



ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ciencias Naturales y Matemáticas

Informe de homicidios en la ciudad de Guayaquil, basado en patrones geoespaciales
mediante análisis topológico durante 2023 y 2024.

PROYECTO INTEGRADOR

Previo a la obtención del Título de:

Matemático

Presentado por:

Jean Pierre Vera Morocho

José Antonio Zambrano Campaña

GUAYAQUIL - ECUADOR

Año: 2025

DEDICATORIA

Dedicado a mis padres, a mi hermano, a mi novia y a todos quienes me acompañaron incluso cuando quise desistir.

Jean Vera M.

DEDICATORIA

Dedicado a mi madre, mi hermano, mi familia, Cásper, Cookie, Caramelo, Luna, Félix, Luigi, Kenneth, Frank y todas las personas con las que he tenido la suerte compartir un fragmento de mi vida, que de una u otra forma aportaron en algo para llegar a este momento.

José Zambrano C.

AGRADECIMIENTOS

Extiendo mi más sincero agradecimiento a los profesores y autoridades de la Facultad de Ciencias Naturales y Matemáticas de la ESPOL por hacer posible un espacio de aprendizaje riguroso. De manera especial, quiero expresar mi gratitud a Cristian Hernandez, M.Sc., por su desinteresado apoyo durante la realización de este proyecto, a Jorge Vielma, PhD. por inspirar el tema del mismo y a Elimar Marchan, PhD., por su continua guía dentro y fuera de las aulas durante el tiempo que cursé la carrera.

Jean Vera M.

AGRADECIMIENTOS

Quiero agradecer a mi familia y a mis grandes por ser un soporte fundamental en mi vida. Agradezco los consejos aportados por mis amigos, compañeros y mis profesores, especialmente al profesor Jorge Vielma, quien encendió la chispa para llevar a cabo este proyecto, y a los profesores Fernando Mejías y Ángel Guale quienes nos compartieron su forma de ver el problema. Así mismo agradezco las indicaciones y la experiencia de la profesora Elimar Marchan y la profesora Liliana Pérez por su brindarnos su conocimiento como investigadoras en este y otros proyectos en los que he tenido la oportunidad de colaborar con a ellas; y al profesor Cristian Hernández por su crítica y su guía que nos ha ayudado a mejorar este proyecto.

José Zambrano C.

Declaración Expresa

Nosotros José Antonio Zambrano Campaña y Jean Pierre Vera Morocho acordamos y reconocemos que:

La titularidad de los derechos patrimoniales de autor (derechos de autor) del proyecto de graduación corresponderá a los autores, sin perjuicio de lo cual la ESPOL recibe en este acto una licencia gratuita de plazo indefinido para el uso no comercial y comercial de la obra con facultad de sublicenciar, incluyendo la autorización para su divulgación, así como para la creación y uso de obras derivadas. En el caso de usos comerciales se respetará el porcentaje de participación en beneficios que corresponda a favor del autor o autores.

La titularidad total y exclusiva sobre los derechos patrimoniales de patente de invención, modelo de utilidad, diseño industrial, secreto industrial, software o información no divulgada que corresponda o pueda corresponder respecto de cualquier investigación, desarrollo tecnológico o invención realizada por nosotros durante el desarrollo del proyecto de graduación, pertenecerán de forma total, exclusiva e indivisible a la ESPOL, sin perjuicio del porcentaje que nos corresponda de los beneficios económicos que la ESPOL reciba por la explotación de nuestra innovación, de ser el caso.

En los casos donde la Oficina de Transferencia de Resultados de Investigación (OTRI) de la ESPOL comunique a los autores que existe una innovación potencialmente patentable sobre los resultados del proyecto de graduación, no se realizará publicación o divulgación alguna, sin la autorización expresa y previa de la ESPOL.

Guayaquil, 29 de mayo del 2025.



Autor 1



Autor 2

EVALUADORES

Luz Elimar Marchan Mendoza

PROFESOR DE LA MATERIA

Cristhian Xavier Hernández Rodríguez

TUTOR

RESUMEN

La proliferación de asesinatos y otros problemas de seguridad interna cada vez más han conmocionado al país en años recientes al punto de que se volvió imperativo plantear enfoques de estudios que ayuden a la comprensión de este fenómeno. Este trabajo se propone diseñar un marco analítico basado en el Análisis Topológico de Datos (TDA) que identifique patrones persistentes en datos delictivos de Guayaquil en pro de optimizar la toma de decisiones en seguridad pública. Este marco se aplicó en datos extraídos de la base de datos del Gobierno Nacional a través de una depuración inicial de los mismos, un muestreo aleatorio y la selección de los atributos a estudiar. Se dió inicio al algoritmo de TDA con esos datos, y los resultados incluyeron: diagramas de persistencias y diagramas de códigos de barra con lo cual se obtuvo indicios de ciertas estructuras topológicas del conjunto de datos.

Palabras Clave: Homicidios, Análisis Topológico de Datos, Persistencia homológica, Propiedades topológicas.

ABSTRACT

The proliferation of killings and other domestic security issues have continuously shocked the country in recent years up to the point setting up study lines which help this phenomena's understanding have become imperative. This work's main goal is to design an analytic framework based in Topological Data Analysis (TDA) which pinpoints persistent patterns in Guayaquil's criminal data in order to improve public security's decision making. This framework was applied to data extracted from National Government's database through an initial refinement, a sampling and an election of the characteristics to study. The TDA algorithm was initialized with such data and its results include persistence diagram and barcode diagrams with which some traces of certain topological structures of the dataset were obtained.

Keywords: *Killings, Topological Data Analysis, Homologic Persistence, Topological properties.*

ÍNDICE GENERAL

RESUMEN	I
ABSTRACT	II
ÍNDICE DE FIGURAS	V
ÍNDICE DE TABLAS	VI
CAPÍTULO 1	1
1. INTRODUCCIÓN	1
1.1 Descripción del problema	1
1.2 Justificación del problema	3
1.3 Objetivos	4
1.3.1 Objetivo General	4
1.3.2 Objetivos Específicos	4
1.4 Marco teórico	4
1.4.1 Antecedentes	4
1.4.2 Conceptos claves	5
1.5 Filtraciones	14
1.5.1 ¿Cómo se construyen filtraciones?	14
1.6 Diagramas de Persistencia	15
1.6.1 Propiedades del TDA	16
1.7 Otros métodos para la obtención de patrones en la data.	17

1.7.1 Agrupamiento mediante k-means.	17
CAPÍTULO 2	18
2. METODOLOGÍA	18
2.1 Formulación de alternativas	18
2.2 Diseño del modelo utilizado	19
2.3 Persistencia Homológica.	20
2.4 Visualización de la persistencia Homológica.	21
2.5 Especificaciones técnicas	24
CAPÍTULO 3	28
3. RESULTADOS Y ANÁLISIS	28
3.1 Homologías y contextualización de los datos	28
3.2 Discusión de resultados	30
3.2.1 Resultados globales	30
3.2.2 Resultados locales	32
CAPÍTULO 4	37
4. CONCLUSIONES Y RECOMENDACIONES	37
BIBLIOGRAFÍA	
APÉNDICES	

ÍNDICE DE FIGURAS

Figura 1.1	Casco convexo de cuatro puntos en $\mathbb{R}^2$. (Elaboración propia)	9
Figura 1.2	Representación grafica de un complejo simplicial formado por triángulos. (Elaboración propia)	13
Figura 2.1	Representación de una serie de una serie de grafos que representan grupos de homología, donde aparecen distintas propiedades topológicas.	22
Figura 2.2	Diagrama de persistencia.(Elaboración propia)	23
Figura 2.3	Diagrama de código de barras. (Elaboración propia)	23
Figura 3.1	De arriba abajo, implementación con complejos Vitori-Rips y Alfa. De izquierda a derecha, diagramas de persistencia y código de barras de toda la ciudad.	31
Figura 3.2	De izquierda a derecha se observan los diagramas de persistencia y el código de barras de los distritos de Florida, Modelo, Esteros, Sur.	33
Figura 3.3	De izquierda a derecha se observan los diagramas de persistencia y el código de barras de los distritos de Ceibos, Pascuales y Progreso.	34
Figura 3.4	De izquierda a derecha se observan los diagramas de persistencia y el código de barras de los distritos hecho mediante una filtración distinta de Ceibos, Esteros, Florida y Modelo.	35
Figura 3.5	Matriz de correlaciones para el <i>dataset</i> .	36

ÍNDICE DE TABLAS

Tabla 1.1	Comparación entre complejos de Vietoris-Rips y Čech	11
Tabla 3.1	Resumen de características de las gráficas	28
Tabla 3.2	Resumen de características del dataset (I)	29
Tabla 3.3	Resumen de características del dataset (II)	29

CAPÍTULO 1

1. INTRODUCCIÓN

1.1 Descripción del problema

La falta de modelos predictivos de actividad delictiva en Guayaquil que empleen Análisis Topológico de Datos (TDA) para estudiar homicidios, revela una brecha crítica en la comprensión de patrones espacio-temporales complejos. Aunque la ciudad ha enfrentado desafíos persistentes de violencia, los enfoques tradicionales han dominado los estudios criminológicos, limitando la exploración de estructuras subyacentes que podrían ser clave para estrategias preventivas. En Guayaquil, los análisis de homicidios suelen basarse en métodos estadísticos clásicos como regresiones espaciales, series temporales, técnicas de machine learning o clustering geográfico, redes neuronales, tal como se evidencia en el trabajo de Correa y León [4] y el Observatorio Ecuatoriano de Crimen Organizado [9].

