n}\sum_{j=1}^{n} 1_{\{x_j \leq t\}}.$$

Es decir, a cada variable X_j, $\widehat{F}_n$ le da un peso de $1/n$. De aquí que para obtener $x_1^b, x_2^b, ..., x_n^b \sim \widehat{F}_n$ basta con generar una muestra aleatoria con reemplazo de $\{x_1, x_2, ..., x_n\}$. Con base en esta muestra calculamos $T_b = T(x_1^b, x_2^b, ..., x_n^b)$. Repetimos el proceso B veces y así podemos obtener $T_1, T_2, ..., T_B$. Con esto, obtenemos una aproximación de la función de distribución de T dada por

$$\widehat{F}_T(t) = \frac{1}{B}\sum_{j=1}^{B} 1_{\{T_j \leq t\}}.$$

Así, definimos el cuantil bootstrap de probabilidad α como $Z_\alpha^{\text{boot}} = \text{ínf}\{T_j : \widehat{F}_T(T_j) \geq \alpha\}$. Haciendo uso de este procedimiento, es posible obtener un cuantil aproximado para la estadística $\sqrt{nh^D}||\widehat{p}_h - p_h||_\infty$.

Sea $X_1', X_2', ..., X_n'$ una muestra de la distribución empírica $\widehat{F}_n$, y sea $\widehat{p}_h'$ el estimador de la densidad construido con base en dicha muestra. Definimos la medida aleatoria

$$J_n(t) = \mathbb{P}\left(\sqrt{nh^D}||\widehat{p}_h - \widehat{p}_h'||_\infty > t | X_1, X_2, ..., X_n\right)$$

y el cuantil bootstrap $Z_\alpha = \text{ínf}\{t : J_n(t) \leq \alpha\}$. Notemos que el signo de menor o igual (en lugar de mayor o igual) se debe a cómo definimos $J_n(t)$, que cuando n tiende a infinito converge a $1 - \mathbb{P}(\sqrt{nh^D}||\widehat{p}_h - p_h||_\infty \leq t)$.

Teorema 2.1.11. *Cuando $n \to \infty$,*

$$\begin{aligned}\mathbb{P}\left[W_\infty(\widehat{\mathcal{P}}_h, \mathcal{P}_h) > \frac{Z_\alpha}{\sqrt{nh^D}}\right] &\leq \mathbb{P}(\sqrt{nh^D}||\widehat{p}_h - p_h||_\infty > Z_\alpha) \\ &= \alpha + O\left(\sqrt{\frac{1}{n}}\right).\end{aligned}$$

La demostración de este teorema viene de resultados estándar en bootstrap. Se recomienda al lector consultar Chazal et al. (2013) y Giné y Guillou (2002) para la caracterización de la demostración. Como es usual, aproximamos Z_α vía Monte Carlo. Sea $T = \sqrt{nh^D}||\widehat{p}_h - p_h||_\infty$ y generamos B muestras bootstrap $x_1^b, x_2^b, ..., x_n^b \sim \widehat{F}_n$, $b = 1, 2, ..., B$ y para cada una calculamos

$$T_b = \sqrt{nh^D}||\widehat{p}_h - \widehat{p}'_{h,b}||_\infty,$$

donde $\widehat{p}'_{h,b}$ es el estimador de la densidad construido a partir de $x_1^b, x_2^b, ..., x_n^b$. Ahora calculamos

$$Z_\alpha = \text{ínf}\{z : \frac{1}{B}\sum_{j=1}^{B} 1_{\{T_j > t\}} \leq \alpha\}.$$

Es importante resaltar que el procedimiento vía bootstrap no es sólo utilizable con la filtración basada en los conjuntos de nivel de la densidad. Es posible realizarlo con una filtración que utilice alguna otra función más general, que tal vez sea relevante para el caso de estudio en el que se trabaja. En efecto, si utilizamos la función f y un estimador $\widehat{f}$, lo que queremos es un cuantil para $\sqrt{nh^D}||\widehat{f} - f||_\infty$ en lugar de $\sqrt{nh^D}||\widehat{p}_h - p_h||_\infty$. Con ello, es posible obtener una banda de confianza para un diagrama generado con base en los conjuntos de nivel de una función más general f.

El paquete TDA de R cuenta con una gama de funciones relacionadas con este último método. Primero, *bootstrapBand* es una función que regresa el cuantil Z_α mediante bootstrap. Cabe mencionar que, como es sabido, los algoritmos que involucran métodos de remuestreo resultan ser muy exhaustivos. La función *bootstrapBand* permite hacer uso de cómputo en paralelo para ayudar a solucionar este aspecto. También, en este paquete se incluye la función *gridDiag*, la cual genera el diagrama de persistencia basado en los conjuntos de nivel de una función cualquiera f, evaluada previamente en una rejilla. Adicionalmente se tiene la función *bootstrapDiagram* obtiene un cuantil bootstrap para un diagrama generado por la función *gridDiag.* A continuación veremos algunos ejemplos sencillos en donde se prueban los primeros dos métodos así como la función *bootstrapBand.*

2.1.5. Implementación y ejemplos

Como se sugiere en Fasy et al. (2014) así como en la literatura de topología computacional, se probaron las implementaciones en algunos ejemplos sencillos. Para cada ejemplo se utilizó la

filtración de Vietoris-Rips y en el último también se utilizó la filtración por medio de conjuntos de nivel de la densidad estimada, así como el método bootstrap para obtener la banda de confianza. Para la estimación de la densidad se utilizó un kernel gaussiano con ancho de banda $h = 0.3$. Se usó el paquete TDA de R tanto para generar muestras uniformes en la circunferencia unitaria, como para generar los diagramas de persistencia. En cada diagrama, los puntos negros representan las componentes conexas y los triángulos rojos los ciclos (agujeros de dimensión uno). A pesar de la sencillez de los ejemplos, estos son buenos para ilustrar la metodología descrita.

Ejemplo 1. En la Figura 18 se muestran los métodos de submuestreo y de concentración de medida aplicados a una muestra uniforme de tamaño 500 en la circunferencia unitaria. Ambos métodos descartan las componentes conexas que surgen como ruido topológico y detectan una componente y un ciclo significativos. En este ejemplo, ambos métodos funcionan perfectamente.

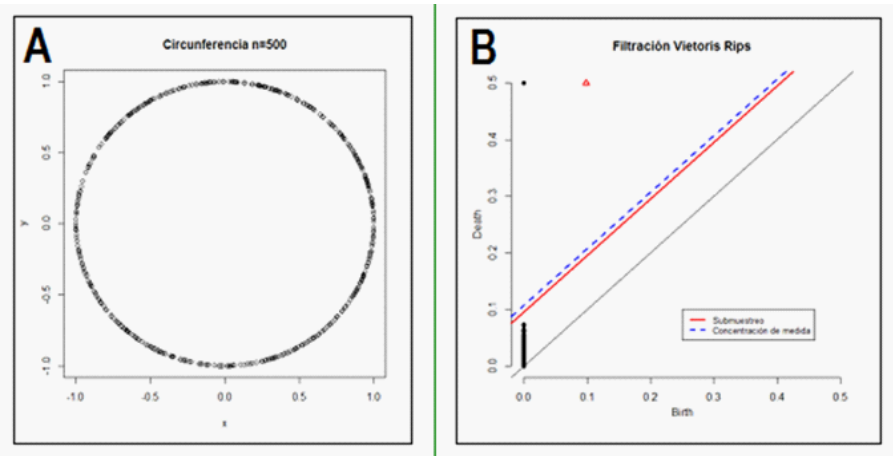

Figura 18. En A se ilustra la muestra de tamaño 500 sobre la circunferencia unitaria. En B se muestra el diagrama de persistencia con base en la filtración de Vietoris-Rips y las bandas de confianza por los métodos de submuestreo y de concentración de medida.

Ejemplo 2. En la Figura 19 se muestran los resultados de la implementación de los dos primeros métodos tomando una muestra uniforme en la circunferencia unitaria y agregándole cierto nivel de ruido normal (a cada punto). Como se puede observar, debido al ruido el diagrama de persistencia reporta ciclos y componentes conexas que en realidad no son parte del objeto teórico. Sin embargo, tanto el método por submuestreo como el de concentración de medida descartan dichas características y rescatan la información relevante verdadera que son una sola componente y un ciclo.

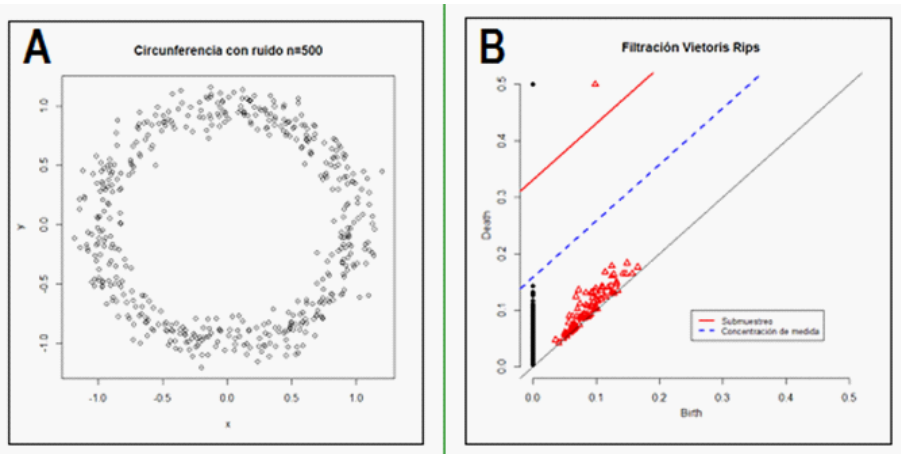

Figura 19. En A se ilustra la muestra de tamaño 500 con ruido en ella. En B se muestra el diagrama de persistencia con base en la filtración de Vietoris-Rips y las bandas de confianza por los métodos de submuestreo y de concentración de medida.

Ejemplo 3. En la Figura 20 se presenta una muestra de tamaño 800 sobre la unión de dos circunferencias. Se tomaron de manera independiente una muestra uniforme de tamaño 400 en cada circunferencia y se unieron estas dos muestras en una sola. Es fácil ver que la muestra completa no tiene distribución uniforme sobre el conjunto completo, en la parte de la unión de ambas circunferencias se presenta mayor concentración de puntos. Debido a esto, el método por concentración de medida detecta componentes conexas de más a pesar de que sí detecta correctamente los dos ciclos. Por otra parte, el método por submuestreo identifica correctamente el ruido topológico.

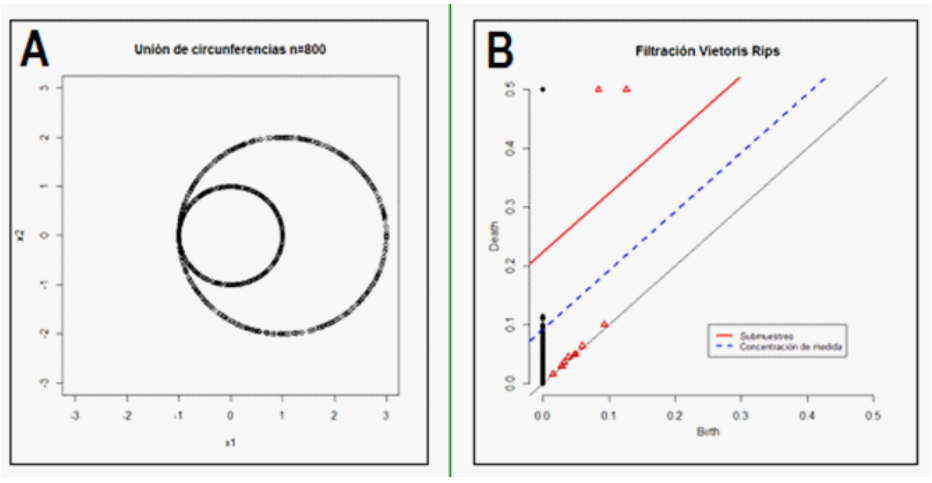

Figura 20. En A se ilustra la muestra de tamaño 800 sobre la unión de dos circunferencias. En B se muestra el diagrama de persistencia con base en la filtración de Vietoris-Rips y las bandas de

confianza por los métodos de submuestreo y de concentración de medida.

Ejemplo 4. En la Figura 21 se presenta una muestra uniforme sobre la circunferencia unitaria contaminada con algunos puntos atípicos. Se obtuvieron dos diagramas, uno con la filtración usual de Vietoris-Rips y otro mediante la estimación de la densidad. Para el primero, se muestra la implementación de los métodos de submuestreo y de concentración de medida. Como se puede observar, el método por concentración de medida detecta como señales verdaderas algunas que en realidad no lo son. Por su parte el método por submuestreo detecta de manera correcta sólo una componente y un ciclo. Por otro lado, en el diagrama con base en la estimación de densidad, la banda bootstrap (obtenida con el paquete TDA) logra capturar la información verdadera.

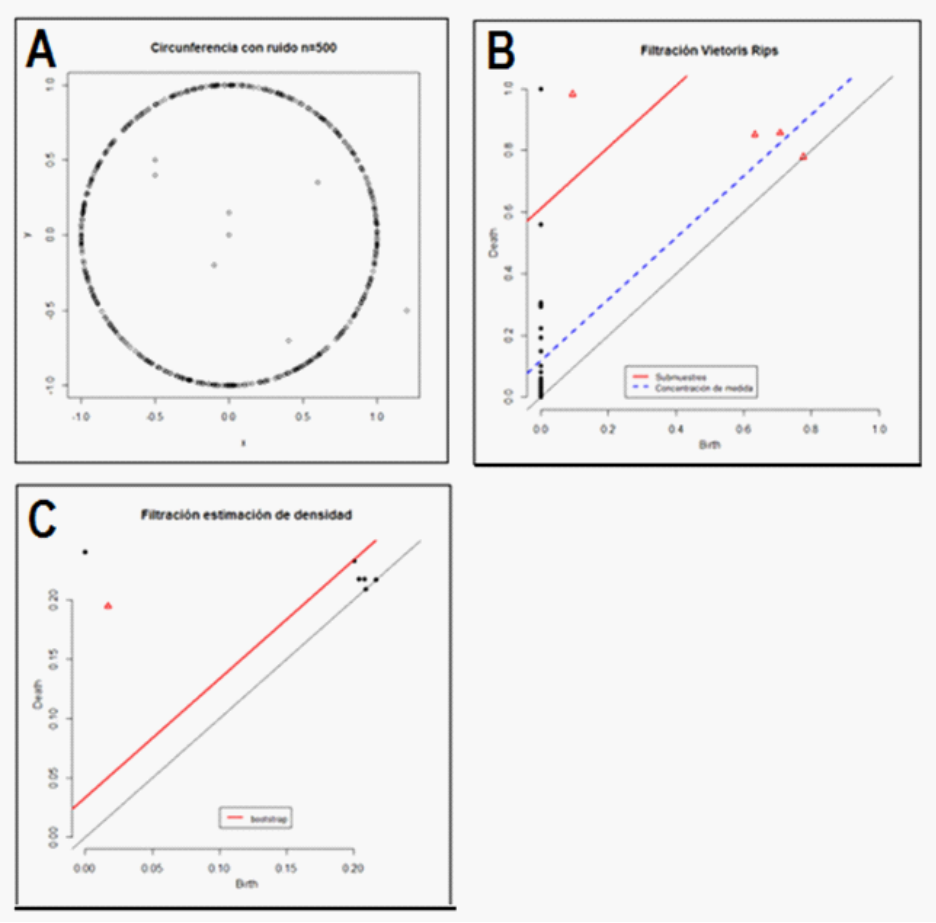

Figura 21. En A se ilustra la muestra de tamaño 500 sobre la circunferencia unitaria con algunos puntos atípicos. En B se muestra el diagrama de persistencia con base en la filtración de Vietoris-Rips y las bandas de confianza por los métodos de submuestreo y de concentración de

medida. En C se muestra el diagrama de persistencia basado en la estimación de la densidad, dotado con una banda de confianza por medio de bootstrap.

