Robustez de la homología persistente: El Teorema de estabilidad

15/06/2021

Alejandro García Castellanos, z17m008

Tutor: Héctor Barge Yañez

Índice

Introducción

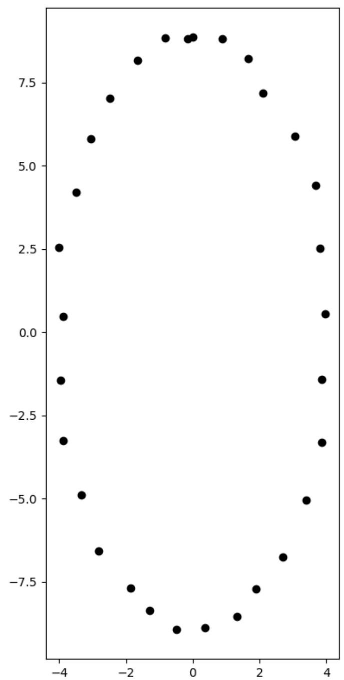
Análisis topológico de datos

El análisis de datos topológicos (TDA) es un enfoque para el análisis de las propiedades cualitativas geométricas de datasets utilizando técnicas de topología.

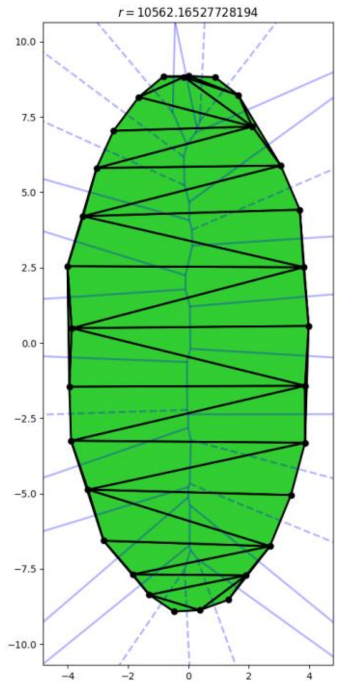
- Propiedades cualitativas geométricas: Componentes conexas, agujeros, cavidades…

- **Ventajas:**
  - Estudiar características de la "forma" de datos de dimensión superior que no se pueden visualizar.
  - No dependen de la elección de una métrica
  - <u>Proporcionan estabilidad frente al ruido</u> → **Teorema de estabilidad**

Pipeline



Nube de puntos



Filtración de complejos simpliciales

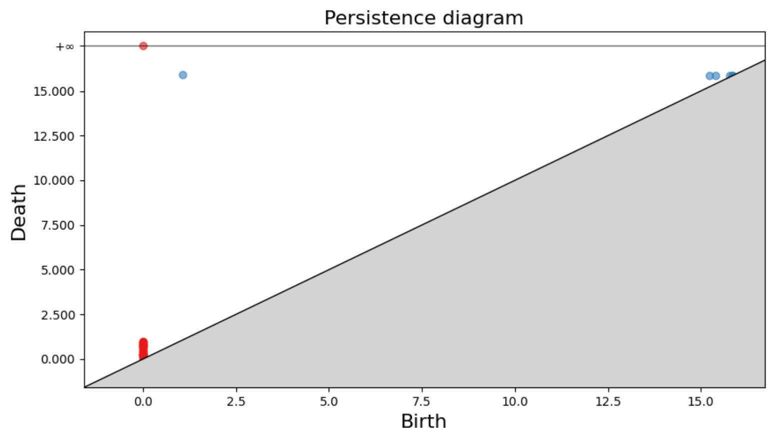
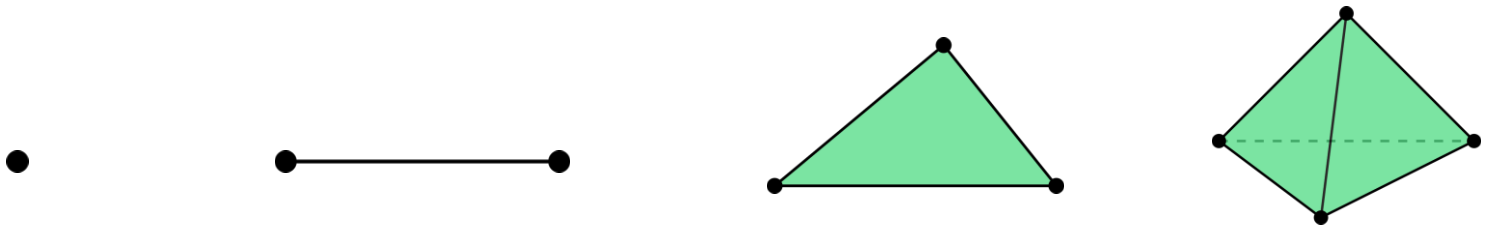