Si bien estos enfoques identifican correlaciones entre variables socioeconómicas, densidad poblacional o presencia policial, presentan limitaciones:

- Enfoque reduccionista: Modelan relaciones lineales o locales, pero no capturan la topología global de los datos.
- Sensibilidad al ruido: Métodos como el clustering jerárquico o K-means dependen de parámetros predefinidos y son vulnerables a outliers, lo que dificulta identificar estructuras

robustas.

- Dinámica temporal ignorada: La mayoría separa el análisis espacial del temporal, omitiendo cómo evolucionan las redes criminales o cómo se propagan los homicidios en el espacio-tiempo. El TDA, mediante herramientas como la homología persistente o los complejos de Vietoris-Rips, permite estudiar la "forma" de los datos espacio-temporales, detectando características invariantes:
- Los agrupamientos persistentes identificarían zonas donde los homicidios forman clústeres estables a través del tiempo, incluso si su ubicación exacta varía levemente.
- Los ciclos y conexiones revelarían si ciertas áreas actúan como "nodos críticos" interconectados en una red de actividad delictiva, facilitando la propagación de violencia.
- La resistencia al ruido al enfocarse en patrones que persisten ante variaciones en la escala o el muestreo, el TDA reduciría falsos positivos causados por datos fragmentados o incompletos, comunes en contextos de inseguridad.

A pesar de su potencial, no existen estudios en la ciudad que apliquen TDA a datos criminológicos. Esto contrasta con avances en otras regiones: Kingsman y Wong en [6] analizaron el problema de la brutalidad policial en Estados Unidos mediante el uso del TDA para descubrir patrones dentro de este tipo de conductas dentro de la policía estadounidense, dicho estudio se enfoca también en comparar este método con métodos tradicionales.

1.2 Justificación del problema

La delincuencia ha sido un tema de interés y preocupación en años recientes debido a que ha llegado a picos sin precedentes. Entre otros aspectos, la mala gestión pública, el abandono estatal y la falta de voluntad política han hecho más grandes las grietas de la desigualdad dentro de la sociedad ecuatoriana, lo que propicia la proliferación de grupos delincuenciales que han mantenido en zozobra al país en menos de un lustro.

Las ciudades costeras, por su proximidad al mar y el acceso a puertos, se convierten en puntos estratégicos para la salida y entrada de mercancía ilegal promovidas por el crimen organizado. Según el Observatorio Ecuatoriano de Crimen Organizado (OECO) [9], Guayaquil es considerada como una de las ciudades más peligrosas, seguida por Quevedo, Machala y Esmeraldas. En este contexto, la recopilación de datos se vuelve crucial para predecir actos delictivos como el asesinato. Si bien existen múltiples causas, en Ecuador las más comunes están ligadas al crimen organizado: enfrentamientos entre bandas, disputas territoriales y extorsión. Por eso, se necesita urgentemente un mecanismo que utilice estos datos para combatir la delincuencia.

La topología es un arma poderosa para el análisis de características comunes entre formas, por lo que recientemente se ha incrementado su uso en el análisis de datos en áreas como la neurociencia y biología junto a métodos tradicionales como regresión lineal, 'k-means', etc. En la misma línea, la ubicación geográfica es un factor clave en la problemática de la seguridad. Por ello, es esencial identificar patrones en la distribución espacial de los crímenes. Esto se logra dotando a los datos de una estructura topológica que permita medir la similaridad de los delitos en diferentes puntos de la ciudad. Al mapear esta actividad criminal, se puede brindar a las

autoridades la información necesaria para una respuesta más eficaz y oportuna.

Aunque esta herramienta no erradicará el problema por completo —ya que este tiene raíces más profundas como la desigualdad social—, sí puede ser una herramienta valiosa para las autoridades en su lucha contra la inseguridad.

1.3 Objetivos

1.3.1 Objetivo General

Diseñar un marco analítico basado en el Análisis Topológico de Datos para identificar patrones persistentes en datos delictivos de Guayaquil durante el periodo 2023-2024

1.3.2 Objetivos Específicos

- Resumir datos históricos geoetiquetados de homicidios en Guayaquil para generar información adecuada para su análisis;
- Aplicar técnicas de homología persistente y complejos simpliciales, para extraer estructuras topológicas de los datos espacio-temporales;
- Analizar correlaciones y su estabilidad entre las características topológicas de homicidios en la ciudad de Guayaquil para producir un informe de patrones generales.

1.4 Marco teórico

1.4.1 Antecedentes

En los últimos años, Ecuador ha experimentado un alarmante incremento en homicidios, concentrados principalmente en la región costera, con Guayaquil como epicentro [9]. Este

fenómeno, según Lalangui et al. en [7], refleja problemas estructurales como el narcotráfico y la debilidad institucional, agravados por recortes presupuestarios y la pandemia.

Aunque las causas exceden el alcance de este estudio, se aborda la problemática desde un enfoque innovador: el Análisis Topológico de Datos (TDA), que permite identificar patrones geoespaciales ocultos en conjuntos complejos; este, aunque poco difundido, ha demostrado su utilidad en estudios similares, como el análisis de violencia policial en EE.UU como apuntan Kinsman y Wong en [6] o patrones de movilidad urbana como en [6]. Su ventaja radica en analizar la "forma" de los datos sin reducir dimensiones, detectando agrupaciones estables (clústeres) y relaciones topológicas.

En este trabajo, se aplica TDA a los homicidios en Guayaquil (2023-2024), complementando enfoques tradicionales como k-means en [7], para revelar estructuras subyacentes en la distribución espacial de crímenes. A continuación, se presentan los conceptos matemáticos clave que sustentan nuestro análisis.

1.4.2 Conceptos claves

El Análisis Topológico de Datos (TDA) consiste en asignar invariantes topológicos a datos. Los datos son típicamente presentados como una muestra discreta cuya topología no es especialmente interesante. Así, se transforma estos datos en un objeto "continuo" que, si es posible, sea topológicamente similar a la forma del objeto geométrico original del que los datos fueron muestreados. Al final, los espacios continuos resultantes deben ser discretizados en pro de ser procesados en un computador.

Este trabajo usa una serie de datos proporcionados por el Ministerio del Interior ecuatoriano, allí se describe el estado de criminalidad que ha azotado al país en los últimos

años, donde constan las coordenadas correspondientes a lugares dentro del país donde se cometieron homicidios y las causas de los mismos. Con el objetivo de analizar la estructura geométrica subyacente detrás de estos datos, a continuación se describen los conceptos que permiten definir las variables de este problema.

Espacios topológicos

Definición 1 (Espacio métrico). (Ver [8]) Sea X un conjunto y $d : X \times X \rightarrow [0, \infty)$ una función que cumple para cualesquiera $x, y, z \in X$

- $d(x, y) \geq 0$ y $d(x, y) = 0$ si y solo si $x = y$;
- $d(x, y) = d(y, x)$;
- $d(x, y) \leq d(x, z) + d(y, z)$.

d es llamada una **métrica** en X y el par (X, d) es llamado un **espacio métrico**.

Definición 2. (Ver [8]) Sea X un conjunto y sea $\tau \subseteq 2^X$ una colección tal que

- $X \in \tau$;
- $\emptyset \in \tau$;
- $U, V \in \tau$, luego $U \cap V \in \tau$;
- $\{U_\lambda : \lambda \in \Lambda\} \subseteq \tau$, luego

$$\bigcup_{\lambda \in \Lambda} U_\lambda \in \tau.$$

La colección τ es llamada una **topología** en X , a los miembros de τ se les llama conjuntos **abiertos** de X y el par (X, τ) es llamado **espacio topológico** [8].

Definición 3. (Ver [8]) Sea (X, τ) un espacio topológico, dado cualquier $x \in X$ se define a $A \subseteq X$ como una **vecindad** de x si existe un $U \in \tau$ tal que $x \in U \subseteq A$.

Definición 4. (Ver [8]) Sean (X, τ) , (Y, η) dos espacios topológicos y sea $f : X \rightarrow Y$ una función, decimos que f es **continua** en $x \in X$ si para cada vecindad V de $f(x)$, $f^{-1}(V)$ es una vecindad de x . f es **continua** si lo es para todo $x \in X$.

Definición 5. (Ver [8]) Dos espacios topológicos (X, τ) , (Y, η) son llamados **homeomorfos** si existen funciones inversas $f : X \rightarrow Y$ y $g : Y \rightarrow X$ tal que f y g son continuas. En tal caso, f y g son llamadas **homeomorfismos** y se dice que f y g definen un **homeomorfismo** entre (X, τ) y (Y, η) .

Observación 1. (Ver [8]) Para dos espacios topológicos X, Y , una función $f : X \rightarrow Y$ es continua si para todo conjunto abierto de Y , su preimagen bajo f es un abierto de X . Si además f es un homeomorfismo, entonces también las imágenes de los abiertos de X son abiertos en Y (i.e., f es una **función abierta**).

Observación 2. (Ver [8]) Si $f : X \rightarrow Y$ es un homeomorfismo, una consecuencia inmediata de la definición es que la inversa $f^{-1} : Y \rightarrow X$ es también un homeomorfismo; por tanto Y también es homeomorfo a X . Además;

- todo espacio topológico es homeomorfo a sí mismo pues la función identidad es un homeomorfismo,
- la composición de dos homeomorfismos es a su vez un homeomorfismo.