Como se describen en Fasy et al. (2014), se presentaron cuatro métodos para separar, en un diagrama de persistencia, ruido topológico de señales verdaderas en la homología de un objeto. Los primeros tres están basados en una función distancia a los datos y el último usa la estimación de la densidad. La ventaja de los primeros es que están conectados de una forma más directa con los datos, mientras que el último es más robusto a ruido y datos atípicos. Es necesario tomar en cuenta el contexto en el que se está aplicando la homología persistente para hacer uso de ellos.

A partir de estos resultados naturalmente surgen varias preguntas y líneas a seguir. A lo largo de este capítulo hemos hablado de incertidumbre en el espacio de diagramas de persistencia. Ideas similares a las descritas a lo largo de esta sección, se pueden utilizar para cuantificar la incertidumbre en los códigos de barras. Por otro lado, existen otros parámetros topológicos de interés, tal es el caso del *grado p total de persistencia* definido por

$$\theta = 2 \sum d(x, \text{Diag})^p,$$

donde la suma es sobre puntos en el diagrama de persistencia cuya distancia a la diagonal es más grande que cierto umbral y Diag denota a la recta identidad. Es de interés construir intervalos de confianza para estimaciones de este tipo de parámetros. Por su parte, en el método por estimación de densidad, el ancho de banda óptimo es una pregunta abierta; se utilizó un ancho de banda fijo, pero uno adaptativo sería de utilidad para hacer inferencias más refinadas. Por último, es necesario también hacer experimentos con mayor detalle para investigar bajo qué condiciones los métodos descritos funcionan bien y bajo cuáles no. Se hace énfasis en estos cuestionamientos debido a que son potencialmente abordables en trabajos futuros.

2.2. Pruebas de hipótesis

La homología persistente es una herramienta primordial dentro del análisis topológico de datos. Se han desarrollado trabajos en donde se construyen estimadores para ciertas características de una colección de diagramas de persistencia. Por ejemplo, en Bubenik y Kim (2006) se estudian

los códigos de barras para ciertos casos particulares de la distribución P con la que se muestrea la nube de datos; o como describimos en la sección anterior, en Fasy et al. (2014) se describen métodos para generar un conjunto de confianza para los diagramas de persistencia. Sin embargo, hacer inferencia a partir de observaciones que son diagramas de persistencia es un tanto limitado hasta el momento. Turner (2013) describe la forma de definir los conceptos de media y mediana para diagramas de persistencia, asi como introduce el desarrollo del concepto de media de Fréchet. Pero en particular, surge la necesidad de construir pruebas que evalúen la evidencia en contra de la afirmación de que dos muestras surgen de una misma población o de un mismo proceso. Para esto, Robinson y Turner (2013) propone el uso de pruebas de permutación. Daremos una pequeña introducción a pruebas de permutación como se detalla en Good (1993) y Wasserman (2004). Se recomienda al lector consultar dichas referencias así como el Capítulo 6 de Welsh (1996), en donde se extienden las propiedades y se estudian a mayor profundidad estas pruebas.

2.2.1. Pruebas de permutaciones

Las pruebas de permutación son pruebas no paramétricas que comúnmente son utilizadas para averiguar si dos distribuciones son la misma. Estas pruebas absuelven de la necesidad de proponer un modelo paramétrico, proporcionando una estimación empírica de la distribución de la estadística de prueba bajo la hipótesis nula. Supongamos que tenemos dos muestras independientes $X_1, X_2, ..., X_m \sim F_X$, $Y_1, Y_2, ..., Y_n \sim F_Y$ y H_0 es la hipótesis de que las dos muestras están idénticamente distribuidas. Más precisamente, queremos probar la hipótesis

$$H_0 : F_X = F_Y \text{ vs. } H_1 : F_X \neq F_Y.$$

Sea $T(X_1, X_2, ..., X_m, Y_1, Y_2, ..., Y_n)$ una estadística de prueba acorde con el contexto del problema, por ejemplo

$$T(X_1, X_2, ..., X_m, Y_1, Y_2, ..., Y_n) = \left|\overline{X}_m - \overline{Y}_n\right|.$$

Sea $N = m+n$ y consideremos las $N!$ permutaciones que podemos formar con $X_1, ..., X_m, Y_1, ..., Y_n$. Con cada una de estas permutaciones calculemos la estadística T y llamemos $T_1, T_2, ..., T_{N!}$ a dichos valores. Bajo la hipótesis nula, se tiene que cada uno de los T_j es igual de probable que ocurra, pues todas las permutaciones son equiprobables. La distribución P_0 que le da igual probabilidad de

ocurrencia a cada T_j, es decir $1/N!$, se le conoce como *distribución de permutación*. Así, el $p-$valor se obtiene como

$$p\text{-valor} = \mathbb{P}_0(T > t_{\text{obs}}) = \frac{1}{N!}\sum_{j=1}^{N!} 1_{\{T_j > t_{\text{obs}}\}},$$

donde $t_{\text{obs}} = T(X_1, X_2, ..., X_m, Y_1, Y_2, ..., Y_n)$ es el valor de la estadística T observado. Notemos que un aspecto relevante de estas pruebas es que son "exactas" en el sentido de que no se basan en aproximaciones de teoría asintótica. Para una muestra grande, las pruebas de permutación arrojan resultados similares a las pruebas basadas en teoría asintótica. Pero para muestras pequeñas y a falta de seguridad en los supuestos de un modelo paramétrico, estas pruebas resultan muy útiles.

Como un ejemplo sencillo para ilustrar una prueba de permutación consideremos la muestra (de X y Y) $(X_1, X_2, Y_1) = (1, 5, 4)$ y sea

$$\begin{aligned} T(X_1, X_2, Y_1) &= \left|\overline{X}_m - \overline{Y}_n\right| \\ &= |3 - 4| \\ &= 1. \end{aligned}$$

Las permutaciones nos dan los siguientes resultados,

permutación	valor de T	probabilidad
$(1, 5, 4)$	1	1/6
$(1, 4, 5)$	2.5	1/6
$(5, 1, 4)$	1	1/6
$(5, 4, 1)$	3.5	1/6
$(4, 1, 5)$	2.5	1/6
$(4, 5, 1)$	3.5	1/6

De aquí que el $p-$valor es $\mathbb{P}(T > 1) = 4/6 = 2/3$. En este ejemplo fue sencillo evaluar las $N! = 6$ permutaciones, pero por lo general nos encontramos con una cantidad muy grande de permutaciones; tan solo si $N = 15$, $N! \approx 1.3 \times 10^{12}$. Es por esto que usualmente se aproxima el p-valor tomando una muestra uniforme de las permutaciones, de tamaño *grande*, y calculamos la fracción de veces con la que ocurre el evento $T_j > t_{\text{obs}}$. Se obtiene así el algoritmo comúnmente utilizado:

1. Calcular el valor observado de la estadística de prueba $t_{obs} = T(X_1, X_2, ..., X_m, Y_1, Y_2, ..., Y_n)$.

2. Aleatoriamente permutamos los datos y calculamos la estadística de prueba con los datos permutados.

3. Repetimos el paso anterior una cantidad grande B veces y obtenemos los valores $T_1, T_2, ..., T_B$.

4. El p-valor aproximado es
$$\frac{1}{B}\sum_{j=1}^{B} 1_{\{T_j > t_{obs}\}}.$$

Habiendo descrito lo que son las pruebas de permutación, hablaremos del enfoque que Robinson y Turner (2013) abordan para probar la hipótesis de que dos conjuntos de diagramas de persistencia corresponden al mismo objeto topológico.

2.2.2. Pruebas de hipótesis vía TDA

Desafortunadamente, el espacio de diagramas de persistencia es geométricamente muy complicado. En Turner (2013) se demuestra que el espacio de diagramas de persistencia dotado con la métrica (L_2-)Wasserstein, la cual definiremos formalmente a continuación, tiene una curvatura no acotada; además, la dimensión de este espacio es infinita. Es por esto que no es plausible hacer uso de algún modelo paramétrico para dotar a un conjunto de diagramas de persistencia y por ende no es posible hacer una prueba de hipótesis que requiere como suposición un modelo paramétrico subyacente. El enfoque que proponen en Robinson y Turner (2013) es el de definir una función de pérdida conjunta relevante (entre ambas muestras) y hacer uso de una prueba de permutación.

Sea D el espacio de diagramas de persistencia. Existen muchas maneras de definir distancias en este espacio, de manera análoga al espacio de funciones. En Turner (2013) se discute una familia de distancias que surge de manera natural con base en las métricas L_p en el espacio de funciones. Vamos a enfocarnos en particular en la distancia $2-$Wasserstein. Para definir la distancia $p-$Wasserstein, consideremos dos diagramas de persistencia X y Y. Sea $\phi : X \to Y$ una biyección. Recordemos que un diagrama de persistencia, a pesar de que tenga un número finito de puntos nacimiento-muerte de clases, contiene además la recta identidad; es por esto, que siempre es posible encontrar al menos una biyección entre cualesquiera dos diagramas de persistencia. Definimos la

distancia $p-$Wasserstein entre X y Y, $W_p(X,Y)$, como

$$W_p(X,Y) = \left[\inf_{\phi: X \to Y} \sum_{x \in X} \|x - \phi(x)\|_p^p \right]^{1/p},$$

donde el ínfimo se toma respecto a todas las biyecciones $\phi : X \to Y$ y $\|(a,b)\|_p = (|a|^p + |b|^p)^{1/p}$. Como ya mencionamos, en Turner et al. (2012) se detalla una prueba, con base en el algoritmo húngaro, de que el ínfimo sobre las biyecciones siempre es alcanzable. Gracias a este resultado podemos asegurar que $W_p(X,Y)$ está bien definida. No ahondaremos sobre las propiedades de esta distancia pues no es necesario para describir la prueba a continuación ni para los resultados que se derivan de ella. Como ya mencionábamos se recomienda consultar Cohen-Steinner et al. (2010) en donde se demuestran dos resultados de estabilidad en el mismo sentido que el teorema de estabilidad para distancia cuello de botella que se enunció anteriormente.

Como se ha venido ilustrando a lo largo de este capítulo, cada diagrama de persistencia está determinado por la función que genera la filtración sobre el objeto en estudio. Si de alguna manera nuestro método para generar la filtración es aleatorio, entonces esta aleatoriedad estará presente en el correspondiente conjunto de diagramas de persistencia generados. Una distribución en el conjunto de filtraciones nos genera una distribución en el conjunto de diagramas. Como resultado de esto tenemos a los diagramas de persistencia como elementos aleatorios.

Por ejemplo, consideremos a la filtración de Vietoris-Rips respecto a una muestra S_n como método para generar los diagramas. Como ya mencionamos en la Sección 1.2.3, en el contexto de funciones como base para general una filtración, haciendo uso de la distancia a la muestra observada generamos dicha filtración. Es decir, los conjuntos de nivel están dados por

$$L_\varepsilon = \{x \in \mathbb{R}^d : d_{S_n}(x) \leq \varepsilon\}.$$

Consideremos ahora un conjunto $X \subset \mathbb{R}^d$ con posible ruido en una vecindad de éste. Tomemos una muestra de n puntos (nube de datos) con la intención de aproximar a K. Respecto a esta muestra, se tiene una filtración de Vietoris-Rips. Si consideramos diferentes muestras bajo cierta distribución, obtendremos una distribución de nubes de datos de tamaño n, ρ_{nubes}. Cada nube de puntos a su vez genera una filtración y con esto tenemos una distribución $\rho_{\text{filtraciones}}$. Entonces, con base en esto tenemos una distribución de diagramas de persistencia ρ_X. En efecto, si tomamos m

muestras para formar nubes de datos, estamos tomando una muestra con distribución ρ_{nubes}. Ahora consideremos otro subconjunto $Y \subset \mathbb{R}^d$ del cual suponemos podemos muestrear puntos en él de manera similar. Si quisiéramos saber si X y Y son diferentes, es acertado proponer como hipótesis nula que son el mismo conjunto. Una condición necesaria, pero no suficiente, es que $\rho_X = \rho_Y$. Lo anterior implica que nuestra hipótesis nula para estudiar los diagramas de persistencia es que las distribuciones subyacentes de las cuales ρ_X y ρ_Y fueron obtenidas, son la misma. Teniendo esto en cuenta, continuemos con la descripción de la prueba en el caso en el que tenemos dos conjuntos de diagramas de persistencia y queremos probar la hipótesis de que describen el mismo objeto.

Supongamos que tenemos una colección de n diagramas de persistencia independientes y un esquema de etiquetas que divide la colección en dos grupos diferentes de tamaño n_1 y n_2: $\mathcal{X}_1 = \{X_{1,1}, X_{1,2}, ..., X_{1,n_1}\}$ y $\mathcal{X}_2 = \{X_{2,1}, X_{2,2}, ..., X_{2,n_2}\}$. Por ejemplo, si nos interesa la estructura topológica de cierto órgano con alguna patología en particular, podemos suponer que un grupo de diagramas correspondientes a órganos de pacientes enfermos y otro de pacientes que no lo están representan grupos diferentes. Nuestro objetivo es evaluar la evidencia de que los procesos que generan $\mathcal{X}_1$ y $\mathcal{X}_2$ son diferentes. Para lograr esto tomamos como hipótesis nula la aseveración de que las etiquetas son intercambiables; es decir, dicho de una manera informal, la configuración actual de etiquetas no es menos probable que la que hubiese ocurrido bajo un etiquetado aleatorio, en relación con la estadística de prueba. Un ejemplo de este razonamiento es el siguiente. Al tirar una moneda balanceada tres veces, cada configuración tiene la misma probabilidad (1/8); sin embargo, considerando como estadística de prueba el número de águilas en los tres lanzamientos, es menos probable obtener tres (cuya probabilidad es 1/8) que obtener dos (la cual es 3/8). De manera similar, esto se cumple para la prueba propuesta: a pesar de que cada configuración de etiquetas es igualmente probable bajo la hipótesis nula, aseguramos que varios de los etiquetados aleatorios producen un valor de la estadística de prueba que difiere mucho del valor observado.