Por tanto, la relación entre espacios topológicos dada por «ser homeomorfos» es una **relación de equivalencia**. A las clases de equivalencia bajo esta relación se les denomina "clases de

homeomorfismo”.

Observación 3. (Ver [8]) *La estructura topológica de un espacio consiste de su colección de conjuntos abiertos. Por tanto, todos los espacios de una misma clase de homeomorfismo tienen exactamente las mismas propiedades topológicas (invariantes), es decir, desde el punto de vista de la topología son todas iguales.*

Combinaciones convexas y casco convexo

Definición 6. (Ver [2]) *Dado un conjunto de puntos $S = \{x_1, \dots, x_n\}$ en $\mathbb{R}^d$, una **combinación convexa** es una expresión de la forma:*

$$x = \sum_{i=1}^n \lambda_i x_i,$$

donde $\lambda_i \geq 0$ y $\sum_{i=1}^n \lambda_i = 1$. Los elementos de S serán identificados como **vértices**, y a cualquier subconjunto de S como **espacio o arista**.

Definición 7. (Ver [2]) *El **casco convexo** $C(S)$ de un conjunto $S = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ es el conjunto de todas las combinaciones convexas de puntos en S :*

$$C(S) = \left\{ \sum_{i=1}^n \lambda_i x_i : x_i \in S, \lambda_i \geq 0, \sum_{i=1}^n \lambda_i = 1 \right\}$$

Ejemplo 1. Para $S = \{(1, 1), (2, 3), (4, 2), (3, 0)\} \subset \mathbb{R}^2$, el casco convexo $C(S)$ es el cuadrilátero mostrado en la Figura 1.1.

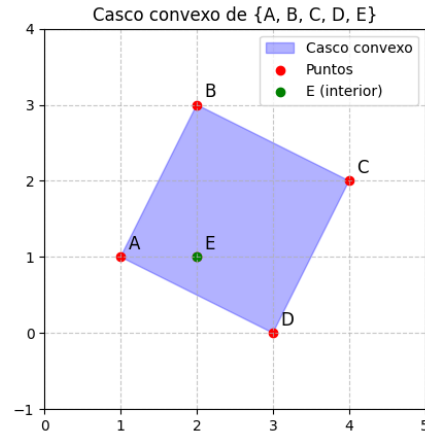


Figura 1.1.
Casco convexo de cuatro puntos en $\mathbb{R}^2$. (Elaboración propia)

k-símplices y Complejos Simpliciales

Definición 8. (Ver [3]) Un ***k*-símplex** σ es el casco convexo de $k + 1$ puntos afínmente independientes $\{v_0, \dots, v_k\} \subset \mathbb{R}^d$:

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

La **dimensión** de σ es k .

Ejemplo 2. Algunos *k*-símplex incluyen;

- 0-símplex: Punto $\{v_0\}$
- 1-símplex: Segmento de recta $\{v_0, v_1\}$
- 2-símplex: Triángulo $\{v_0, v_1, v_2\}$
- 3-símplex: Tetraedro $\{v_0, v_1, v_2, v_3\}$

Definición 9. (Ver [5]) Un **complejo simplicial** $\mathcal{K}$ es una colección finita de **símplices** en $\mathbb{R}^d$ que **satisface**:

1. Si $\sigma \in \mathcal{K}$ y τ es una cara de σ ($\tau \subset \sigma$), entonces $\tau \in \mathcal{K}$,
2. Para cualesquiera $\sigma_1, \sigma_2 \in \mathcal{K}$, $\sigma_1 \cap \sigma_2$ es vacío o es una cara común.

Ejemplo 3. Para $S = \{A, B, C, D\}$, con $A = (1, 1)$, $B = (2, 3)$, $C = (4, 2)$ y $D = (3, 0)$, el complejo simplicial completo contiene:

$$\{\emptyset, \{A\}, \{B\}, \{C\}, \{D\}, \{A, B\}, \{A, C\}, \dots, \\ \{A, B, C\}, \{A, B, D\}, \dots, \{A, B, C, D\}\}$$

con **dimensión máxima 3**.

Definición 10. (Ver [3]) Dado $P \subset \mathbb{R}^d$ y $r \geq 0$, el complejo $VR(P, r)$ contiene todos los subconjuntos $\sigma \subset P$ con $\text{diam}(\sigma) \leq 2r$:

$$VR(P, r) = \{\sigma \subset P : \|x - y\| \leq 2r \ \forall x, y \in \sigma\}$$

Definición 11. (Ver [3]) El complejo $\check{C}(P, r)$ contiene $\sigma \subset P$ si $\bigcap_{p \in \sigma} B_r(p) \neq \emptyset$:

$$\check{C}(P, r) = \left\{ \sigma \subset P : \bigcap_{p \in \sigma} B_r(p) \neq \emptyset \right\}$$

Tabla 1.1.*Comparación entre complejos de Vietoris-Rips y Čech*

Propiedad	Vietoris-Rips	Čech
Base geométrica	Diámetro $\leq 2r$	Intersección de bolas
Computación	Más eficiente	Más precisa
Relación	$\check{C}(P, r) \subseteq VR(P, r) \subseteq \check{C}(P, 2r)$	

Homología Simplicial

Definición 12. (Ver [5]) Sea $\mathcal{K}$ un complejo simplicial y d su dimensión. Una **d -cadena** es la suma de los d -simplexes en $\mathcal{K}$. La notación usada en este contexto será $c = \sum_i a_i \sigma_i$ donde σ_i son d -simplexes de $\mathcal{K}$ y a_i son coeficientes.

Los coeficientes en este caso, son tomados de un campo. El ejemplo más común de esto es el campo de los números reales $\mathbb{R}$. Sin embargo, se pueden usar otros campos como puede ser el que corresponde al campo de las clases de equivalencia módulo 2 que, por su relación con la naturaleza binaria del computador, es usado extensivamente en el álgebra computacional. Siguiendo el mismo esquema para sumar polinomios, se suman las d -cadenas, esto es: si p y q son dos d -cadenas tales que $p = \sum_i a_i \sigma_i$ y $q = \sum_i b_i \sigma_i$, entonces:

$$p + q = \sum_i (a_i + b_i) \sigma_i$$

Definición 13. (Ver [5]) La **frontera** de un d -simplex es la suma de sus $d - 1$ caras. Si tenemos una cara $\sigma = [a_0, a_1, \dots, a_d]$, entonces su frontera es:

$$\partial_d \sigma = [a_1, \dots, a_d] + [a_0, a_2, \dots, a_d] + [a_0, a_1, \dots, a_{j-1}, a_{j+1}] + \dots + [a_1, \dots, a_{d-1}].$$

Es decir, es la suma de todos los complejos simpliciales formados por los vértices de σ donde es

omitido el vértice j -ésimo.

∂_p define un homomorfismo entre los complejos simpliciales. $\partial_n : C_n \rightarrow C_{n-1}$. ∂_n es llamado homomorfismo de frontera o mapa de frontera definido por las cadenas. El complejo de cadenas es una sucesión de grupos de cadenas conectados por el homomorfismo de frontera:

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

Las d-cadenas definidas anteriormente junto con su respectiva suma, forman un grupo de d-cadenas denotado como $(C_d, +)$, si no hay confusión sobre la operación, se denotará como $C_d(\mathcal{K})$. Dado un complejo simplicial $\mathcal{K}$, su **homología simplicial** en dimensión n es el grupo cociente:

$$H_n(\mathcal{K}) = Z_n(\mathcal{K}) / B_n(\mathcal{K})$$

donde:

- $Z_n(\mathcal{K}) = \ker \partial_n$ (ciclos)
- $B_n(\mathcal{K}) = \text{im } \partial_{n+1}$ (bordes)

Considere la figura 1.2, donde se tiene el complejo simplicial formado por los vértices $e_1, e_2, e_3, e_4, e_5, e_6$

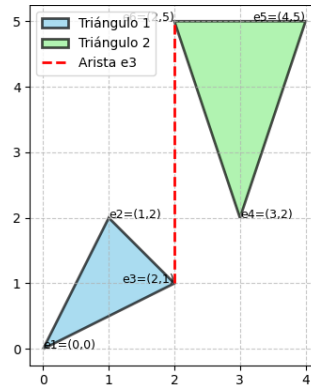


Figura 1.2.
Representación grafica de un complejo simplicial formado por triángulos. (Elaboración propia)

Tomando este esquema, se puede dar una interpretación geométrica a las definiciones. C_0 corresponderá a las combinaciones lineales de Vértices. C_1 serán las combinaciones lineales de aristas y C_2 son las combinaciones lineales de triángulos. Si se tiene una arista, denotada como $a = [e_1, e_2]$ entonces $\partial_1(a) = e_2 - e_1$, esto es el borde de una arista, es la resta de sus puntos.

Persistencia Topológica

Definición 14 (Filtración). (Ver [5]) Una **filtración** de $\mathcal{K}$ es una secuencia anidada:

$$\mathcal{K}_1 \subseteq \mathcal{K}_2 \subseteq \dots \subseteq \mathcal{K}_m = \mathcal{K}$$

Definición 15 (Código de barras). (Ver [5]) El **código de barras** de una filtración es la colección de intervalos $[a, b)$ que representan el nacimiento (a) y muerte (b) de clases de homología.

La homología de persistencia es una herramienta fundamental en el Análisis Topológico de Datos (TDA) que estudia la evolución de características topológicas (como componentes

conexas, agujeros o cavidades) a través de múltiples escalas. Su objetivo principal es identificar qué características persisten y cuáles son efímeras cuando se varía un parámetro de escala.

1.5 Filtraciones

Definición 16. (Ver [5]) Una **filtración** es una familia anidada de espacios topológicos o complejos simpliciales $\{F_r\}_{r \in T}$ donde $T \subseteq \mathbb{R}$, tal que para $r \leq r'$ se cumple $F_r \subseteq F_{r'}$. Como se mencionó en líneas anteriores, Si tenemos un complejo simplicial $\mathcal{K}$ y una familia de subcomplejos provenientes del mismo, se tiene que la filtración definida a partir de este subcomplejo, es una familia de subcomplejos tales que:

$$\mathcal{K}_1 \subseteq \mathcal{K}_2 \subseteq \dots \subseteq \mathcal{K}_m = \mathcal{K}$$

Ejemplo 4. En la figura 1.2 se observa que si se tiene el complejo simplicial generado por los puntos $e_1, e_2, e_3, e_4, e_5, e_6$, entonces se forma el complejo $\mathcal{K} = [e_1, e_2, e_3, e_4, e_5, e_6]$. Del cual tenemos los subcomplejos: $\mathcal{K}_1 = [e_1, e_2]$, $\mathcal{K}_2 = [e_1, e_2, e_3]$, $\mathcal{K}_3 = [e_1, e_2, e_4, e_5]$. Así la filtración generada es:

$$\mathcal{K}_1 \subseteq \mathcal{K}_2 \subseteq \mathcal{K}_3 \subseteq \mathcal{K}$$

1.5.1 ¿Cómo se construyen filtraciones?

Si los elementos en $\mathcal{D}$ están en un espacio métrico (como podría ser puntos en $\mathbb{R}$), entonces $\mathcal{D}$ es compacto. El conjuntos de complejos de Vietoris-Rips, $(Rips_r(\mathcal{D}))_{r \in \mathbb{R}}$ y los complejos de Čechs, son filtraciones.

Definición 17 (Módulos de persistencia). Un **módulo de persistencia** P en un subconjunto T de números reales, es una familia indexada de espacios vectoriales $(V_r | r \in T)$ y una familia doble

indexada de mapas lineales $h_s^r : V_r \rightarrow V_s | r \leq s$ el cual satisface la ley de composición $h_t^s h_s^r = h_t^r$ cuando $r \leq s \leq t$, y h_t^r es la identidad de V_r .

Chazal y Michel en [3] indican que si se define una filtración, como es el caso de las filtraciones generadas por los complejos de Vietoris-Rips, es posible definir módulos de persistencia sobre estas mediante los grupos de homología producidos sobre la filtración subyacente. De la misma forma que se indica en la definición, se inducen funciones lineales sobre los módulos de persistencia de esta filtración. Una vez definido el modulo de persistencia (el cuál es un espacio vectorial), es posible descomponer el espacio vectorial en la suma directa de intervalos modulares $\mathbb{I}_{(x,y)}$. Como se acostumbra en la terminología matemática, x es el ínfimo del intervalo, mientras que y es el supremo.

Los códigos de barras se forman a partir de estos intervalos y permiten ver la duración de las características topológicas que persisten a medida de que se avance a la derecha. Así mismo, se forma el diagrama de persistencia: en el plano cartesiano ubicamos la recta $f(x) = x$, sobre esta diagonal se van a ubicar los ínfimos y los supremos de cada intervalo. Como se observará todos los puntos en el diagrama de persistencia están por encima de la recta identidad. Formalmente para un diagrama de persistencia se tiene:

1.6 Diagramas de Persistencia

Definición 18 (Diagrama de persistencia:). *El **diagrama de persistencia** $dgm(V)$ es una representación multiconjunto de intervalos $[b, d)$ donde:*

- *b : Parámetro donde nace una característica topológica*
- *d : Parámetro donde muere (o $d = \infty$ si persiste).*

Ejemplo 5. Para la homología en dimensión 0 (H_0):

- Cada componente conexa nace en algún $r = b$ (birth)
- Dos componentes se fusionan en $r = d$ (death), donde $b < d$
- La componente "más joven" (aparecida más recientemente) es la que muere

1.6.1 Propiedades del TDA

Definición 19 (Estabilidad). [3] Pequeñas perturbaciones en los datos producen pequeños cambios en el diagrama (medidos por la bottleneck distance).

Definición 20 (Ruido topológico). [3] Puntos cercanos a la diagonal $\Delta = \{(b, b) | b \in \mathbb{R}\}$ representan características efímeras

Ejemplo 6. Considere la función $f(x) = \sin(x)$ en $[0, 2\pi]$:

- Componentes conexas (H_0):
 - Nacen en mínimos locales ($\pi/2$)
 - Mueren al unirse en máximos ($3\pi/2$)
- Ciclos (H_1):
 - Un ciclo nace al rodear el "agujero" del seno
 - Muere cuando el nivel supera el máximo global

1.7 Otros métodos para la obtención de patrones en la data.

1.7.1 Agrupamiento mediante *k-means*.

El advenimiento del *Machine Learning* está íntimamente ligado al estudio y a la estructura de datos, por lo cual se han desarrollado técnicas para descubrir patrones dentro de estos. Una forma de hacerlo es etiquetando y agrupando datos. La distancia, en el sentido topológico, entre los puntos de un grupo de datos permite que estos sean agrupados y etiquetados por afinidad. De esta motivación nacen los *clústeres* o agrupamiento de datos. En general, cuando se parte de un conjunto de datos, el cual está en un espacio métrico, tomamos esta métrica para segmentar los datos. La letra K en el nombre del algoritmo nos dice el número de clusters que se quieren.

En el Apéndice A se encuentra una versión del algoritmo de Lloyd (1957) que primero fue usado para la modulación de pulso-código.

Las entradas para este algoritmo son un conjunto de datos tomados de $\mathbb{R}^d$, con la distancia usual. El conjunto de datos son los puntos $S = \{x_1, x_2, \dots, x_M\}$. Este algoritmo coloca a cada uno de estos punto en alguno de los k -clústeres definidos. El vector e_i con $i \in \mathbb{N}$, representa al vector que tiene 1 en la i -ésima coordenada y 0 en el resto de coordenadas. Este vector dentro del algoritmo, se encarga de asignar a cada vector de S a su respectivo clúster. El algoritmo retorna r_j , el cual es el vector de S que va al clúster j y nos da un conjunto con cada una de las medias μ_k de cada uno de los k -clústeres [1].

CAPÍTULO 2

2. METODOLOGÍA

2.1 Formulación de alternativas

Estudiar el desarrollo del crimen en el Ecuador o, en particular, en Guayaquil, no es un problema nuevo. Se ha partido de diferentes enfoques para llegar a una solución del problema.

Lalangui et al. en [7] calcularon las tasas de homicidio en Latinoamérica con especial énfasis en Ecuador con datos que van desde el 2019 al 2022, usando el método de *k-means*. Para este fin, calculan el riesgo excesivo o relativo y agrupan los datos en segmentos, que son producidos por el algoritmo del método usado. Para ello utilizan una métrica de este tipo de análisis conocida como *riesgo excesivo* el cual mide la mortalidad relativa en una población. Los autores señalan que se quiere investigar la mortandad vista como un asunto de Salud Pública.

Dentro del mismo artículo resaltan el trabajo del *INEC* (Instituto Ecuatoriano de Estadísticas y Censos), que se encarga de recopilar estos datos, al mismo tiempo que se hacen análisis descriptivos de los mismos, por lo cual este tipo de frecuencias o métricas, suele aparecer en este tipo de análisis previo. Lo cual es una ventaja al querer, aplicar otro tipo de técnicas estadísticas para el análisis de datos.

Dentro de su investigación, los investigadores señalan que a pesar de los resultados obtenidos es importante abordarlos bajo otros enfoques. Reconociendo que dentro de su

estudio, al menos en el caso de Ecuador, la forma en como los datos fueron obtenidos y registrados, son una fuente de limitaciones. Mencionaron los problemas de codificación de la causa de la muerte, es decir, los errores que surgen al asignar códigos estandarizados a las causas de fallecimiento en los certificados de defunción. Sin embargo, declaran que para efectos prácticos, los datos son "representativos y adecuados para el análisis de homicidios a nivel cantonal y nacional".

En cuanto a la técnica usada para obtener una tendencia dentro de una serie de tiempo, la prueba de Mann-Kendall tiene una mayor robustez y precisión si se lo usa en secciones grandes de tiempo. Sin embargo, el estudio solo considera un periodo de siete años. Además, usan prueba de hipótesis para validar sus hipótesis, que a pesar de ser un procedimiento estándar dentro de la estadística, corta parte de los datos.

Por este motivo se escogió el TDA haciendo un uso extensivo de la persistencia homológica y a diferencia de otros métodos de análisis, el presentado en este proyecto se destaca por su estabilidad, por el uso intensivo de todas las características de cada punto en el *dataset* y por ser inmune al cambio de parámetros, lo cual nos permite preservar la integridad de los datos.

2.2 Diseño del modelo utilizado

Se consideró en enfocar el análisis a la extracción de patrones de otros crímenes como robos o extorsiones pero se decidió finalmente escoger los homicidios por la desafortunada relevancia actual que estos tienen como problemática social en la ciudad de Guayaquil y por la existencia de bibliografía con la cual comparar directamente los resultados de este trabajo.