Usualmente, las pruebas de permutación que se utilizan para comparar dos muestras numéricas se enfocan en alguna función de la distancia entre las medias de ambas. A pesar de que es posible definir el concepto de media de Fréchet en el espacio de diagramas de persistencia, es computacionalmente costoso obtener las medias en cada permutación, además de que esta media no necesariamente es única. En su lugar, se propone una función que mida las distancias dentro

de los grupos como estadística de prueba.

La pérdida conjunta de cualquier configuración de etiquetas puede ser expresada como la suma de las varianzas de cada grupo. Así, se plantea que tomando la media o la suma de las varianzas de los dos grupos sería una estadística de prueba sensible. La expresión usual de la varianza muestral (para un conjunto de números reales) es

$$\sigma_X^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x}_n)^2.$$

Sin embargo, es posible obtener el mismo resultado sin necesidad de calcular la media $\overline{x}_n$, mediante

$$\sigma_X^2 = \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2.$$

La ventaja de hacer uso de ésta última expresión es que las distancias $x_i - x_j$ sólo se deben calcular una vez y pueden ser guardadas en una matriz de distancias. Así, cuando se tiene algún etiquetado, sólo es necesario consultar esta matriz haciendo el trabajo de cómputo más ligero. Para diagramas de persistencia con etiquetado L dentro de los conjuntos $\mathcal{X}_1 = \{X_{1,1}, X_{1,2}, ..., X_{1,n_1}\}$ y $\mathcal{X}_2 = \{X_{2,1}, X_{2,2}, ..., X_{2,n_2}\}$ la estadística de prueba análoga es

$$\sigma^2_{\mathcal{X}_{1,2}}(L) = \sum_{m=1}^{2}\frac{1}{2n_m(n_m-1)}\sum_{i=1}^{n_m}\sum_{j=1}^{n_m}W_2(X_{m,i}, X_{m,j})^2,$$

donde $W_2(\cdot,\cdot)$ es la distancia 2−Wasserstein.

Esta estadística de prueba $\sigma^2_{\mathcal{X}_{1,2}}$ no es la única opción para la prueba de permutación, sino es un ejemplo de una función de pérdida conjunta. Una función de pérdida es una función que mapea los datos a números reales que intuitivamente representa algún "costo" asociado al modelo y a los datos. Una función de pérdida conjunta es la suma de dos o más funciones de pérdida que corresponde al costo total. Ya que existen diferentes métricas para el conjunto de diagramas de persistencia, se pueden definir diferentes funciones de pérdida conjunta. Algunas de ellas son de la forma

$$F_{p,q}(\mathcal{X}_1, \mathcal{X}_2) = \sum_{m=1}^{2}\frac{1}{2n_m(n_m-1)}\sum_{i=1}^{n_m}\sum_{j=1}^{n_m}W_p(X_{m,i}, X_{m,j})^q,$$

donde $p \in [1, \infty]$, $q \in [1, \infty)$ y $W_p(\cdot, \cdot)$ es la distancia $p-$Wasserstein, teniendo en cuenta que para $p = \infty$ ésta es la distancia cuello de botella. Ya que tenemos definida la estadística de prueba $T(L) = \sigma^2_{\mathcal{X}_{1,2}}(L)$, veamos un resultado importante y el algoritmo propuesto en Robinson y Turner (2013) para estimar el $p-$valor de la prueba de permutación.

Lema 2.2.1. *Sean $X_{1,1}, X_{1,2}, ..., X_{1,n_1}$ y $X_{2,1}, X_{2,2}, ..., X_{2,n_2}$ diagramas de persistencia independientes e idénticamente distribuidos (hipótesis nula) y sea α la proporción de todas las configuraciones de etiquetas L tales que $T(L) \leq T(L_{obs})$. Entonces para todo $p \in [0, 1]$ se tiene que $\mathbb{P}(\alpha \leq p) \leq p$.*

La demostración de este lema es muy sencilla de entender, pero no la incluiremos en esta tesis; para consultarla referimos al lector a Robinson y Turner (2013). Este resultado nos asegura que, bajo la hipótesis nula, un valor pequeño de α es muy poco probable. Así se tiene que valores cercanos a cero de α presentan evidencia en contra de la hipótesis nula. A continuación se muestra un algoritmo para obtener un estimador insesgado de α. Como ya mencionamos, computacionalmente es muy costoso realizar todas las permutaciones.

1. Calcular el valor observado de la estadística de prueba $t_{obs} = \sigma^2_{\mathcal{X}_{1,2}}(X_{1,1}, ..., X_{1,n_1}, X_{2,1}, ..., X_{2,n_2})$. Fijamos una cantidad grande B de veces para permutar los diagramas.

2. Aleatoriamente permutamos los datos y calculamos la estadística de prueba con los datos permutados.

3. Repetimos el paso anterior $B - 1$ veces y obtenemos los valores $T_1, T_2, ..., T_{B-1}$.

4. Una estimador insesgado para α es

$$\frac{1}{B} \sum_{j=1}^{B-1} 1_{\{T_j \leq t_{\text{obs}}\}}.$$

Notemos que α no es precisamente un p-valor, como podemos recordar, un p-valor está definido como

$$p - valor = \mathbb{P}(T \geq T_{\text{obs}}).$$

Además, una condición necesaria es que bajo la hipótesis nula debe tener distribución uniforme en el intervalo $[0, 1]$. Phipson y Smyth (2010) exponen las razones por las cuales α no es un verdadero p-valor. Sin embargo, la propiedad de que bajo la hipótesis nula $\mathbb{P}(\alpha \leq p) \leq p$ para cualquier $p \in [0, 1]$, nos permite hacer uso de α para dar información en contra de la hipótesis cuando éste toma un valor pequeño.

Cuando estudiamos pruebas de hipótesis, sale a relucir el concepto de potencia, esencial para la "evaluación" de las mismas. La potencia de una prueba describe el número de errores tipo I y tipo II en términos de la comparación del p-valor con el umbral α. Una prueba *más potente* es mejor rechazando la hipótesis nula cuando esta en realidad es falsa. El tamaño de una prueba se define como la potencia evaluada en la hipótesis nula. Para H_0 la hipótesis nula, y H_1 la hipótesis alternativa se define la *potencia* de la prueba de hipótesis como

$$\text{potencia} = \mathbb{P}(\text{rechazar } H_0 | H_1 \text{ es verdadera}).$$

En general, incluso cuando la prueba es insesgada (es decir, el tamaño se alcanza en la hipótesis nula), es deseable que la prueba sea *uniformemente más potente* (UMP) o al menos *localmente más potente* (LMP) en la región de la hipótesis nula. En este caso se utiliza una distribución de permutación para aproximar el comportamiento de la estadística de prueba bajo la hipótesis nula, así que no es posible tomar consideraciones de UMP o LMP. Sin embargo, es posible evaluar de manera aproximada diferentes estadísticas de prueba mediante simulaciones, tomando la proporción de veces que la hipótesis es rechazada a un nivel dado. Se pueden comparar estas proporciones entre diferentes estadísticas de prueba para tener una idea de los méritos relativos de las mismas. En Robinson y Turner (2013) ilustran un ejemplo sencillo al respecto: simulan puntos en dos objetos parecidos (una circunferencia y un anillo) ambos contaminados con la misma cantidad de ruido gaussiano, y para diferentes niveles de ruido se calcula la potencia de la prueba.

En Robinson y Turner (2013) se exponen dos aplicaciones específicas del enfoque. Por un lado, dado un conjunto de siluetas de objetos se plantea evaluar la hipótesis de que se trata del mismo, considerando siluetas rotadas y/o deformadas. También, aplicaron este enfoque a datos de imágenes de resonancia magnética; la idea era estudiar la homología de ciertas partes del cerebro para pacientes con trastorno por déficit de atención con hiperactividad (ADHD por sus siglas en

inglés). Para su objetivo tomaron en cuenta grupos de pacientes con el trastorno y grupos de control, tanto de hombres como de mujeres. La implementación de este algoritmo es sencilla, más no el exhaustivo trabajo computacional que esto conlleva. En el paquete TDA de R se encuentran disponibles funciones para calcular la distancia de Wassertein, la distancia cuello de botella, y como ya mencionamos, para generar diagramas de persistencia.

A partir de lo descrito a lo largo de esta sección, se puede extender el enfoque en diferentes sentidos, además de que surgen algunas líneas por las cuales continuar. Por un lado, se puede realizar la prueba con más de dos grupos de observaciones; si queremos examinar si $k > 2$ grupos difieren, se puede utilizar la estadística

$$\sigma^2_{\mathcal{X}_k} = \sum_{m=1}^{k} \frac{1}{2n_m(n_m-1)} \sum_{i=1}^{n_m}\sum_{j=1}^{n_m} W_2(X_{m,i}, X_{m,j})^2.$$

En Robinson y Turner (2013) se plantea el uso de pruebas de permutación cuando tenemos conjuntos de diagramas de persistencia como muestras, sin embargo, existe la posibilidad de explorar otros métodos no paramétricos. Por otro lado, es de interés estudiar bajo qué condiciones se puede garantizar la potencia de la prueba de distinguir diferentes distribuciones en los diagramas de persistencia, ya sea de manera teórica o por medio de simulaciones. Se hace mención de estos aspectos debido a su potencial desarrollo en un futuro.

Ahora bien, en el siguiente capítulo hablaremos sobre la aplicación de análisis topológico de datos que se realizó en ecología. Primero se dará una introducción a los conceptos básicos que se encuentran inmersos en el contexto de estudio para después pasar a describir el análisis que se llevó a cabo. Como ya hemos mencionado, la idea es estudiar la homología de nichos ecológicos de ciertas especies y evaluar si la estructura topológica de los mismos tiene relación con cierta clasificación. Además, se plantea estudiar dos regiones geográficas comparables en el sentido ecológico a través de la homología de sus ambientes disponibles y ver si se existen diferencias significativas. Los nichos ecológicos se han estudiado desde una perspectiva inferencial, con la intención de obtener una estimación del mismo. Se busca con esta aplicación ver si TDA tiene algo que aportar en el estudio, tanto de nichos ecológicos como de ambientes disponibles.

Capítulo 3

Aplicación a ecología

3.1. Nichos y distribuciones

Dentro de la disciplina de ecología se encuentra el área de biogeografía, la cual se encarga de estudiar la distribución de especies y ecosistemas en el espacio geográfico mundial y a través del tiempo. La distribución de especies está directamente relacionada con el concepto de nicho ecológico. Una especie tiene posibilidad de habitar cierta región geográfica si ésta tiene características ambientales que le favorecen. Para ligar una región geográfica con sus correspondientes características ambientales surge el concepto de ambiente disponible. Partiendo de esto, para estudiar la distribución de especies en cierta región geográfica o a escala global, se concibe el concepto de nicho ecológico inmerso en el ambiente disponible. A continuación formalizaremos estas nociones.

Dada una región geográfica G, cada punto $x \in G$ tiene asociado un vector de características ambientales correspondientes a n variables (Ver Figura 22).

Definición 3.1.1. *Se define el* ***espacio ambiental disponible*** *de la región G como*

$$E_G = \{(x_1, x_2, ..., x_n) \in \mathbb{R}^n : \exists x \in G \text{ tal que } x \text{ tiene asociado a } (x_1, x_2, ..., x_n)\}.$$

Si definimos la función $f : G \to \mathbb{R}^n$ respecto a esta asignación de características ambientales, tenemos que E_G es la imagen de G bajo f; ver Figura 22.

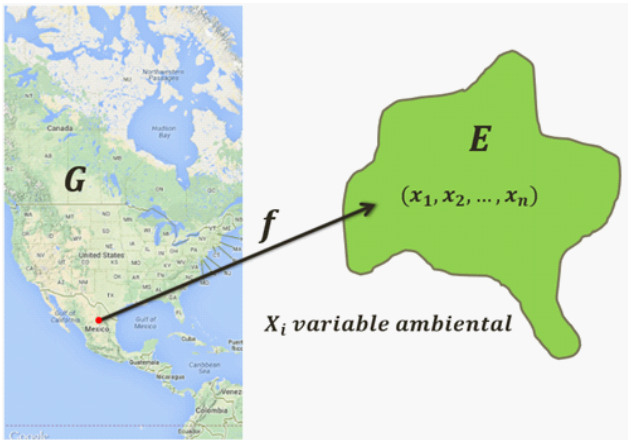

Figura 22. Espacio ambiental disponible E de la región geográfica G.

Esta caracterización del espacio ambiental disponible, como función del espacio geográfico, será importante para establecer filtraciones del espacio ambiental haciendo uso del espacio geográfico G. El espacio E_G guarda una relación importante con la distribución de individuos dentro del espacio G, así como también en la *abundancia* y *riqueza* de especies. La abundancia de una especie es la cantidad de individuos que hay de la misma en un ecosistema o región geográfica; por otro lado, la riqueza de especies se define como la cantidad de especies diferentes que viven en un mismo lugar. Se busca explorar con las herramientas que nos proporciona TDA si la estructura topológica del ambiente disponible está ligada a estos conceptos ecológicos.

Cuando nos referimos al *lugar* de una especie podemos referirnos, en un sentido ecológico, a varios aspectos como lo son el hábitat, la ubicación geográfica, el papel que desempeña dentro de un ecosistema, las interacciones con otros factores bióticos y abióticos, etc. El concepto de *nicho* busca describir el lugar que ocupa una especie dentro de un espacio de características ambientales que le permitan subsistir. A lo largo de la historia se ha generado una discusión de lo que el término nicho debe englobar. Grinell (1917), estudiando el cuitlacoche californiano (una especie de ave) definió su nicho como los requerimientos climáticos y de entorno expresados geográficamente; Elton (1924) define el nicho de un animal como el lugar que ocupa en el ambiente biótico y su relación con la comida y sus enemigos naturales. Estas definiciones nos ejemplifican la ambigüedad que se presenta al determinar las características en las que debe estar basado el término; la primera nos refiere a un espacio geográfico y la segunda a las relaciones y las implicaciones que se tienen dentro del ecosistema. Por otro lado, Hutchinson (1957) da una definición más consistente: un

hipervolumen de variables ambientales, todo punto del cual corresponde a un estado del ambiente el cual permite a la especie existir indefinidamente. A partir de esto podemos dar una definición del concepto en un lenguaje matemático.