2.3 Persistencia Homológica.

Como se la ha descrito en líneas anteriores, la persistencia homológica produce una estructura de naturaleza algebraica la cual es llamada grupos de homología. Como se ha señalado, los grupos de homología encapsulan distintas propiedades topológicas. En general, estos grupos de homología son denotados como H_i , donde $i \in \mathbb{N}$, es la dimensión de este subespacio. De manera más intuitiva, la información que brindan estos grupos es:

- H_0 describe el número de componentes conexas. En el caso de nuestro dataset formado por puntos en $\mathbb{R}^5$, cada punto es en sí mismo una componente conexa.
- H_1 describe el número de huecos.
- H_2 describe el espacio vacío en superficies.

Cada uno de estos grupos de homología se tiene una dimensión $n + 1$. También pueden existir grupos de homología de dimensiones superiores, pero a diferencia de grupos en estas tres dimensiones, es más complicada su visualización geométrica.

El diseño conceptual de la metodología se estructuró como un flujo de trabajo secuencial, concebido para transformar los datos crudos de homicidios en una estructura matemática coherente y apta para el TDA. El proceso partió de un conjunto de datos tabular, del cual se seleccionaron cinco variables consideradas fundamentales para el análisis: tipo de muerte, las dos coordenadas geográficas (x, y) , la presunta motivación y la motivación observada. Esta selección inicial se orientó a reducir la dimensionalidad del problema y enfocar el análisis en los aspectos espaciales y causales del delito. Una vez aisladas estas características, se ejecutó una fase crítica de preprocesamiento para garantizar la calidad de los datos. Este paso incluyó la

corrección y estandarización de entradas textuales que presentaban inconsistencias o errores tipográficos, asegurando así que categorías conceptualmente idénticas fueran tratadas de manera uniforme.

Posteriormente, se procedió con la etapa de transformación, que fue esencial para preparar los datos para los algoritmos matemáticos. Todas las variables categóricas, como el tipo de muerte y las motivaciones, fueron convertidas a una representación numérica mediante un mapeo sistemático, donde a cada categoría textual única se le asignó un valor entero distintivo. Este proceso de codificación fue fundamental, ya que los algoritmos de TDA operan sobre nubes de puntos en espacios vectoriales. Finalmente, el flujo concluyó con una fase de validación y purificación de datos. En esta última etapa, se verificó la integridad numérica de todas las celdas y se eliminaron de forma sistemática aquellos registros (filas) que, tras la conversión, presentaban datos faltantes o corruptos. El resultado de este diseño fue la generación de una matriz numérica final, limpia y completa, donde cada fila representaba un homicidio como un vector en $\mathbb{R}^5$. Esta matriz constituyó la nube de puntos de entrada, lista para ser sometida al análisis topológico con el fin de descubrir sus patrones estructurales subyacentes.

2.4 Visualización de la persistencia Homológica.

Para la visualización de persistencia homológica se fijará en la siguiente secuencia de imágenes, el cual constituye un ejemplo de juguete, que simula la creación de grupos de homología representados como grafos, en los que aparecerán características topológicas como componentes conexas y aristas.

Esto último es una simplificación del proceso de creación para los complejos simpliciales.

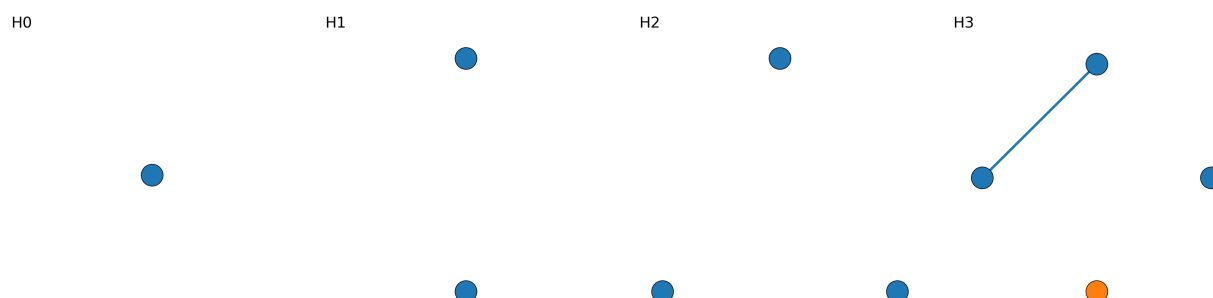


Figura 2.1.

Representación de una serie de grafos que representan grupos de homología, donde aparecen distintas propiedades topológicas.

Como se mencionó en el capítulo uno, para este fin se necesitará un conjunto de datos, del cual se construirá un cascarón convexo, sobre el cual se extraerá grupos simpliciales o filtraciones, las mismas que serán formadas a través de los complejos de Vietori-Rips o alpha. Una vez formadas estas filtraciones, puede ocurrir que dos vértices se conecten o que se separen. De esta manera vemos como se van creando o más bien aparecen ciertas estructuras topológicas en los datos. En la figura 2.1, se observa que cada vértice aislado es en sí mismo una componente. A medida que continúa H_1 , se añade otra componente, manteniendo la que estaba al inicio. Si se avanza a H_4 , se forma una arista, fusionando dos vértices. Esta arista es en sí misma una característica topológica que consumió la primera que estaba al inicio, y el segundo vértice que aparece en H_2 . Esta información se puede condensar si se grafica un plano en el intervalo $[0, 4]$, y se coloca sobre la duración la que persisten las características topológicas. por ejemplo el al vértice de H_0 , continua hasta H_4 donde se funde con otro vértice, se asignará el punto $(0, 6)$.

Se obtiene un gráfico de persistencia, como el visto en la Figura 2.2: encima de la función identidad se grafican los puntos correspondientes a donde inician y terminan los

atributos topológicos de nuestro complejo simplicial en la Figura 2.1.

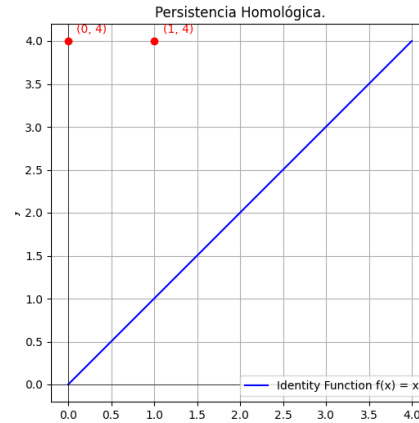


Figura 2.2.
Diagrama de persistencia. (Elaboración propia)

Por otra parte, como se señaló en el capítulo anterior, el código de barras encapsula esta información graficando barras que corresponden a la duración de una determinada característica topológica. Dicha información puede ser extraída directamente del diagrama de persistencia, al observar cada par (a, b) del plano de persistencia como un intervalo. En el diagrama de código de barras, como el visto en la Figura 2.3, se aprecia la persistencia de características topológicas, generadas por cada uno de los complejos simpliciales; al observar cada par (a, b) del plano de persistencia como un intervalo según su duración o persistencia en el diagrama de persistencia.

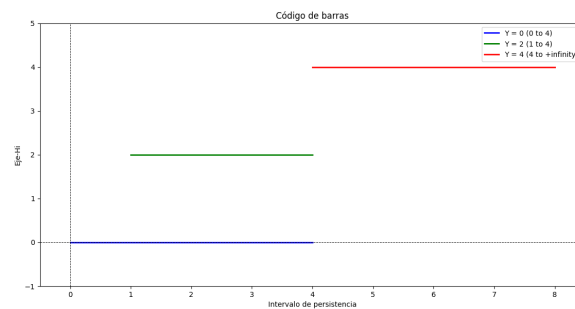


Figura 2.3.
Diagrama de código de barras. (Elaboración propia)

2.5 Especificaciones técnicas

Para la ejecución del proyecto, se configuró un entorno de análisis en *Python* (versión 3.9) utilizando librerías especializadas como *Pandas* para la manipulación de datos y *Gudhi* para la implementación del análisis topológico. El procedimiento se detalla a continuación.

Se extrajo un documento de formato *.xlsx* de la base de datos abiertos del gobierno nacional donde figura información histórica de homicidios y asesinatos en el país desde enero del 2023 hasta diciembre del 2024: *DatosHomicidios2023-2024.xlsx*. Para optimizar el rendimiento y evitar posibles errores de memoria, se trabajó con una muestra aleatoria de dos mil datos para la construcción de la estructura topológica. Para esta tarea la muestra fue producida por el método *MaxMin* que intencionalmente selecciona puntos que están lejos entre sí, creando una muestra que cubre el *dataset* en toda su extensión, a continuación se describe su funcionamiento

1. Escoger un punto aleatorio para iniciar y agregarlo a la muestra.
2. Encontrar el punto del *dataset* más lejano del punto escogido y agregarlo a la muestra.
3. Para cualquier otro punto del *dataset*, calcular la distancia mínima a cada uno de los puntos en la muestra.
4. Escoger el punto que tiene el máximo de estas distancias mínimas. Es decir, hallar el punto que está más lejos de todos los puntos de muestra seleccionados.
5. Repetir los dos pasos anteriores hasta tener dos mil puntos.

Este método fue escogido porque escoge puntos en *clusters*, atípicos y extremos de los datos que preservan la estructura geométrica a larga escala, que es lo que el TDA analiza.

Luego, se importó el archivo actualizado `DatosHomicidios2023-2024.xlsx` a un *dataframe* de Pandas. De este marco de datos inicial, se seleccionaron y aislaron estas ocho columnas de interés para el análisis

- `tipo_muerte`,
- `coordenada_x`,
- `coordenada_y`,
- `presunta_motivacion`,
- `presun_motiva_observada`,
- `edad`,
- `genero` y
- `ocupacion_registro_civil`.

Algunas de estas columnas contienen datos categóricos que para la implementación del TDA debieron ser transformados. Antes de la transformación numérica, se realizó un proceso de limpieza para asegurar la consistencia de esos datos categóricos, además de la identificación de diversas inconsistencias textuales y errores tipográficos en algunas columnas que fueron corregido aplicando un mapeo de reemplazo.

Con esta nueva base de datos estandarizados, se procedió a la transformación de todas las variables categóricas en valores numéricos. Se crearon diccionarios de mapeo para cuatro de las columnas; por ejemplo para `tipo_muerte`, se asignaron enteros únicos a cada categoría, así "ASESINATO" se convirtió en 1.

La fase final del preprocesamiento se centró en garantizar la integridad y completitud del conjunto de datos. Primero, se validó que los valores en las columnas fueran de tipo numérico, convirtiendo cualquier entrada no válida en un valor nulo, `NaN`, para luego ser reemplazada por la moda de los datos de la respectiva columna.

El resultado de este proceso fue una matriz numérica limpia, la cual fue guardada en un archivo binario de *NumPy* nombrado como `M_Numerica.npy`. Esta matriz, que representa la nube de puntos final, se convirtió en la entrada directa para la aplicación de los algoritmos de TDA.

La siguiente fase del proceso se centró en la aplicación del algoritmo topológico para extraer características estructurales ocultas en la nube de puntos multidimensional. Se importó el archivo binario `M_Numerica.npy` utilizando la librería *NumPy*, cargando la matriz numérica directamente en la variable llamada `datosCargados_Homicidios`.

La librería *Gudhi* se empleó para construir un Complejo de Rips, se inicializó `gudhi.RipsComplex` utilizando la muestra de datos. El parámetro `max_edge_length` se configuró en 86, determinando el radio máximo para la formación de aristas en el complejo; este valor fue ajustado empíricamente para capturar la conectividad relevante entre los puntos.

A partir del Complejo de Rips, se generó un árbol de símlices asignado a la variable `simplex_tree` a través del método `rips_complex.create_simplex_tree(max_dimension=5)`, una estructura de datos eficiente que almacena la información de los complejos simpliciales. Además, se realizó una implementación del algoritmo basada en Complejos Alfa generando, similarmente, un árbol de símlices. En este contexto, la `max_dimension` se limitó a 5 para optimizar el uso de memoria y computación, enfocándose en características topológicas de bajas dimensiones como nodos, bucles, cavidades, etc.

La persistencia topológica se calculó invocando el método `simplex_tree.persistence()`. El resultado es un conjunto de pares (nacimiento, muerte) para cada característica, indicando las escalas en las que aparecen y desaparecen. Para interpretar los patrones identificados, se generaron dos visualizaciones clave utilizando la librería *matplotlib* y las funciones de trazado de *Gudhi*: el **Diagrama de Persistencia** y **Diagrama de Código de Barras**.

Finalmente, se realizó la implementación de ambos complejos a subconjuntos del *dataset* correspondiente a cada uno de los ocho sectores de la ciudad de Guayaquil. Todas estas visualizaciones, generadas mediante el código descrito, permitieron una interpretación visual de la estructura subyacente de los datos de homicidios, revelando patrones y conectividades.

CAPÍTULO 3

3. RESULTADOS Y ANÁLISIS

En esta sección se escribe exhaustivamente la interpretación de los diagramas de persistencia y de código de barras producidos de la aplicación del algoritmo de TDA descrito en la sección anterior.

3.1 Homologías y contextualización de los datos

Para el desarrollo del trabajo se tomó los datos abiertos del Gobierno Nacional. Este *dataset* recopila las muertes en la ciudad de Guayaquil durante los años 2023 y 2024 en donde se listan las muertes y las características del evento. Los peritos reconocen 34 características, sin embargo, se tomó 8 de ellas para el fin del presente trabajo; tipos de crimen, 2 coordenadas espaciales, edad, género, presunta motivación, presunta motivación observada y oficio en el registro civil, pues estas relacionan las características del crimen con las de la víctima.

Tabla 3.1.
Resumen de características de las gráficas

Homologías		
dimensión 0	dimensión 1	dimensión 2
arista	bucle	huecos

La tabla 3.1 resume los conceptos de homologías producidos por la implementación del TDA en pro de tener un lenguaje común durante el análisis de resultados de los diagramas de persistencia y código de barras.

Por otro lado, junto al *dataset*, se tiene un diccionario de términos en donde se resume la información del primero, de la cual se destaca la proporción de cada característica dentro de una columna. Las tablas 3.2 y 3.2 resumen las proporciones más destacables de 5 de esas columnas.

Tabla 3.2.

Resumen de características del dataset (I)

Características		
tipo de muerte	4255 asesinatos	62 homicidios
presunta motivación	4146 casos de delincuencia común	143 casos de violencia comunitaria
presunta motivación observada	2007 casos de microtráfico	1896 casos de amenaza
género	3999 personas de género masculino	294 personas de género femenino
profesión en registro civil	2005 estudiantes	308 bachilleres

Tabla 3.3.

Resumen de características del dataset (II)

Características		
tipo de muerte	18 femicidios	7 sicarios
presunta motivación	36 casos de violencia intrafamiliar	9 caso de terrorismo
presunta motivación observada	99 riñas	1 abigeato
género	2 personas trans	2 personas transexuales
profesión en registro civil	187 empleados	33 personas que realizan quehaceres domésticos

3.2 Discusión de resultados

De ahora en adelante, a las características de dimensiones superiores las denotaremos como espacios n -dimensionales entre superficies.

3.2.1 Resultados globales

Ambas implementaciones del TDA vistas en la figura 3.1 muestran que las características más persistentes son las aristas, lo que implica que en el *dataset* los crímenes son cometidos de manera aislada y aquellos que no, serán asimilados en un *clúster* a medida que avanzan etapas en la filtración, aquí se destaca que en la implementación de filtraciones basadas en complejos Vitori-Rips no aparecen ni huecos ni espacios de más de una dimensión entre superficies, en donde los huecos en su mayoría mueren instantáneamente al no despegarse de la diagonal del gráfico. Los diagramas producido por filtraciones basadas en complejos Alfa muestran también a los huecos muriendo instantáneamente y en la mismas etapas que los bucles salvo algunas excepciones. Además, ambos diagramas de código de barras muestran que los nodos son más persistentes: "engulliendo" a homologías de dimensiones superiores a lo largo de cada intervalo.

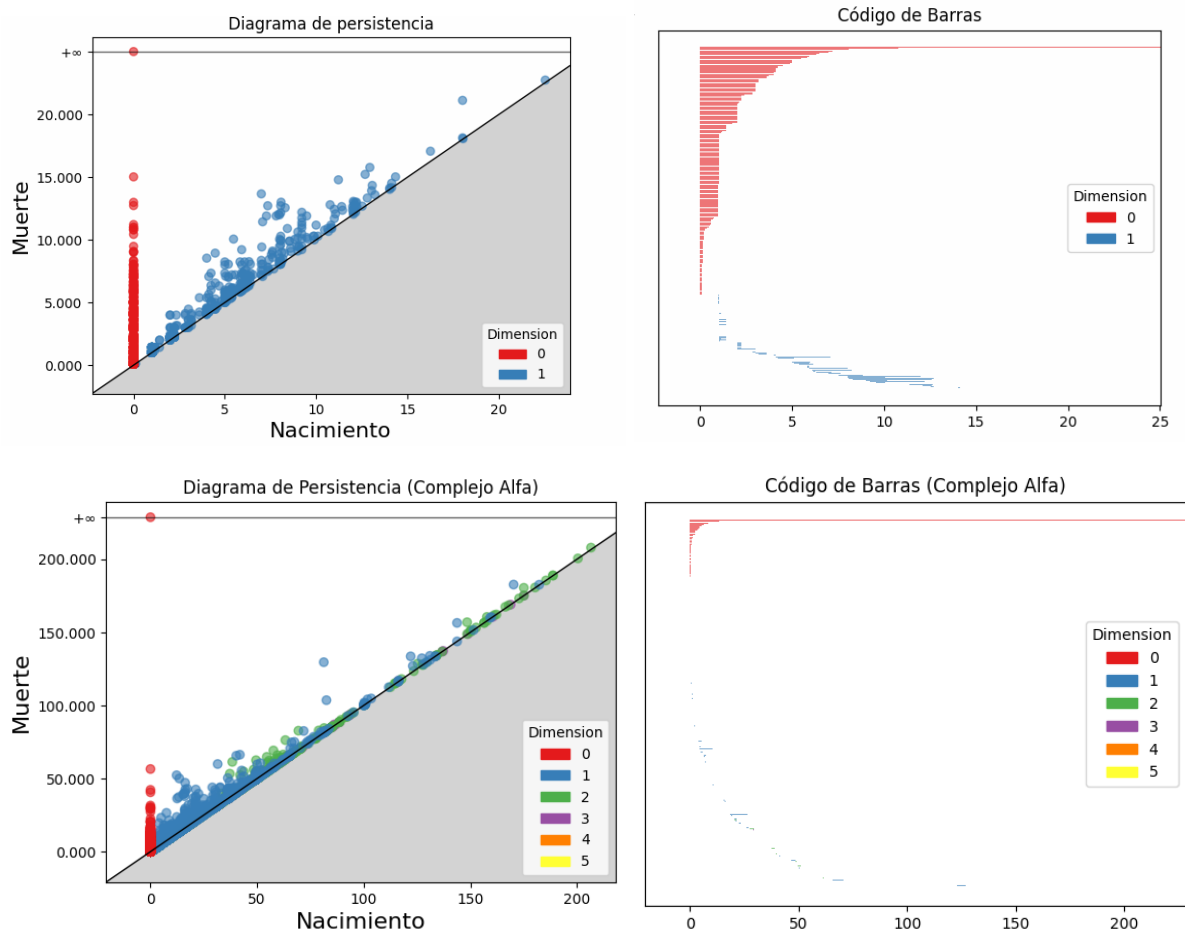


Figura 3.1.