Definición 3.1.2. *Dadas n variables ambientales de interés para una especie. Se define el* ***nicho ecológico fundamental*** *de dicha especie como el subconjunto* $N \subset \mathbb{R}^n$,

$$N = \{x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n : \text{ la especie puede subsistir con las características } x_i\}.$$

Es importante recalcar tres puntos en esta definición. Las variables X_j representan variables como la temperatura, la precipitación anual, la humedad relativa, el tipo de clima, el tipo de relieve, etc. Con esto tenemos que sus rangos R_j son subconjuntos de $\mathbb{R}$, uniones de intervalos o un conjunto finito de categorías. Por otro lado, la selección de variables que deben tomarse en cuenta sigue siendo algo un tanto ambiguo, pero esta confusión se elimina con el contexto en el que estemos situados y el objetivo con el que queremos estudiar el nicho; depende también de la especie en cuestión, habrá variables que son más significativas, en el sentido biológico, que otras. Por último, notemos que el subconjunto N es un subconjunto de características de las cuales no necesariamente todas se encuentran disponibles en el planeta Tierra. Es por ello que surge la siguiente definición.

Definición 3.1.3. *Dadas n variables ambientales de interés para una especie. Se define el* ***nicho ecológico existencial*** *de dicha especie como el subconjunto* $N_E \subset \mathbb{R}^n$,

$$\begin{aligned} N_E \;=\; & \{x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n : \text{ la especie puede subsistir con las características } x_i \text{ y existe} \\ & \text{algún punto en la Tierra con dichas características}\}. \end{aligned}$$

Actualmente los problemas sobre nichos ecológicos se sitúan en la tarea de inferirlo como subconjunto de $\mathbb{R}^n$, ya sea el nicho fundamental o el existencial. Referimos al lector a Peterson et al. (2006), en su Capítulo 1 describen la teoría subyacente a los nichos ecológicos y a la distribución geográfica de especies; en el Capítulo 2 exponen los procedimientos desarrollados en la práctica para la modelación de estos objetos y se adentran en los métodos comúnmente utilizados. Por su

parte, para conocer más sobre estos métodos para estimación de nichos también se recomienda Ortega-Huerta y Peterson (2008). En esta referencia se hace una discusión de seis de ellos: BioMapper, FloraMap, Domain, Weights of evidence, GARP y MaxEnt. Para ejemplificar la importancia de estudiarlos, introduzcamos los conceptos de *distribución realizada* y *distribución potencial* de especies.

Definición 3.1.4. *Dada una región geográfica G, la* ***distribución realizada*** *de una especie en esta región es el área en la que se encuentra presente la especie. La* ***distribución potencial*** *de una especie, con nicho N, se define como el subconjunto de G*

$$G \cap f^{-1}(N \cap E_G),$$

que son los puntos en G que tienen características favorables para la especie.

El investigar el nicho ecológico y el ambiente disponible de alguna región geográfica de interés nos permite estudiar la distribución potencial de dicha especie en la región en cuestión. La distribución realizada es un concepto primordial para los métodos de estimación de nichos. Como se explica en el Capítulo 2 de Peterson et al. (2011), estos métodos se basan principalmente en datos de presencia-ausencia y de sólo presencia, los cuales, como su nombre lo dice, son obtenidos con base en la distribución realizada de la especie en estudio. Por otra parte, el Capítulo 3 de Peterson et al. (2011) está destinado a exponer varias de las áreas en las que estudiar nichos ecológicos es de gran interés. A continuación daremos algunos ejemplos de dichas áreas y enunciaremos algunos trabajos específicos.

Dentro del estudio sobre la biodiversidad, los modelos para nichos proveen una base para la inferencia de la diversidad de especies y su distribución en el espacio geográfico y en el tiempo. La aplicación tal vez más simple en este ámbito es la idea de utilizar dichos modelos para guiar al investigador hacia el descubrimiento de poblaciones de especies que hasta ahora eran desconocidas, esto gracias a la inferencia de la distribución potencial. En Guisan et al. (2006) se describe la estimación de la distribución potencial de una especie rara de planta en Suiza (*Eryngium alpinum*), con lo cual revelaron la existencia de varias poblaciones que se desconocían hasta el momento,

mejorando así la eficiencia de su búsqueda. Por otra parte, una de las investigaciones más sorprendentes en esta área es la que se presenta en Raxworthy et al. (2003) sobre el análisis de camaleones en Madagascar. Los autores propusieron modelos de nichos para 11 especies de camaleón, y para cada especie identificaron áreas en la estimación en las cuales no se tenía conocimiento de presencia. En estudios de campo subsecuentes en estas áreas se encontraron siete nuevas especies de camaleón, en contraste con los estudios en otras áreas, los cuales revelaron a lo más una nueva especie. Este par de ejemplos señalan la importancia de considerar el análisis de la distribución potencial de especies con la intención de descubrir mayor biodiversidad.

En el campo de la conservación biológica, la tarea principal es el de brindar orientación científica para detener o relentizar la curva de extinción y degradación de la biodiversidad del planeta. Para lograr este objetivo, los biólogos buscan contestar algunas preguntas fundamentales tales como "¿qué conservar?", "¿dónde conservarlo?" y "¿cómo conservarlo?". Con el estudio del nicho existencial y con ello la distribución potencial de especies, los modelos para nichos proporcionan gran ayuda en dicho quehacer sobre todo para resolver el *qué* y el *dónde* en estas cuestionantes. En Rojas-Soto et al. (2008) se estudió al gorrión de la Sierra Madre (*Xenospiza bailey*) para evaluar sus características distribucionales. En particular para esta especie, una cuestión a resolver era si en verdad las tres poblaciones conocidas en México estaban separadas o bien, si existían poblaciones adicionales situadas en regiones comprendidas entre las tres. Por otro lado, se han utilizado los modelos de nichos para la agilización en la reinserción de especies en peligro de extinción. En Martínez-Meyer (2006) evaluaron lugares en México para la reintroducción potencial del cóndor californiano (*Gymnogyps californianus*) y del lobo mexicano (*Canis lupus baileyi*); en la Figura 23 se muestra el resultado para el lobo mexicano en el área de Baja California.

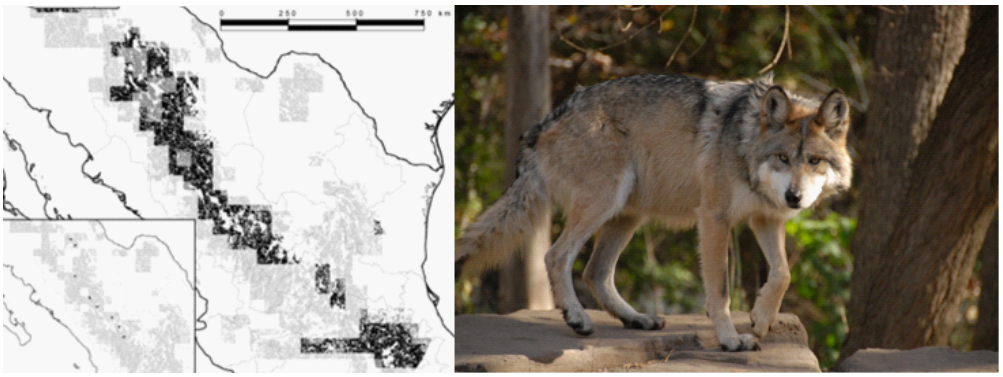

Figura 23. En la parte izquierda, imagen tomada de Martínez-Meyer et al. (2006), se muestra la distribución potencial estimada para el lobo mexicano. En gris claro se muestran las regiones que son propicias pero no lo serán en el futuro, en gris oscuro están las áreas que sólo lo serán en el futuro, y por último en negro están las regiones que presentan factibilidad para la especie tanto en el presente como en el futuro. A la derecha se muestra un ejemplar de la especie.

Por otro lado, las especies invasivas representan un fenómeno mundial con consecuencias a escalas colosales, tanto en el ámbito biológico como en el económico. En Benedict et al. (2007) se hace un análisis de la invasión mundial del mosquito tigre (*Aedes albopictus*), una especie considerada altamente dañina, con la intención de descubrir las regiones potencialmente habitables en América del Norte. También, en el estudio de transmisión de enfermedades se ha introducido el uso de modelos de nichos para predecir regiones potenciales de hábitat. En Peterson et al. (2005) se estudió al mosquito del Dengue (*Aedes aegypti*), y se estimó la distribución potencial del mismo a lo lardo de todo México, esto con la intención de proporcionar información para el desarrollo de políticas en salud pública; ver Figura 24. En Pimentel et al. (2004) exponen las áreas económicas que mayor afectan tales como la productividad agrícola y la transmisión de enfermedades, además calculan que el costo anual de la invasión de especies es de aproximadamente $120 millones de dólares tan sólo en los Estados Unidos. Es por esto que existe mucha actividad de investigación en este ámbito.

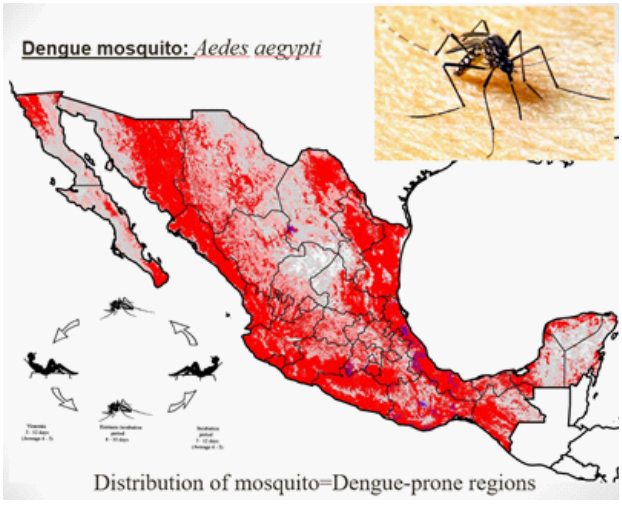

Figura 24. Imagen tomada de Peterson et al. (2005). En escala de rojo se muestra la distribución potencial del mosquito del dengue, siendo más propicias las regiones más oscuras.

Como se acaba de describir, el poder inferir el nicho ecológico de especies y con ello estudiar la distribución potencial de las mismas, representa una enorme ventaja en muchas áreas de la ecología. En cambio, estudiar el nicho desde el enfoque de homología, la estructura topológica del mismo podría revelar aspectos importantes de la especie en cuestión. Dos cualidades (contrarias) determinantes para el desarrollo de una especie es el de ser *generalistas* y *especialistas*; una especie generalista es aquella que presenta mayor tolerancia a cambiar de ambiente mientras que una especialista es una especie la cual tiene un rango pequeño de características que le permiten subsistir. Específicamente surge la idea de explorar la relación entre esta clasificación de especies y la homología de los nichos vistos como espacios topológicos.

En conjunto con el Dr. Enrique Martínez Meyer y la M. C. Edith Calixto Pérez se planeó abordar el estudio de siete especies de las cuales, cuatro son generalistas y tres especialistas. Se buscaba encontrar relación entre esta clasificación y la complejidad en sentido de homología que presentan los nichos de las especies. Por otro lado, se planteó la comparación de dos regiones geográficas por medio de sus ambientes disponibles, se tomaron regiones que se consideran comparables en el sentido ecológico. Un poco más adelante describiremos a detalle el planteamiento del estudio en ambos casos. A continuación describiremos la base de datos de la cual fueron extraídos los datos analizados, así como las consideraciones que fueron estipuladas en el trabajo.

3.2. Datos

3.2.1. Base de datos

Worldclim, como lo mencionan en su sitio web http://www.worldclim.org/, es una base de datos de capas de clima (aproximadas por rejillas del espacio geográfico mundial) con una resolución espacial de aproximadamente un kilómetro cuadrado, en su partición más fina. Es decir, dichas capas de clima representan una aproximación del ambiente disponible mundial. Las variables con las que cuenta la base de datos son derivadas de la temperatura mensual y de la precipitación con la intención de obtener variables que sean biológicamente más relevantes. En Guisan y Zimmermann (2000) se explican los distintos factores que actuan sobre una especie para poder estudiarla. Si bien existen 3 tipos de factores, los directos, los indirectos y los recursos, los más importantes

(los directos) en el caso de especies animales son tales como la precipitación, la temperatura, la nubosidad, la radiación, viento, entre otros. La tabla que se muestra en la Figura 25 muestra la lista completa de variables con las que cuenta Worldclim. Parte de estos datos son los que se utilizaron para hacer el análisis que describiremos en la siguiente sección.

BIO1	**Temperatura media anual**
BIO2	Rango medio diurno (Media mensual*(temp máx-temp min)
BIO3	Isotermalidad ((BIO2/BIO7)*100)
BIO4	Temperatura estacional (desviación estándar*100)
BIO5	Temperatura máxima del mes más cálido
BIO6	Temperatura mínima del mes más frío
BIO7	**Rango de la temperatura anual (BIO5-BIO6)**
BIO8	Temperatura media del cuarto más húmedo
BIO9	**Temperatura media del cuarto más seco**
BIO10	Temperatura media del cuarto más cálido
BIO11	Temperatura media del cuarto más frío
BIO12	**Precipitación anual**
BIO13	Precipitación del mes más húmedo
BIO14	Precipitación del mes más seco
BIO15	Precipitación estacional (coeficiente de variación)
BIO16	Precipitación del cuarto más húmedo
BIO17	Precipitación del cuarto más seco
BIO18	Precipitación del cuarto más cálido
BIO19	Precipitación del cuarto más frío

Figura 25. Variables ambientales con las que cuenta la base de datos Worldclim. Las variables resaltadas en negro son las que fueron utilizadas para el análisis de nichos ecológicos.

Para el análisis de los siete nichos, los biólogos recomendaron las variables: temperatura media anual (BIO1), el rango de temperatura anual (BIO7), temperatura media del cuarto más seco (BIO9) y la precipitación anual (BIO12), las cuales tienen mayor relevancia para las especies en cuestión. Esto fue basado en su experiencia propia en el estudio de estas especies, así también como la utilización del software MaxEnt (Máxima entropía) para la estimación del nicho. En general el método por máxima entropía es un algoritmo utilizado comúnmente en el área de aprendizaje máquina para obtener predicciones o hacer inferencias de información incompleta. En Phillips et al. (2006) muestran las principales funciones de dicho software. Utiliza datos de sólo presencia pero se pueden utilizar datos de presencia y ausencia; las variables ambientales pueden ser continuas o categóricas, y además MaxEnt puede incorporar interacciones entre variables. Algoritmos determinísticos y eficientes permiten estimar una distribución de probabilidad por

máxima entropía, que debido a su definición matemática es posible hacer una interpretación de cómo las variables ambientales actúan en el modelo. A nivel mundial, la temperatura media anual (BIO1) y la precipitación (BIO12) se ilustran en la Figura 26.