De arriba abajo, implementación con complejos Vitori-Rips y Alfa. De izquierda a derecha, diagramas de persistencia y código de barras de toda la ciudad.

3.2.2 Resultados locales

Con el motivo de comparar los resultados globales con los de locales, se realizó una sectorización de la ciudad de Guayaquil en sus distritos: Esteros, Florida, Ceibos, Sur, Modelo, Pascuales, etc. La realidad de cada uno de estos sectores es muy distinta entre sí y a priori se esperaban resultados distintos, sin embargo, la data muestra algo diferente.

En la figura 3.2 se observa que en estas parroquias su alarmante número de muertes en los dos años analizados. En Florida se registraron 321 muertes, en Modelo 151, en Esteros 462, y en el Sur de Guayaquil Metropolitano 688. En estas parroquias, la violencia no es ajena a ellas y esta se ha ido incrementando paulatinamente.

Algo que puede verse a priori en los diagramas es que, en síntesis, las características de dimensión uno mueren a penas nacen se ve en la figura 3.1 mientras que las características de dimensión cero persisten por mucho más tiempo. Esto puede sugerir que la escala en la que se presentan los datos o la distancia en la que están dispersos los datos puede ser muy grande, lo cual hace que la mayor cantidad de ellos se queden aislados.

Los datos señalan que lo común en estos sectores es la violencia que se genera dentro de ellos pero donde se presenta mayores índices de violencia es en el sur de Guayaquil, donde está ubicado un punto estratégico importante para los distintos grupos delincuenciales que operan en la ciudad: el puerto marítimo. Como se observa en el diagrama de persistencia y en el código de barras en el extremo derecho del de la figura 3.2, las características de dimensión uno duran más que en los anteriores gráficos. Si bien están cerca de la diagonal, muchas de las características de dimensión uno se despegan de esta, lo cual puede indicar que se forman bucles por mucho más tiempo respecto a los otros sectores. Esto puede ser un indicativo de la densidad de violencia en

el sur de Guayaquil o de la existencia de características similares entre cada crimen destacando la ubicación, la edad y la motivación observada.

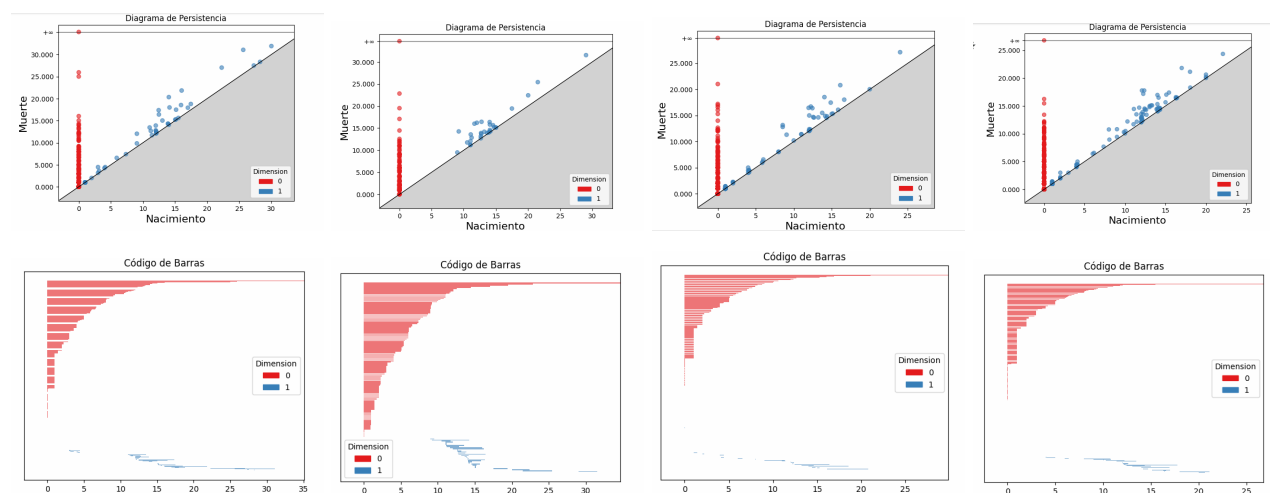


Figura 3.2.

De izquierda a derecha se observan los diagramas de persistencia y el código de barras de los distritos de Florida, Modelo, Esteros, Sur.

Por otra parte, las zonas presentadas en la figura 3.3 destacan por sus distintas condiciones geográficas y socioeconómicas. Pascuales es una parroquia dentro de Guayaquil ubicada en el norte de la ciudad, Progreso es una parroquia rural del cantón y Ceibos es una zona de un alto crecimiento económico, a priori se podría esperar que los resultados en estos sectores difieran.

En efecto, los datos dan evidencia de que en los años 2023 y 2024, en Ceibos hubieron 49 homicidios, 607 en Pascuales y 329 casos de homicidio en Progreso. Al igual que en resultados anteriores, se observa que las características de dimensiones superiores a cero duren menos. Lo que es perenne en los diagramas de persistencia, donde las características de dimensión uno están cerca de la diagonal, es decir son absorbidas rápidamente por las de dimensión cero.

Aunque en Ceibos existe una diferencia importante en el número de homicidios respecto a los otros sectores de Guayaquil, se conserva el patrón de persistencia de las características de dimensión cero. En este sentido, como antes se destaca la importancia de la escala pues en este caso, la distancia de cada punto, puede ser más grande.

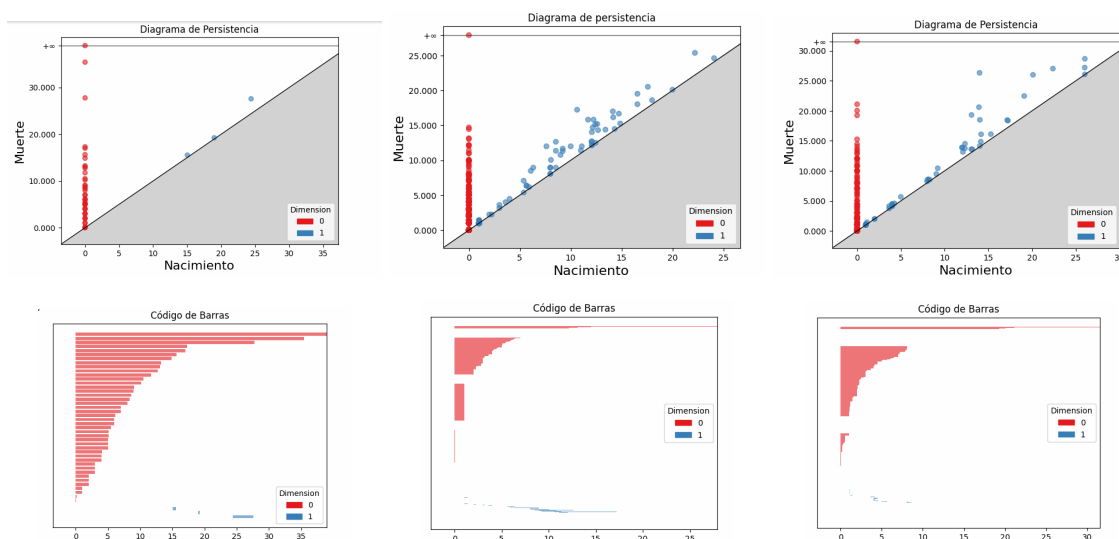


Figura 3.3.

De izquierda a derecha se observan los diagramas de persistencia y el código de barras de los distritos de Ceibos, Pascuales y Progreso.

Al igual que en las dos gráficas 3.3 y 3.2, la figura 3.4 muestra que las características de dimensión cero prevalecen frente a las de dimensión superior a uno. No obstante, se observa que algunos bucles persisten más que en otros gráficos que a priori implica que cada crimen del bucle comparte al menos una característica por determinado tiempo.

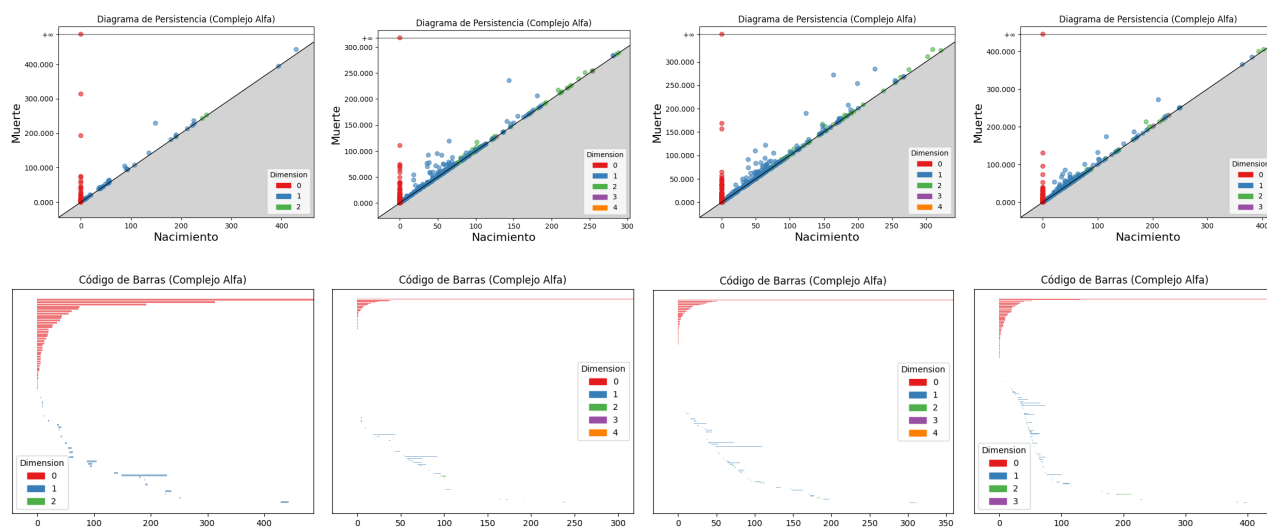


Figura 3.4.

De izquierda a derecha se observan los diagramas de persistencia y el código de barras de los distritos hecho mediante una filtración distinta de Ceibos, Esteros, Florida y Modelo.

La matriz de correlaciones vista en la Figura 3.5 mide la proporción lineal entre dos variables, en ese sentido se destaca que las columnas de coordenadas exhiben la correlación positiva más fuerte y que la mayoría es cercana a cero, positiva o negativa. Este hecho explica el por qué no es posible concluir, por ejemplo, la razones exactas por la que un bucle se rompe o por las que es casi imposible la formación de homologías de dimensión superior a dos, así se conjetura la presencia de cualidades en los datos que no se aprecian en las columnas del *dataset* que afectan a la construcción de los complejos simpliciales, se resalta el reemplazo de características cualitativas a cuantitativas como una de estas posibles cualidades. Además, esto podría explicar la disposición natural de las homologías en aristas y el colapso de homologías de dimensión superior a cero en las de orden cero y uno.

En este sentido, es destacable el hecho de que los diagramas de persistencia y de códigos de barras de barras encapsulan la información estadística que se puede hallar por otros medios como la presentada en la matriz de correlaciones.

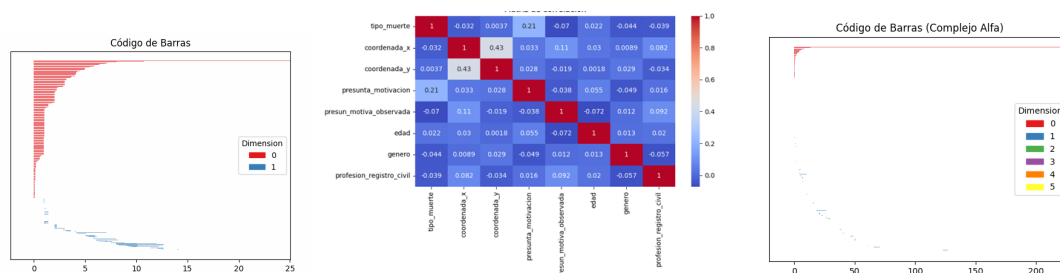


Figura 3.5.
Matriz de correlaciones para el dataset.

CAPÍTULO 4

4. CONCLUSIONES Y RECOMENDACIONES

En este trabajo se utilizó el TDA para analizar varias características persistentes de un *dataset* donde los elementos corresponden a las víctimas mortales de la ciudad de Guayaquil durante 2023 y 2024. Cada uno de estos datos contiene ocho atributos a los cuales se les aplicó el TDA, este resumió información de los mismos presentada a través de diagramas que ilustran las relaciones subyacentes en el conjunto de datos.

Conclusiones

- La implementación de filtraciones que determinan los subsecuentes complejos simpliciales, como los de Vietori-Rips y Alfa utilizados en este trabajo, producen perspectivas diferentes sobre la estructura subyacente de los datos.
- Las características de dimensión cero persisten en el tiempo más que otro tipos de características, esto es; los bucles, huecos, etc son engullidos por las aristas y nodos. Además, se sospecha que existen propiedades en los datos ajenas al *dataset* que afectan a la creación de complejos simpliciales y podrían explicar los patrones descritos.
- La información producida por el TDA vista en los diagramas se acopla bien a la la presentada en la matriz de correlaciones mostrando que el método junto a análisis estocásticos están en una relación sinérgica.

Recomendaciones

- Basado en los distintos resultados que los análisis con complejos Vitori-Rips y Alfa arrojaron, se recomienda la realización de un estudio comparativo y sistemático de diferentes métodos de construcción de complejos simpliciales. Dicho estudio se sugiere, no debería centrarse únicamente en la eficiencia computacional, sino en cómo cada método afecta la estabilidad e interpretabilidad de las características topológicas resultantes. El objetivo de tal análisis debería ser en tal caso, desarrollar un marco de trabajo que guíe a investigadores a seleccionar el complejo más adecuado para sus datos y objetivos, permitiendo así la obtención de conclusiones más robustas y reveladoras sobre las estructuras subyacentes del fenómeno estudiado.
- En el presente estudio, los datos cualitativos fueron reemplazados por un número correspondiente al orden en el que aparecían en la columna respectiva del *dataset*. Sin embargo, como se ha señalado antes, el tipo de asignación escogido podría afectar a la creación de los complejos simpliciales y a los resultados de persistencia. Así, se recomienda un tratamiento adecuado para datos de carácter cualitativo de forma que se pueda refinar el análisis.
- El análisis topológico y el análisis estocástico se pueden complementar en el estudio de una colección de datos. El análisis topológico provee una perspectiva global de estos, mientras que un estudio estadístico y probabilístico, podría dar una perspectiva local en la colección de datos, proveyendo al investigador de una vista más completa de la información. Se recomienda hacer un análisis estocástico de la duración de las características, tomando el

promedio y otras medidas descriptivas de las mismas que complemente al TDA.

- Expandir la aplicabilidad del TDA en estudios posteriores a través de la combinación de las recomendaciones anteriores con el objetivo de analizar un fenómeno determinado de la manera más completa posible. Además se exhorta a los investigadores a explorar la posibilidad de traducir las característica topológicas al mundo real, lo que permite a las organizaciones correspondientes tomar mejores decisiones de recursos estatales y de seguridad.

BIBLIOGRAFÍA

- [1] Adams, R. P. (2018). K-means clustering and related algorithms. *Princeton University*.
- [2] Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (2006). *Nonlinear programming: theory and algorithms*. John Wiley & sons.
- [3] Chazal, F. and Michel, B. (2021). An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in artificial intelligence*, 4:667963.
- [4] Correa Cuadrado, D. D., León Cazar, G. J., et al. (2022). Análisis de correlación de homicidios y rendimiento académico: estudio enfocado en unidades educativas de guayaquil, durán y samborondón período 2013-2014. *ESPOL. FCSH*.
- [5] Edelsbrunner, H. and Harer, J. (2010). *Computational topology: an introduction*. American Mathematical Society.
- [6] Kingsman, D. and Wong, T. A. (s.f). The homological persistence of police violence: analysis and limitations. Manuscrito disponible bajo solicitud.
- [7] Lalangui, K., Sánchez-Carrillo, C., and Quentin, E. (2025). Evolución espacio-temporal de los homicidios en ecuador de 2015 a 2022. *Investigaciones Geográficas*.
- [8] Mendelson, B. (2012). *Introduction to Topology*. Courier Corporation, 3 edition.
- [9] Observatorio Ecuatoriano de Crimen Organizado (2024). Boletín anual de homicidios 2023: Variación de los homicidios intencionales en ecuador (enero - diciembre 2023). Reporte técnico,

Pan American Development Foundation (PADF), Quito, Ecuador. Coordinado por Alicia Contero y Renato Rivera.

APÉNDICES

APÉNDICE A

Algoritmo A.1. Algoritmo K-Means Clustering (Algoritmo de Lloyd)

Entrada: Vectores de datos $\{x_m\}_{m=1}^M$, número de clusters K

Salida: Asignaciones $\{c_m\}_{m=1}^M$, medias de clusters $\{\mu_k\}_{k=1}^K$

```
1 for  $m \leftarrow 1$  hasta  $M$  do
2    $c_m \leftarrow [0, 0, \dots, 0]$ ; // Inicializar responsabilidades
3    $k' \leftarrow \text{EnteroAleatorio}(1, K)$ ; // Asignación aleatoria inicial
4    $c_{mk'} \leftarrow 1$ ;
5 end
6 return  $\{c_m\}_{m=1}^M, \{\mu_k\}_{k=1}^K$ ; // Retornar asignaciones y centroides
7 repeat
8   for  $k \leftarrow 1$  hasta  $K$  do
9      $M_k \leftarrow \sum_{m=1}^M c_{mk}$ ; // Número de puntos en el cluster  $k$ 
10     $\mu_k \leftarrow \frac{1}{M_k} \sum_{m=1}^M c_{mk} x_m$ ; // Calcular nueva media
11  end
12  for  $m \leftarrow 1$  hasta  $M$  do
13     $c_m \leftarrow [0, 0, \dots, 0]$ ; // Resetear responsabilidades
14     $k' \leftarrow \arg \min_k \|x_m - \mu_k\|^2$ ; // Encontrar cluster más cercano
15     $c_{mk'} \leftarrow 1$ 
16  end
17 until ninguno de los  $c_m$  cambie;
```