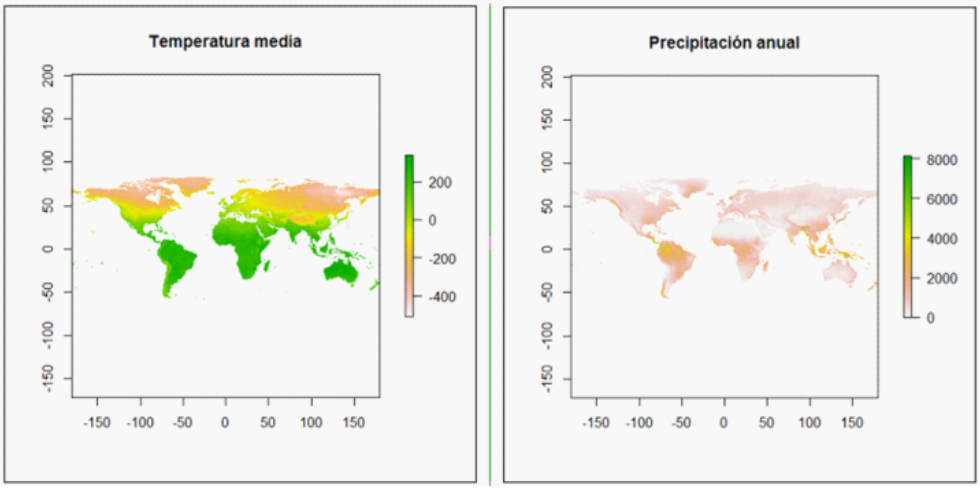

Figura 26. Mapas con la ilustración de la temperatura media mensual y la precipitación anual en el mundo.

La base de datos se encuentra dividida en tres tipos de condiciones climáticas respecto a la época. Cuenta con las condiciones actuales, las cuales fueron obtenidas mediante interpolaciones de datos observados en años que comprenden desde 1950 al 2000. También contiene extrapolaciones derivadas de simulaciones haciendo uso de modelos de clima global, correspondiente a circunstancias climáticas del Holoceno (periodo alrededor de hace 6000 años), del último periodo glacial máximo (aproximadamente hace 22000 años) y del último interglaciar (periodo alrededor de hace 120000 y 140000 años). Por último, cuenta con proyecciones de datos futuros para los años 2050 y 2070, esto para cuatro posibles escenarios respecto a la concentración de gases de efecto invernadero en la atmósfera. Aunque para nuestro objetivo sólo son necesarias las condiciones actuales, la existencia de los datos restantes ofrece la posibilidad de hacer un análisis respecto a cambio climático o de comparación de ambientes disponibles en las épocas utilizables.

Todos los datos se encuentran disponibles en cuatro resoluciones distintas. La más burda es de 10 arco-minutos, que equivale a aproximadamente 18 km mientras que la más fina es de 30 arco-segundos, correspondiente a aproximadamente 0.8 km. Estas consideraciones son importantes

a la hora de hacer el análisis con R pues como se mencionó en los capítulos anteriores, las librerías existentes soportan un número reducido de datos si no se cuenta con mucha memoria. Es por esto que fue necesario hacer uso de la resolución de 10 arco-minutos y además hacer un muestreo uniforme sobre la rejilla. Esto lo explicaremos con un poco de mayor detalle más adelante. Para mayor información sobre los métodos utilizados para generar las capas de clima, así como la descripción de algunas de las variables ambientales que se consideran y las unidades en las que están medidas, referimos al lector a Hijmans et al. (2005). En esta referencia se describe la forma de obtener las capas respecto a ciertas variables como precipitación media mensual, temperatura mínima y máxima, entre otras.

Esta base de datos se encuentra en formato *raster*, que en computación es el término para denominar una estructura de matriz de puntos que generalmente representa una rejilla rectangular de pixeles. Para manipularlos, en R se encuentra disponible el paquete "raster", el cual provee un conjunto de funciones y clases que permiten manejar datos geográficos en este formato. En conjunto con los paquetes pHom y TDA son los que fueron utilizados para hacer el análisis topológico.

Como se mencionó en el Capítulo 1, para poder hacer uso de la homología persistente es necesario que los datos se encuentren representados en un espacio métrico adecuado. En nuestro caso, vamos a describir a qué nos vamos a referir con el término "adecuado". Rencher (2002) nos describe al respecto: en un contexto univariado, la distancia entre dos números x y y es medida por el valor absoluto de su diferencia $|x - y|$. En un sentido estadístico, esta distancia no es muy informativa; por ejemplo, no nos interesa saber cuántas *unidades* difieren dos medias, sino cuántas *desviaciones estándar* difieren entre sí. Es por esto que consideramos estadísticas de estudio como

$$\frac{|\mu_1 - \mu_2|}{\sigma}.$$

En el caso multivariado, para obtener una distancia útil debemos tomar en consideración no sólo las varianzas de las variables sino también las covarianzas o correlaciones entre ellas. Mahalanobis (1936) introduce la distancia

Definición 3.2.1. *Dada una colección de variables aleatorias $X_1, X_2, ..., X_n$ con matriz de varianzas y covarianzas Σ invertible y dados dos vectores aleatorios $x = (x_1, ..., x_n)^T$ y $y = (y_1, ..., y_n)^T$*

con distribución $x, y \sim (X_1, X_2, ..., X_n)$*, se define la distancia de Mahalanobis entre* x *y* y *como*

$$d_M(x,y) = \sqrt{(x-y)^T \Sigma^{-1}(x-y)}.$$

Notemos que

$$\begin{aligned} d_M(x,y) &= \sqrt{(x-y)^T \Sigma^{-1}(x-y)} \\ &= \sqrt{(x-y)^T (\Sigma^{1/2}\Sigma^{1/2})^{-1}(x-y)} \\ &= \sqrt{\left[(\Sigma^{1/2})^{-1}(x-y)\right]^T \left[(\Sigma^{1/2})^{-1}(x-y)\right]} \\ &= \sqrt{z^T z}, \end{aligned}$$

en donde $z = \left[(\Sigma^{1/2})^{-1}(x-y)\right]$. En particular si $\Sigma = I$, la distancia de Mahalanobis resulta ser la distancia euclidiana. La importancia en el uso de esta distancia sale a relucir cuando tenemos variables con distinto rango de valores; por ejemplo, si tenemos dos variables X_1 y X_2 que miden la temperatura media en unidades $°C$x10 (10 veces los grados centígrados) y la precipitación media anual medida en mm^3 respectivamente. Por un lado, X_1 tiene un rango de valores relativamente pequeño, digamos $X_1 \in [130, 250]$ que representan temperaturas de 13 a 25 $°C$; en cambio, los valores que toma la precipitación media anual pueden rondar desde los 800 a los 4000 mm^3, dependiendo de la región. De aquí que las variables X_1 y X_2 tienen distinta varianza, y además es de esperarse que su correlación no sea nula. En el sentido geométrico, las bolas $B_r(x)$ respecto a esta distancia resultan ser elipsoides con centro en x como se muestra en la Figura 27. Las variables que se utilizaron para el análisis presentan rangos muy distintos entre ellas, como se ilustra en la Figura 26. Es por esto que es necesario hacer uso de la distancia de Mahalanobis para dotar a los datos de una estructura de espacio métrico.

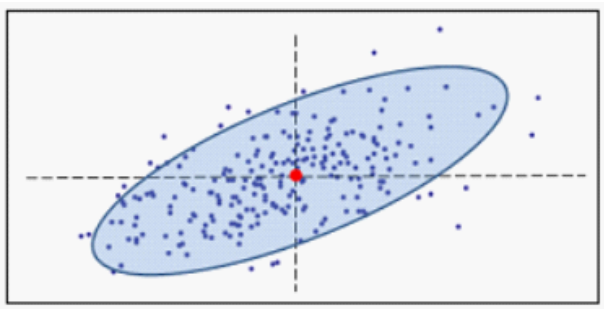

Figura 27. Ilustración de las vecindades en $\mathbb{R}^2$ con la distancia de Mahalanobis.

Ya que hemos descrito una forma conveniente de colocar nuestros datos en un espacio métrico, nos falta definir filtraciones para poder implementar el algoritmo de homología persistente. Para el estudio de nichos ecológicos y ambientes disponibles, salen a relucir tres filtraciones para el espacio con una interpretación muy clara en el sentido biológico.

Filtración por franjas en G. Comenzaremos dando un ejemplo para mostrar la intuición detrás de la filtración. Sea G el área geográfica que conforma América. Supongamos que una especie atraviesa el continente americano en un movimiento de migración, iniciando por la parte del estrecho de Bering, continuando hacia el sur por América del Norte, pasando Centroamérica y terminando en la parte más baja de Argentina. A lo largo del intervalo de tiempo que le tome a la especie hacer esto $[t, T]$, podemos definir una sucesión creciente de subconjuntos de G, con una partición del intervalo $t = t_1 < t_2 < \cdots < t_n = T$, definimos

$$\begin{aligned} G_j &= \{x \in G : x \text{ tiene una latitud mayor o igual al punto con menor} \\ &\qquad \text{latitud recorrido por la especie al tiempo } t_j\}. \end{aligned}$$

Tenemos que $G_1 \subset G_2 \subset \cdots \subset G_n$, lo cual se observa mejor en la Figura 28. Esta filtración del espacio geográfico G nos inducirá una filtración del espacio ambiental disponible de G; en efecto, definiendo

$$E_j = f(G_j),$$

tenemos que $E_1 \subset E_2 \subset \cdots \subset E_n = E$. Esta filtración tiene un sentido muy claro en el ámbito de dinámica de poblaciones, pues en este caso la filtración está directamente relacionada con el movimiento de migración de una especie o grupo de especies. El parámetro de "tiempo" en la filtración tiene que ver con la profundidad de la migración en dirección sur. Lamentablemente, en el sentido computacional esta filtración no es sencilla de implementar debido a que los subconjuntos crecientes se construyen de manera no parametrizada.

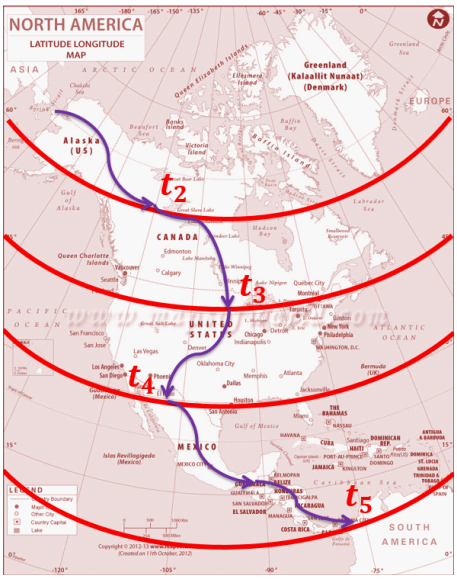

Figura 28. Ilustración de la filtración por franjas en el espacio geográfico, basada en un movimiento de migración comenzando por el Estrecho de Bering.

Filtración por bolas en G. Esta filtración es muy parecida a la descrita anteriormente pero en una escala local. Supongamos que un individuo de cierta especie de árbol se encuentra en una montaña. Al esparcir su semilla, a lo largo del tiempo el árbol logrará conquistar territorio en la montaña e incluso fuera de ella. Definamos

$$G_j = \{x \in G : x \text{ tiene una distancia al árbol menor o igual a la semilla más alejada al tiempo } t_j\},$$

es decir, las G_j son vecindades alrededor del árbol como se puede apreciar en la Figura 29. Tenemos que esta filtración al estar relacionada con el esparcimiento de la semilla de un árbol, tiene una interpretación directa en el sentido de dispersión de individuos en el espacio geográfico. Por desgracia es igual de complicada de computar que la filtración anterior. Si no es posible parametrizarla de alguna manera, resulta computacionalmente muy complicada de implementar. A continuación veremos que la filtración de Vietoris-Rips tiene una explicación en el sentido de lo que vamos a conocer como *tolerancia ambiental*.

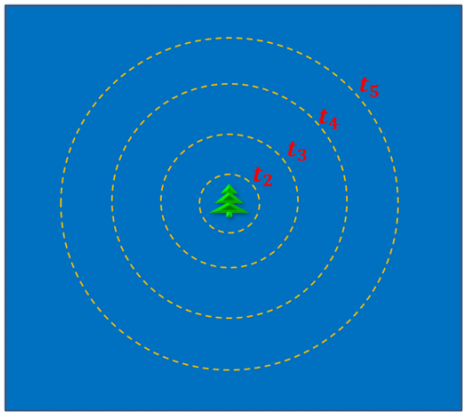

Figura 29. Ilustración de la filtración por vecindades de un individuo de alguna especie de árbol.

Filtración por bolas en N. Las dos filtraciones anteriores representan el recorrido del espacio ambiental inducido por movimiento (de migración o de dispersión) en el espacio geográfico. Ahora consideremos una filtración respecto a una distancia dentro del espacio ambiental, aplicando directamente la filtración de Vietoris-Rips al conjunto $S_n \subset N$. Dicha filtración representa una exploración del subespacio que representa el nicho N directamente en el espacio ambiental. Es importante mencionar que un individuo de cierta especie toma en cuenta principalmente las características ambientales que tiene un lugar y que éstas le permitan subsistir, dejando en segundo plano las características correspondientes al espacio geográfico. Con base en esto definimos el término tolerancia ambiental como la capacidad que tiene la especie para cambiar de ambiente. Para ilustrar a lo que nos referimos, observemos en la Figura 30 que si una especie está situada en el punto x entonces tiene menor tolerancia a cambiar de ambiente que si estuviera en el punto y.

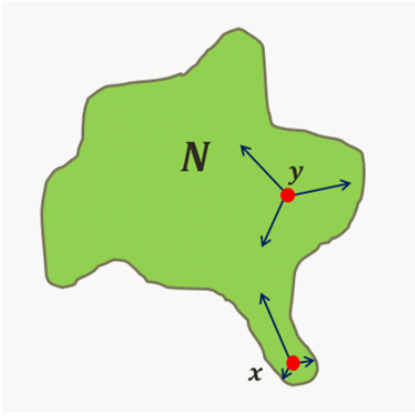

Figura 30. Interpretación de la filtración de Vietoris-Rips en el espacio ambiental, directamente

aplicada a una muestra en el nicho ecológico. Ésta está directamente ligada a la tolerancia ambiental que presenta la especie.

El término tolerancia ambiental se encuentra directamente ligado con presencia de una especie en una región geográfica. Si una especie es especialista y su hábitat está ubicado principalmente en zonas tropicales, dicha especie posiblemente no podrá resistir un cambio a ambientes menos húmedos y fríos. Asimismo, si una región geográfica tiene un ambiente disponible que presenta una estructura muy accidentada es de esperarse que tenga menor diversidad de especies que una que no. Debido a esto, esta última filtración se utilizó para llevar a cabo el análisis que describiremos en la siguiente sección.

3.3. Casos de estudio

3.3.1. Nichos ecológicos

Como mencionábamos anteriormente, surge la idea de estudiar al nicho ecológico de una especie como subespacio de $\mathbb{R}^d$ y ver si existe relación entre su complejidad topológica y la clasificación entre especialista y generalista. En términos de homología se busca observar diferencias en el número de agujeros presentes en los nichos. En el contexto de nichos ecológicos, actualmente los modelos se aplican principalmente para inferir el conjunto como tal, de él obtener la distribución potencial de la especie y con base en esta información generar conclusiones que sustenten la creación de políticas de salud, económicas, de conservación, entre otras. El objetivo de aplicar la homología persistente en este contexto es evaluar si este enfoque produce información relevante en el sentido de tolerancia ambiental.

Para el análisis se utilizaron los nichos de siete especies situadas en la región ilustrada en la Figura 31, la cual comprende la parte sureste de México, todo América Central y la parte norte de América del Sur. Cuatro especies generalistas: mono araña (*Ateles*), zorra (*Urocyon*), pantera (*Panthera onca*) y mono aullador (*Alouatta palliata*), ver Figura 32; y tres especialistas: musaraña (*Cryptotis*), murciélago (*Balantiopteryx io*) y mono aullador (*Alouatta pigra*), ver Figura 33. A continuación daremos una pequeña reseña de cada una de las especies.

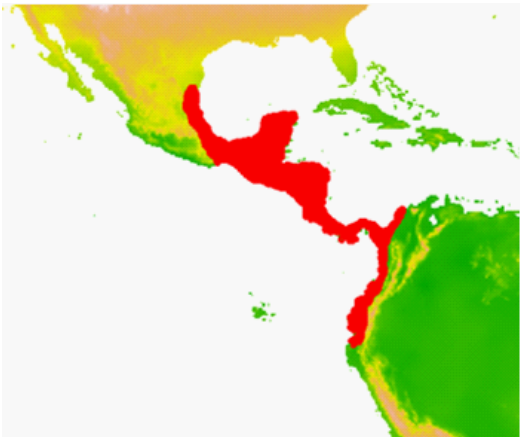

Figura 31. Región en donde se sitúan las siete especies que fueron estudiadas.

Como se menciona en Townsend (2003), una especie generalista es aquella que es capaz de desarrollarse en una amplia variedad de condiciones ambientales y le es posible hacer uso de diferentes recursos como alimento para poder vivir. Con ello presenta mayor capacidad de adaptarse a nuevos lugares. El mono araña se encuentra situado en bosques tropicales de Centroamérica y América del Sur, se caracteriza principalmente por sus largos brazos y extremidades convirtiéndolo en el mono más grande del denominado *Nuevo Mundo*. Esta especie se alimenta principalmente de frutas y con menor medida de hojas, flores e insectos, lo cual es una de las razones por las que es considerado generalista. La zorra es un mamífero carnívoro que se extiende a lo largo del continente americano y habita en zonas boscosas. Es capaz de trepar árboles, lo cual le permite mayor variedad de alimentación, cazando aves y ardillas. La pantera es una variedad negra del jaguar (*Panthera onca*) debido al melanismo. Este felino, si bien prefiere las selvas densas y húmedas, se puede adaptar a áreas boscosas o abiertas. Fundamentalmente es solitario y se alimenta de animales pequeños. Por último, el mono aullador habita en diversidad de ambientes que comprenden bosques (húmedos, semi-caducifolios, secos, de montaña, entre otros) y manglares. Su dieta es a base de frutas y hojas tiernas principalmente. A pesar de encontrarse vulnerable debido a la deforestación, por su tipo de dieta y la capacidad de vivir en espacios muy reducidos es muy adaptable y es capaz de vivir en bosques fragmentados e intervenidos. En el contexto de esta clasificación, se busca explorar si la homología del nicho de una especie generalista presenta alguna particularidad.

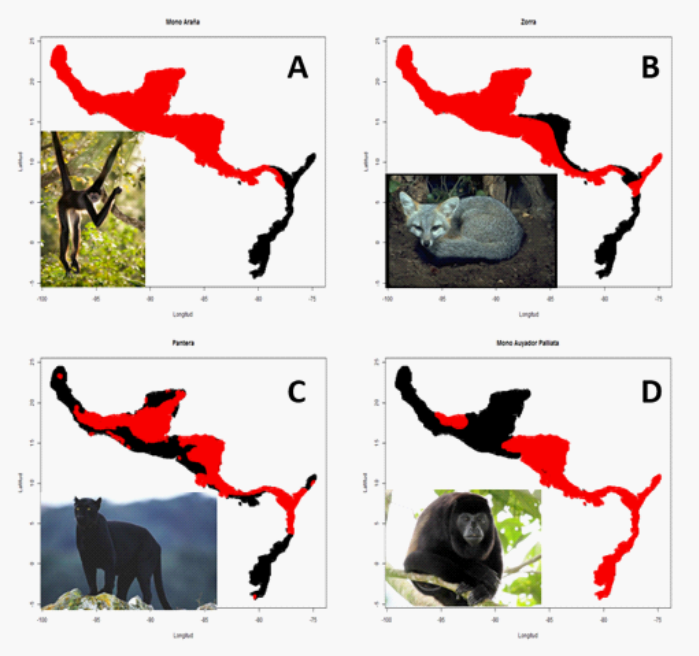

Figura 32. Especies generalistas. Se muestra en rojo la región correspondiente al nicho del mono araña (*Ateles*), de la zorra (*Urocyon*), de la pantera y del mono aullador (*Alouatta palliata*) en A, B, C y D respectivamente.

Al contrario de una especie generalista, una especie especialista tiene un rango restringido de condiciones ambientales que le son favorables o bien tienen una dieta muy específica, tal como se explica en Townsend (2003). De esta definición se entiende que una especie especialista presenta poca tolerancia ambiental. La musaraña es un mamífero de tamaño pequeño, de entre 4 y 15 centímetros, que se caracteriza por sus ojos pequeños y sus bigotes que son muy sensibles. Es muy parecido a un ratón, pero recientemente se ha sugerido que tienen un parentesco cercano con los erizos. Son los mamíferos más activos que existen, cazan de día y de noche y pueden llegar a morir si pasan más de cuatro horas sin comer; esta condición los hace difíciles de adaptarse. El murciélago de saco del sureste es una especie presente en Belice, Guatemala y principalmente en el sureste de México. Este murciélago se alimenta únicamente de insectos. Como se menciona en Chiroptera Specialist Group (1996), a pesar de que esta especie presenta una distribución relativamente amplia, los hábitats de la zona están en una situación de extrema fragilidad por la actividad turística y el vandalismo en las cuevas donde duermen, por lo que la especie está en peligro. Por último, el mono aullador negro guatemalteco es una especie de mono oriundo de

una región comprendida entre Belice, Guatemala y principalmente la península de Yucatán y los estados de Tabasco y Chiapas. Tiene su hábitat en selvas tropicales perennes, semi-caducifolias y de tierra baja. Es endémico de la selva tropical centroamericana y es la única especie de aullador que vive en la península de Yucatán. A continuación enunciaremos las consideraciones que se adoptaron para aplicar TDA en estos nichos, tanto de especies generalistas como de especialistas.

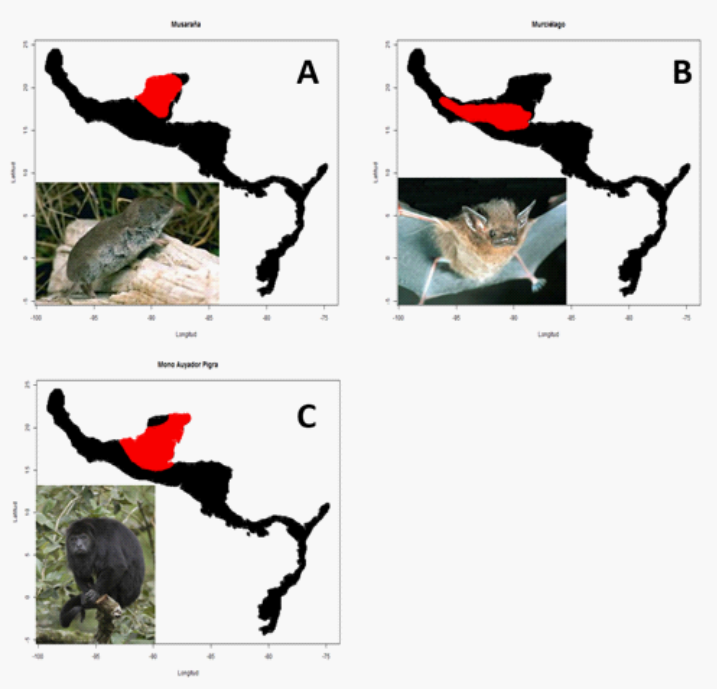

Figura 33. Especies especialistas. Se muestra en rojo la región correspondiente al nicho de la musaraña (*Cryptotis*), del murciélago (*Balantiopteryx io*) y del mono aullador (*Alouatta pigra*) en A, B y C respectivamente.

Los nichos están con base en las variables temperatura media anual (BIO1), el rango de temperatura anual (BIO7), temperatura media del cuarto más seco (BIO9) y la precipitación anual (BIO12). En la Figura 34 se muestran los valores mínimo y máximo de cada variable a nivel mundial. Aunque en cada nicho se presenten rangos de valores más reducidos, estos siguen siendo muy diferentes entre sí. Es por ello que tomamos el nicho como subespacio de $\mathbb{R}^d$ dotado de la distancia de Mahalanobis. Como ya se mencionó, se utilizó el paquete pHom para generar los diagramas de persistencia, basados en la filtración de Vietoris-Rips y se investigó la homología de dimensión 0, 1 y 2. Además, debido a que la base de datos, en su resolución más burda, tenía alrededor de

15000 datos, fue necesario hacer un muestreo uniforme con tamaño 300; R con el paquete pHom (y no con muchos cambios en el paquete TDA) sólo permite una base de datos de alrededor de 800, cuando la homología es sencilla, y este número se reduce a 400 o menos si se tiene presencia de muchos agujeros, ya sean ruido topológico o señales verdaderas. Se hicieron varios muestreos uniformes pero todos fueron consistentes con los resultados que se presentarán a continuación. En la Figura 35 y Figura 36 se muestran los resultados del análisis mediante los diagramas de persistencia obtenidos y dotados de una banda de confianza.

Variable	min	máx
BIO1	-513	338
BIO7	53	725
BIO9	0	65535
BIO12	0	9916

Figura 34. Rangos de las variables que se utilizaron en la construcción de los nichos.

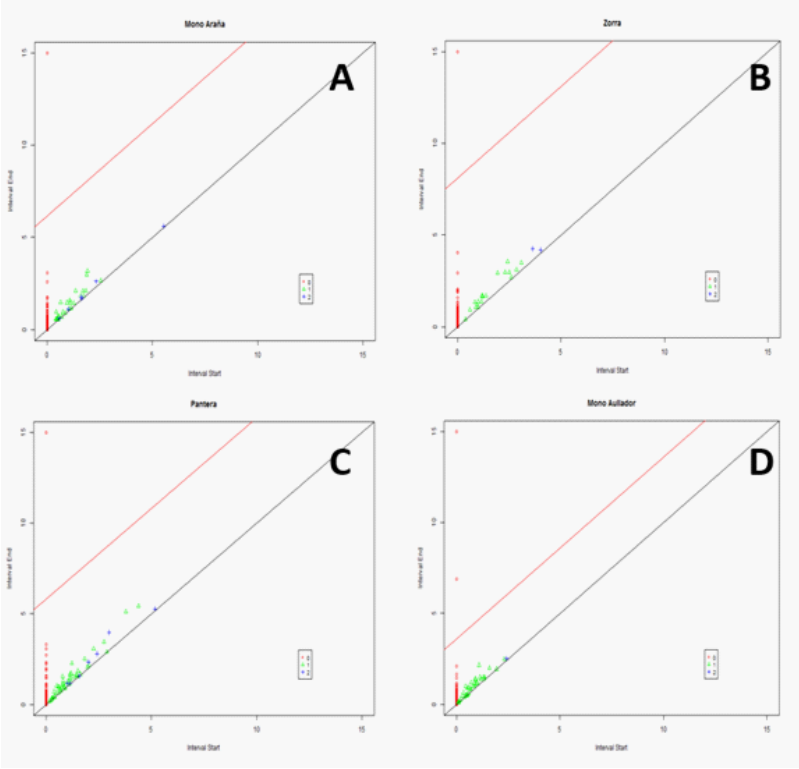

Figura 35. Diagramas de persistencia de las especies generalistas. Los puntos rojos indican las componentes conexas, los triángulos verdes son los agujeros de dimensión 1 y las cruces azules los agujeros de dimensión 2.

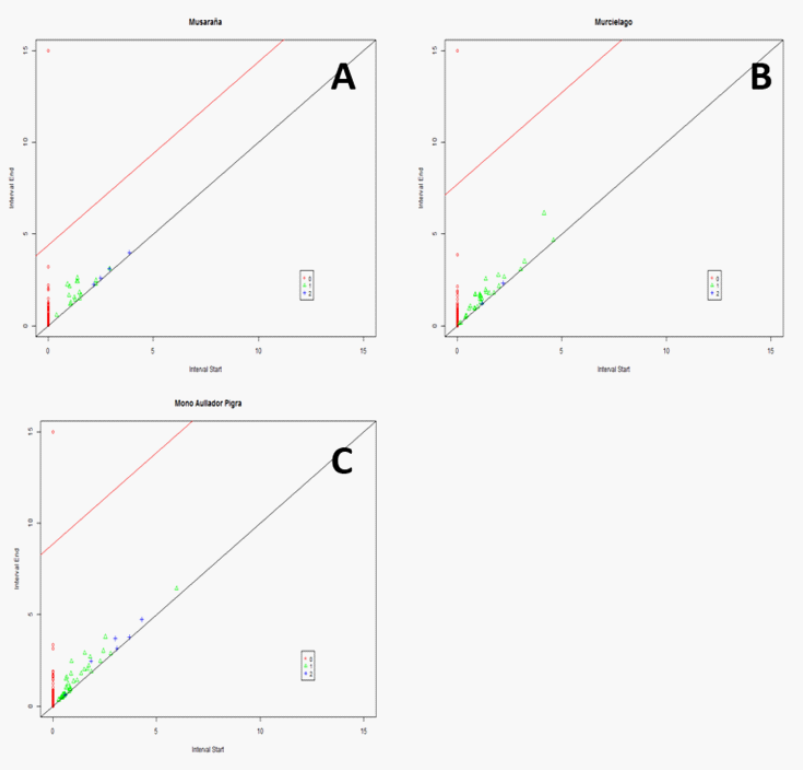

Figura 36. Diagramas de persistencia de las especies especialistas. Los puntos rojos indican las componentes conexas, los triángulos verdes son los agujeros de dimensión 1 y las cruces azules los agujeros de dimensión 2.

De los resultados obtenidos, se puede concluir que no es posible hacer una distinción en la homología de los nichos entre estas especies generalistas y especialistas. A excepción del nicho del mono aullador Palliata, todos los nichos presentan una homología trivial (sin agujeros) en dimensiones 1 y 2, y sólo tienen una componente conexa. Por su parte, el nicho del mono Aullador Palliata (especie generalista) está dividido en dos componentes. Como se puede observar en las Figuras 32 y 33, esta especie abarca regiones con características ambientales que no permite ninguna otra de las especies consideradas. Un hecho importante es que investigaciones recientes sugieren que el nicho fundamental de una especie forma un elipsoide en el espacio de variables ambientales. Esto se debe a que en dichos trabajos, tales como Osorio-Olvera (2016) y Broennimann et al. (2012), presentan evidencia de que la abundancia de una especie decrece conforme aumenta la distancia al centroide del nicho climático de la especie, de donde consideran una normal multivariada para modelar el nicho fundamental de especies virtuales. Esto es consistente con los resultados que obtuvimos. Por

otro lado, es necesario reflexionar que el nicho existencial es un subconjunto del nicho fundamental, y depende directamente del ambiente disponible a nivel global. Como podremos corroborar con los resultados en los ambientes disponibles estudiados, existen regiones en el planeta cuyo ambiente disponible presenta una homología que no es la trivial, al menos en dimensiones 0 y 1.

Existen varias direcciones en las cuales dirigir nuestra atención respecto a este trabajo. Por un lado, es factible considerar estudiar nichos de especies que se encuentren presentes en otras regiones geográficas del planeta y tomar en cuenta nichos de un número más grande de especies. Es interesante para este objetivo, y en cierto sentido necesario, estudiar el ambiente disponible a nivel global para conocer su homología. Sin embargo, para esto es necesario contar con mayor poder de cómputo, debido a que, como ya hemos comentado, los algoritmos disponibles hasta el momento requieren de mucha memoria conforme el conjunto de datos que se analiza contiene muchos elementos. Considerar el ambiente disponible de todo el mundo, a pesar de que se haga un muestreo uniforme, para poder asegurar una muestra representativa, esta muestra debe de tener una cantidad de datos con el orden de miles. Para ilustrar que el ambiente disponible global contiene una homología que no es del todo sencilla, enseguida se mostrará el estudio realizado en el tema de ambientes disponibles.

3.3.2. Ambientes disponibles

El ambiente disponible de una región geográfica juega un papel importante en el sentido de riqueza y distribución de especies dentro del área en cuestión. En Chen (2013) se hace un breve análisis de la relación que hay entre ciertas variables espacio/ambientales y la riqueza de especies; se muestra que dichas variables (que versan entre temperatura, precipitación, elevación, entre otras) son determinantes en el índice de riqueza en el espacio geográfico. Surge la inquietud de estudiar el ambiente disponible como subespacio topológico de $\mathbb{R}^d$ y analizar la posible relación entre su complejidad en el sentido de homología y los conceptos ya mencionados. Por una parte, en un sentido exploratorio en este contexto se realizó el análisis de pequeñas regiones, una en la parte central de Canadá y otra en África central. Además, en un sentido ecológico India y Australia presentan características que sugieren su comparación. A continuación se presentan ambos análisis.

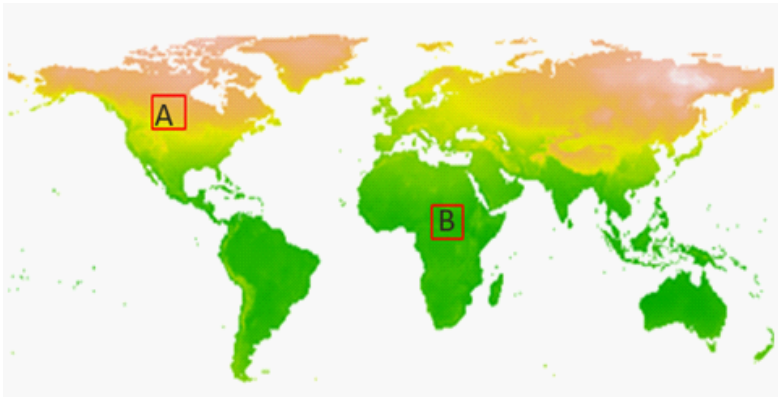

Figura 3.1: Figura 37. Los recuadros marcados son las regiones que fueron analizadas con TDA. La región A corresponde a Canadá y B a la parte central de África.

En un principio, se planteó la idea de estudiar dos regiones que en los mapas que ilustran las variables ambientales parecieran diferentes. Como se puede apreciar en la Figura 37, las regiones marcadas con los recuadros rojos presentan características ambientales distintas. De esto surgió la percepción de compararlas mediante la homología de sus ambientes disponibles. Para el análisis, se hizo uso de la filtración de Vietoris-Rips, que como ya describimos anteriormente, está relacionada con la tolerancia que tiene una especie a cambiar de ambiente dentro de esta región. Las variables utilizadas fueron las mismas que en el análisis de nichos y por consecuencia también se tomó en cuenta la distancia de Mahalanobis. En la Figura 38 se muestran los diagramas de persistencia obtenidos.

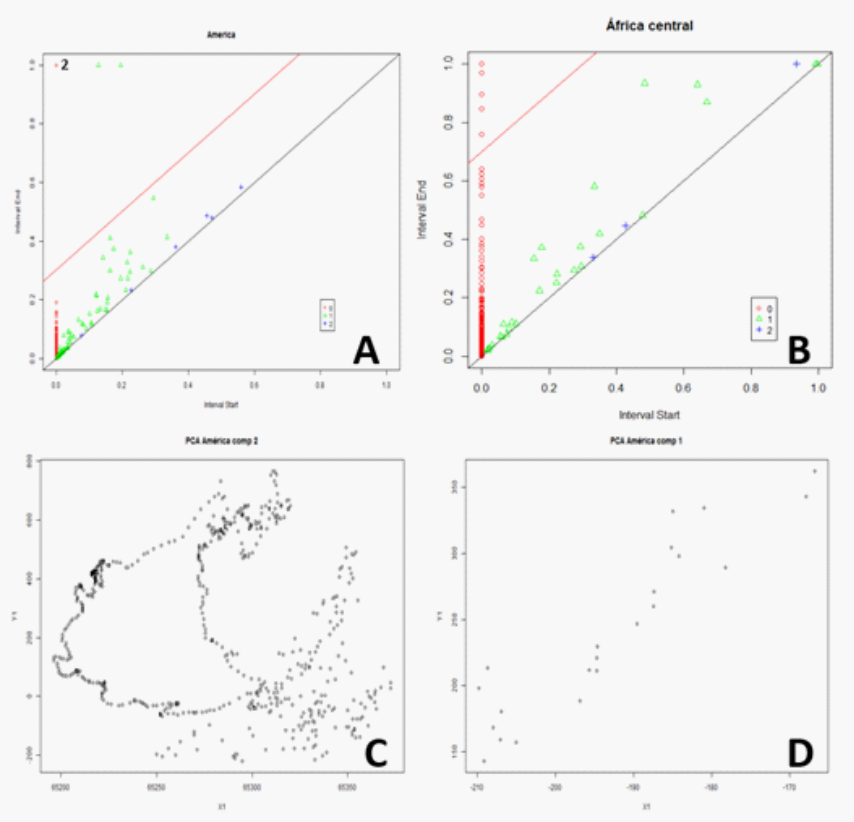

Figura 38. En A y B se muestran los diagramas de persistencia de Canadá y de África Central respectivamente. En el diagrama de persistencia en A, el número 2 indica la multiplicidad de la clase. En C y D se muestran las proyecciones de las dos componentes conexas del ambiente disponible de la región central de Canadá.

De estos diagramas de persistencia, se concluye que los ambientes disponibles presentan una estructura topológica diferente. Por un lado, a pesar de que el ambiente disponible correspondiente a la región en África central está compuesto por varias componentes conexas, éste tiene como grupos de homología de dimensión 1 y 2 al grupo trivial. Un resultado interesante es la homología de dimensión 1 del ambiente disponible correspondiente a la región central de Canadá, debido a que dicho grupo de homología contiene dos clases significativas. Como estudio complementario, se propuso hacer un análisis por componentes principales para intentar visualizar estos agujeros. En la Figura 38, en C y D, se muestran las proyecciones a las componentes principales de los dos pedazos de los que está compuesto (como se observa en la Figura 38 A, hay dos agujeros de dimensión 0 significativos). Cabe destacar que una de las componentes conexas, que se muestra en la Figura 38 D, tiene una cantidad considerablemente pequeña de puntos; analizando los datos respecto

a las variables ambientales, dicha componente corresponde a características presentes en puntos extremos como lo son las partes altas de montañas. De la proyección de la otra componente conexa se puede apreciar claramente los ciclos que detecta la homología persistente. De estos resultados, sale a relucir la inquietud de estudiar otras regiones geográficas desde la homología de su ambiente disponible, y a mayor escala, la homología del ambiente disponible global.

17 de los 190 países que hay en el mundo, en conjunto poseen cerca del 70 % de la biodiversidad global. Cada uno de ellos es considerado como *mega diverso*, debido a que, como es de esperarse, individualmente también contienen una gran cantidad de especies diferentes, tanto animales como vegetales. Australia e India son dos de estos países mega diversos. Por un lado, como mencionan en Chapman (2009), Australia alberga a alrededor de 147,579 especies diferentes; mientras que India, como lo indica la página web oficial http://indiabiodiversity.org/, tiene en su territorio a 91,000 especies animales y 45,500 de plantas. Ambos países comparten un tamaño territorial muy parecido. De acuerdo a la definición de la RAE, un subcontinente es una gran extensión territorial que se considera subdivisión del continente al que pertenece. India y Australia son considerados subcontinentes de Asia y Oceanía respectivamente. Todas estas, son algunas de las razones por las cuales, en un sentido ecológico surge la noción de compararlos. En el contexto del análisis topológico, se planteó la pregunta ¿la homología de sus ambientes disponibles produce información relevante? En la Figura 39 se muestran los resúmenes topológicos de ambos ambientes disponibles.

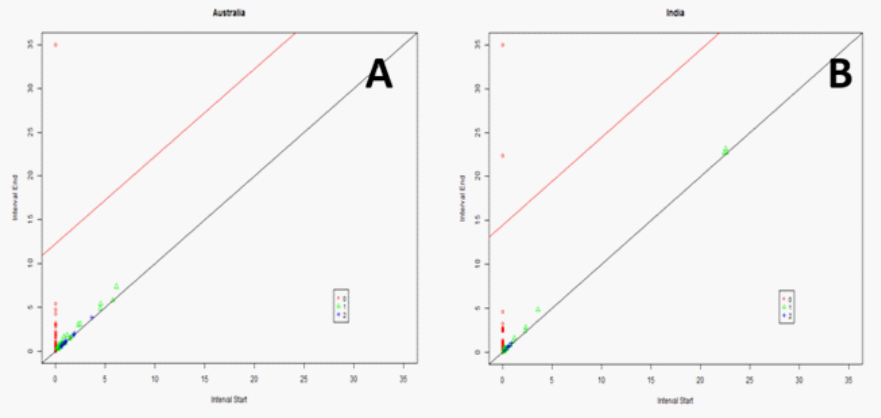

Figura 39. Diagramas de persistencia correspondientes a Australia (A) y a India (B). Los puntos rojos indican las componentes conexas, los triángulos verdes son los agujeros de dimensión 1 y las cruces azules los agujeros de dimensión 2.

Como podemos apreciar en ambos diagramas de persistencia, la homología de los dos ambientes disponibles es relativamente sencilla. Por una parte, el ambiente disponible de Australia presenta una sola componente conexa y no contiene agujeros de dimensiones 1 y 2. De igual manera el ambiente disponible correspondiente a India tiene como grupos de homología de dimensión 1 y 2 al grupo trivial, pero por otro lado, está dividido en dos componentes conexas. De forma parecida a las componentes surgidas en el ambiente disponible de la región central de Canadá, una de las componentes conexas del ambiente disponible de la India surge debido a características que corresponden a regiones extremas, la cual contiene una cantidad muy pequeña de puntos respecto a la otra componente. De estos resultados podemos concluir que ambos ambientes disponibles presentan una estructura topológica muy similar. Esto nos permite aseverar que India y Australia, poseén una estructura semejante de características ambientales, lo cual corrobora que presenten propiedades parecidas en el sentido de riqueza y abundancia de especies.

Del análisis realizado, surge la motivación de estudiar otras regiones geográficas a través de la homología de sus ambientes disponibles. Además, existen varias preguntas que sería interesante contestar. Por un lado, se concluye que en el espacio ambiental a nivel mundial se tiene la presencia de agujeros de dimensión 1, pero en el sentido biológico no es muy claro aún qué significan estos agujeros ni qué repercusiones tienen respecto a la dispersión de especies en esas regiones o respecto a alguna otra característica ecológica del lugar. Por otro lado, es posible accesar a datos ambientales correspondientes a condiciones pasadas y futuras, analizar la homología a nivel mundial del ambiente disponible en estos escenarios y comparar dichos resultados, con la intención de estudiar los posibles efectos del cambio climático en el sentido de homología. También, considerar nuevas filtraciones para estudiar con homología persistente en este contexto es una posible tarea a seguir. Se plantean estos desafíos para su posible tratamiento en trabajos futuros.

Conclusiones

El análisis topológico de datos es una rama emergente dentro del área de topología computacional. La teoría desarrollada inmersa en la topología algebraica, la cual sustenta los algoritmos descritos en la homología persistente, ha surgido apenas en las últimas dos décadas. En contraste la aplicación como tal es mucho más reciente, sólo de algunos años a la fecha. Por ello, la técnica que engloba TDA aún presenta varias limitaciones que versan en el sentido computacional y en la interpretación de resultados.

Por una parte, nos encontramos con la necesidad de mejorar los algoritmos que son utilizados para obtener la homología persistente. Un potencial uso de este enfoque de análisis es la aplicación a grandes conjuntos de datos: Big Data. Sin embargo, por el momento los paquetes desarrollados de software permiten un número limitado de datos, que dependiendo de la complejidad estructural de la correspondiente nube, el trabajo computacional puede llegar a ser imposible de efectuar. Además, dentro del contexto de estudio, frecuentemente surgen filtraciones de manera natural con una interpretación relevante asociada directamente al problema en cuestión. Pero estas filtraciones usualmente no son sencillas de implementar, lo cual reduce el posible aprovechamiento de concebir a los datos dentro de un espacio topológico. Estas son algunas de las razones por las cuales es imprescindible seguir explorando los aspectos computacionales de este enfoque.

Por otro lado, como se mencionó en el Capítulo 1, analizar una nube de datos mediante el análisis topológico es en un principio sencillo. Basta con dotar al conjunto de puntos que conforman la nube con una distancia adecuada dentro del fenómeno que dan lugar a los datos. Pero como acabamos de comentar, surge un problema en el momento de intentar implementar una filtración directamente relacionada con el contexto del análisis. Por su parte, existen filtraciones estándar, como la de Vietoris-Rips, que siempre pueden ser utilizadas para aplicar el enfoque. Sin embargo,

los resultados que proporciona el análisis con estas filtraciones muchas veces no resultan sencillos de interpretar. Con la intención de solucionar esta situación, es esencial una estrecha colaboración entre quien analiza los datos y el dueño de los mismos.

A pesar de las limitaciones presentes en el análisis topológico, TDA posee un fuerte potencial para la exploración inicial de datos complejos. El adjetivo *complejo* se sitúa principalmente en la dificultad asociada para realizar un análisis, dificultad que se refleja en tres propiedades relacionadas directamente con los datos: el tamaño de la nube, el espacio en el que se encuentra y la estructura inherente de la nube en ese espacio. Por un lado, si una base de datos es demasiado grande, muchas de las técnicas usuales de análisis resultan no ser adecuadas para obtener información confiable. Pero por otro lado, tener una muestra pequeña tampoco nos asegura facilidad de análisis. El espacio en donde está inmersa la nube de datos, o bien el subespacio que representa por sí misma, presenta cada vez con mayor frecuencia características complicadas de abordar con la teoría estadística clásica. Ejemplos de estas características son la alta dimensionalidad y la geometría propia del espacio, lo que sugiere la concepción del conjunto de datos como subconjunto de objetos matemáticos más generales como lo son las variedades topológicas. El análisis topológico es una herramienta que provee de información relevante respecto a la estructura de dicho espacio. Debido a esto, TDA representa un instrumento exploratorio para la identificación de la forma que posee la nube de datos, proporcionando un preámbulo respecto a la manera en la que debe ser abordado el análisis.

Del trabajo realizado en nichos ecológicos y en ambientes disponibles fue posible obtener varias conclusiones en diferentes sentidos. Primero, de manera directa surgieron resultados en el sentido ecológico haciendo distinciones entre ambientes disponibles y obteniendo consistencia en nichos, ambos casos respecto a su homología. Por su parte, como requisito indispensable la técnica debe ser articulada con ingredientes estadísticos. Como se detalla en el Capítulo 2, los resúmenes topológicos resultantes, deben ser tratados y estudiados como estadísticas obtenidas a partir de los datos. También, referente directamente al procedimiento del análisis topológico, surge el requerimiento de complementar el análisis con técnicas de remuestreo o de generar filtraciones de manera que sea posible analizar bases de datos con un número considerable de elementos. Estas conclusiones se extienden un poco más a continuación.

En el contexto ecológico, los resultados obtenidos en el estudio de los nichos de las siete especies mostraron que no existe diferencia notoria en su estructura en sentido de homología, esto sin importar su clasificación como especialistas o generalistas. Además, estos resultados también sugieren simplicidad en la estructura de los nichos, concebidos como subconjuntos de un espacio topológico, teniendo una homología sencilla en todos los casos. Por otro lado, del análisis de ambientes disponibles surgen diferentes aseveraciones. En el estudio de los ambientes disponibles de Australia e India, no se encontraron diferencias relevantes entre ambos. Esto es consistente en el sentido de abundancia y riqueza de especies, debido a que los dos países presentan características similares respecto a estos conceptos. Por último, de los resultados obtenidos en Canadá y África, sale a relucir la existencia de agujeros en el ambiente disponible mundial. Se encontró que la homología de una región en Canadá presentaba agujeros de dimensión 1, mientras que el ambiente disponible de una región central de África tiene una homología sencilla. Todo esto induce las posibles líneas a seguir dentro de este contexto, dentro de las que se encuentran el estudio del ambiente disponible global y el análisis de nichos de especies situadas en otras regiones del mundo. Además, se ilustra la caracterización de diferencias mediante aspectos topológicos, con la posibilidad de implementar dicha caracterización en diferentes problemas surgidos en ecología.

La importancia de abastecer el análisis topológico con elementos estadísticos, sale a relucir con el hecho de que intrínsecamente lo que se está resolviendo es un problema de inferencia. Lo que se busca es inferir a partir de una nube de datos, la homología del subespacio topológico del cual se supone está muestreada. De esto sale a relucir la necesidad de proporcionar herramientas estadísticas para obtener inferencias más completas y confiables, como por ejemplo, los conjuntos de confianza para las estadísticas obtenidas, que en específico en este trabajo fueron los diagramas de persistencia, pero también es necesario para otras opciones como lo son los *panoramas de persistencia* y los códigos de barras. Así también, es necesario desarrollar, o bien extender, pruebas de hipótesis para este tipo de inferencias, las cuales dependerán de la naturaleza de los espacios de las estadísticas mencionadas. Debido a la complejidad del espacio de diagramas de persistencia, en el sentido geométrico, es complicado obtener resultados de interés en el ámbito estadístico, lo cual ilustra el requerimiento de considerar diferentes alternativas.

A pesar de que en los últimos años se ha trabajado con esta inquietud estadística, aún queda

mucho por hacer. Dentro de los trabajos ya realizados, es necesario enriquecer los procedimientos construidos, esto debido a que, como se ejemplifica con algunos de los métodos en el Capítulo 2, presentan características computacionales que no son favorables en el sentido de que son complicados de implementar. Es por esto, que se sugiere y se incita a continuar con el camino que conlleva el desarrollo del área del análisis topológico de datos, poniendo en conjunto el conocimiento topológico, estadístico y computacional.

Bibliografía

[1] Adams, H. y Carlsson, G. (2014), Evasion paths in mobile sensor Networks, *International Journal of Robotics Research*, 34, 90-104.

[2] Benedict, M. Q., Levine, R. S., Hawley, W. A. y Lounibos, L. P. (2007), Spread of the tiguer: Global risk of invasion by the mosquito Aedes albopictus, *Vector-Borne and Zoonotic Diseases*, 7, 76-85.

[3] Boissonnat, J. D., y Teillaud, M. (2004), *Effective Computational Geometry for Curves and Surfaces*, Springer.

[4] Broennimann, O., Fitzpatrick, M., Pearman, P. B., Petitpierre, B., Pellissier, L., Yoccoz, N. G., Thuiller, W., Fortin, M. J., Randin, C., Zimmermann, N., Graham, C. H. y Guisan, A. (2011), Measuring ecological niche overlap from occurrence and spatial environmental data, *Global Ecology and Biogeography*, 21, no. 4, 481-497.

[5] Bubenik, P. (2015), Statistical topological data analysis using persistence landscapes, *Journal of Machine Learning Research*, 16, 77-102.

[6] Bubenik, P. y Kim, P. T. (2006), A statistical approach to persistent homology, *Homology, Homotopy and Applications*, 9, 337-362

[7] Chan, J. M., Carlsson, G. y Rabadan, R. (2013), Topology in viral evolution, *Proceedings of the National Academy of Sciences*, 110, 18566-18571.

[8] Chapman, A. D. (2009), Number of Living Species in Australia and the World, *Report for the Australian Biological Resources Study.*

[9] Chazal, F., Fasy, B. T., Lecci, F., Rinaldo, A. y Singh, A. (2013a), On the bootstrap for Persistence diagrams and landscapes, *arXiv:1311.0376*.

[10] Chazal, F., Glisse, M., Labrure, C. y Michel, B. (2013b), Optimal rates of convergence for persistence diagrams in topological data analysis. *arXiv preprint 1305.6239*

[11] Chazal, F., de Silva, V., Glisse, M. y Oudot, S. (2016), *The structure and Stability of Persistece Modules*, Springer.

[12] Chen, Y. H. (2013), Global environment- and space-richness ranking relationships: The effects of interaction and high-order terms of explanatory variables, *Open Journal of Ecology*, 3, 389-394

[13] Chiroptera Specialist Group (1996), Balantiopteryx io, *IUCN Red List of Threatened Species*.

[14] Cohen-Steinner, D., Edelsbrunner, H. y Harer, J. (2007), Stability of persistence diagrams, *Discrete & Computational Geometry*, 37, 103-120.

[15] Cohen-Steinner, D., Edelsbrunner, H., Harer, J. y Mileyko, Y. (2010), Lipschitz functions have l_p-stable persistence, *Foundations of Computational Mathematics*, 10, 127-139

[16] Cuevas, A. (2009), Set estimation: Another bridge between statistics and geometry, *Boletín de Estadística e Investigación Operativa*, 25, 71-85

[17] Cuevas, A. y Fraiman, R. (1998), On visual distances in density estimation: the Hausdorff choice, *Statistics & probability letters*, 40, 333-341

[18] Devroye, L. y Wise, G. L. (1980), Detection of abnormal behaviour via nonparametric estimation of support, *SIAM Journal on Applied Mathematics*, 38, 480-488.

[19] Efron, B. (1979), Bootstrap methods: another look at the Jackknife, *Annals of Statistics*, 7, vol. 1, 1-26.

[20] Elton, C. S. (1927), Animal ecology, *Sidgwick and Jackson*, London.

[21] Espinoza, M. E. (2015), *Notas en Homología persistente*, CIMAT.

[22] Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S. y Singh, A. (2014), Confidence sets for persistence diagrams, *Annals of Statistics*, 42, 2301-2339

[23] Ghrist, R. (2008), Barcodes: The persistent topology of data, *Bulletin-American Mathematical Society*, 45, 61.

[24] Giné, E. y Guillou, A. (2002), Rates of strong uniform consistency for multivariate kernel density estimators, *In Annales de l'Institut Henri Poincare (B) Probability ans Statistics*, vol. 38. ELsevier, 907-921

[25] Good, P. (1993), *Permutation Test: A practical Guide to Resampling Methods for Testing Hypotheses*, Springer-Verlag.

[26] Grinell, J. (1917), The niche-relationships of the California thrasher, *Auk*, 34, 427-433.

[27] Guisan, A., Broennimann, O., Engler, R., Vust, M., Yoccoz, N. G., Lehman, A. y Zimmermann, N. E. (2006), Using niche-based models to improve the sampling os rare species, *Conservation Biology*, 20, 501-511.

[28] Guisan, A., y Zimmermann, N. E. (2000), Predictive habitat distribution models in ecology, *Ecological Modelling*, 135, 147-186.

[29] Harer, J. y Edelsbrunner, H. (2010), *Computational Topology: An Introduction*, American Mathematical Society.

[30] Hatcher, A. (2002), *Algebraic Topolog,* Cambridge University Press.

[31] Hijmans, R. J., Cameron, S. E., Parra, J.L., Jones, P. G. y Jarvis, A. (2005), Very high resolution interpolated climate surfaces for global land areas, *International Journal of Climatology*, 25, 1965-1978.

[32] Hutchinson, G. E. (1957), Concluding remarks*, Cold spring Harbor Symmposia in quantitative biology*, 22, 415-427.

[33] Mahalanobis, P. C. (1936), On generalized distance in statistics, *Proceedings of the National Institute of Science of India*, 12, 49-55.

[34] Mileyko, Y., Mukherjee, S. y Harer, J. (2011), Probability measures on the space of persistence diagrams, *Inverse Problems*, 27, 124007.

[35] Martínez-Meyer, E., Peterson, A. T., Servín, J. I. y Kiff, L. F. (2006), Ecological niche modelling and prioritizing areas for species reintroductions, *Oryx*, 40, 411-418.

[36] Munkres, J. R. (1984), Elements of algebraic topology, volume 2, Addison-Wesley Reading.

[37] Nicolau, M, Levine, A. J. y Carlsson, G. (2011), 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*, 108, 7265–7270.

[38] Ortega-Huerta, M. A. y Peterson, A. T. (2008), Modeling ecological niches and predicting geographic distributions: a test of six presence only methods, *Revista Mexicana de Biodiversidad*, 79, vol. 1, 205-216.

[39] Osorio-Olvera, L. A., Falconi, M. y Soberón, J. (2016), Sobre la relación entre idoneidad del hábitat y la abundancia poblacionalbajo diferentes escenarios de dispersión, *Revista Mexicana de Biodiversidad*, http://dx.doi.org/10.1016/j.rmb.2016.07.001.

[40] Peterson, A. T., Martínez-Campos, C., Nakazawa, Y. y Martínez-Meyer (2005), Time-specific ecological niche modeling predict spatial dynamics of vector insects and human dengue cases, *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 75, 9-15.

[41] Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M. y Bastos, M. B. (2011), *Ecological Niches and Geographic Distributions*, Princeton University Press.

[42] Phillips, S. J., Anderson, R. P. y Schapire, R. E. (2006), Maximum entropy modeling of species geographic distributions, Ecological Modelling, 190, 231-259.

[43] Phipson, B. y Smyth, G. K. (2010), Permutation $p-$values should never be zero: calculating exact p-values when permutations are randomly drawn, *Statistical Applications in Genetics and Molecular biology*, 9, vol. 1.

[44] Pimentel, D., Zuñiga, R. y Morrison, D. (2004), Update on the enviromental and economic costs associated with alien-invasive species in the United States, *Ecological economics*, 52, 273-288.

[45] Politis, D. N., Romano, J. P. y Wolf, M. (1999), *Subsampling*, Springer.

[46] Raxworthy, C. J., Martínez-Meyer, E., Horning, N., Nussbaum, R. A., Schneider, G. E., Ortega-Huerta, M. A. y Peterson A. T. (2003), Predicting ditributions of known and unknown reptile species in Madagascar, *Nature*, 426, 837-841.

[47] Robinson, A. y Turner, K. (2013), Hypothesis Testing for Topological Data Analysis, *arXiv:1310.7467v2*.

[48] Rojas-Soto, O. R., Martínez-Meyer, E., Navarro-Sigüenza, G., Oliveras de Ita, A., Gómez de Silva, H. y Peterson, A. T. (2008), Modeling distributions of disjunct populations of the Sierra Madre Sparrow, *Journal of Field Ornithology*, 79, 245-253.

[49] Romano, J. P. y Shaikh, A. M. (2012), On the uniform asymptotic validity of subsampling and the bootstrap, *Annals of Statistics*, 2765-3175.

[50] Seber, G. A. F. (1984), Multivariate observations, J. Wiley & Sons.

[51] Niyogi, P., Smale, S. y Weinberger, S. (2008), Finding the homology of submanifolds with high confidence from random samples, *Discrete and Computational Geometry*, 39, 419-441.

[52] Rencher, A. C. (2002), *Methods of Multivariate Analysis*, J. Wiley & Sons.

[53] Townsend, C., Begon, M. y Harper, J. (2003), *Essentials of Ecology (2nd edition),* p.54-55 Blackwell.

[54] Turner, K. (2013), Means and medians of sets of persistence diagrams, *arXiv:1307.8300*.

[55] Turner, K., Mileyko, Y., Mukherjee, S. y Harer, J. (2012), Frechet means for distributions of persistence diagrams. *arXiv preprint 1206.2790*.

[56] Wasserman, L. (2004), *All of Statistics, A concise course in statistical inference*, Springer.

[57] Welsh, A. H. (1996), *Aspects of statistical inference*, J. Wiley & Sons.