Diagrama de persistencia

Conocimientos previos

Complejo simplicial

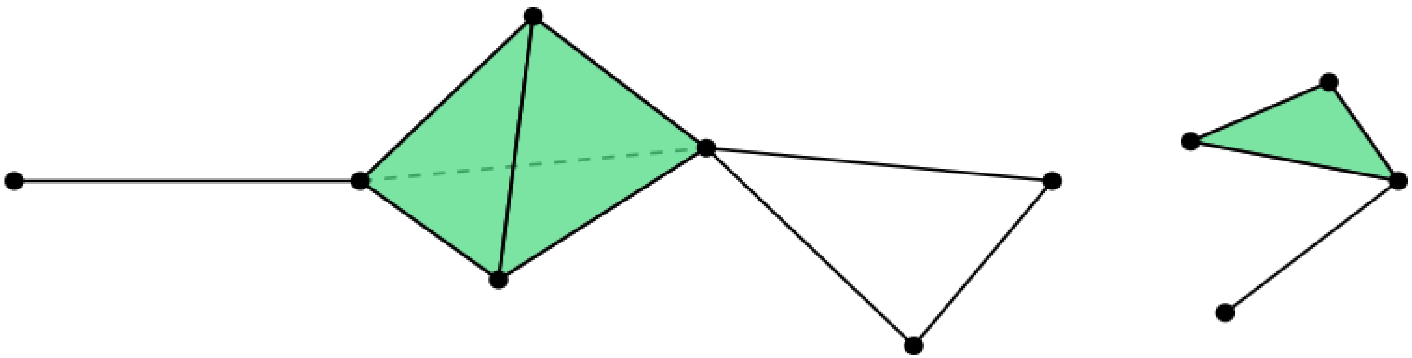
- <u>Def:</u> Un k-*símplice* σ en $\mathbb{R}^d$ con $d \geq k$ es un triángulo k-dimensional.



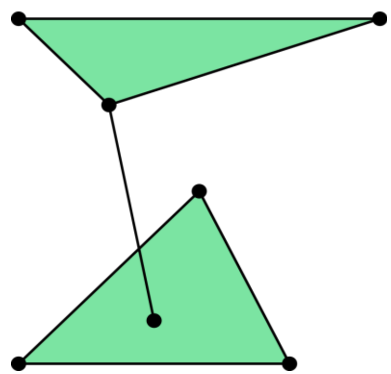
0-símplice 1-símplice 2-símplice 3-símplice

- <u>Def:</u> Un ***complejo simplicial*** es una colección finita de símplices K que satisface que las intersecciones (no vacías) entre los símplices son símplices de menor dimensión, pertenecientes al complejo simplicial K.

Complejo simplicial



Es un complejo simplicial



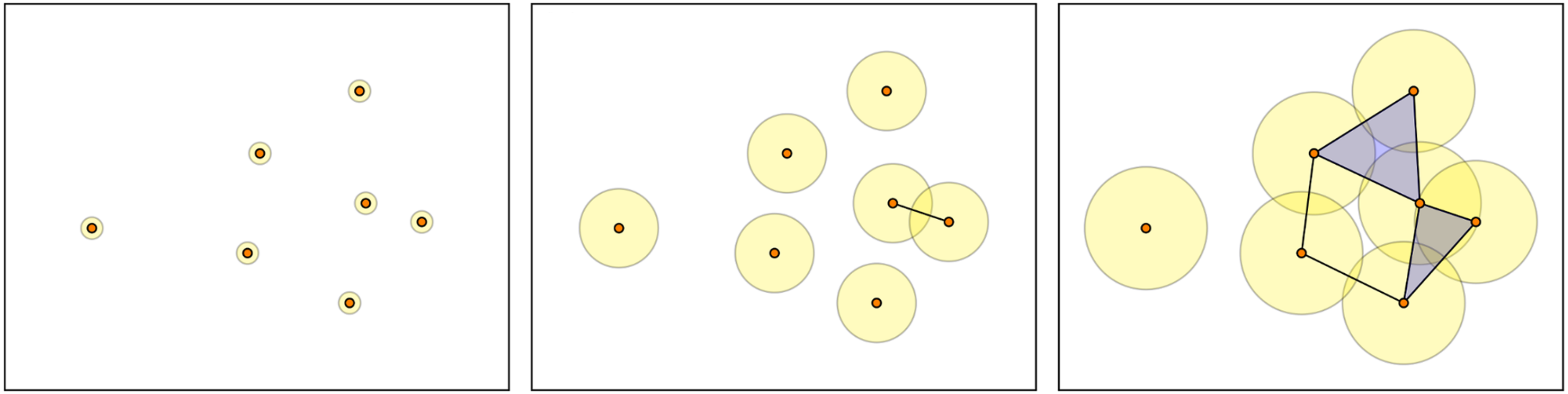
No es un complejo simplicial

Complejo de Vietoris-Rips

- Def: Sea $S \subset \mathbb{R}^d$ un conjunto finito de puntos. Llamamos ***complejo de Vietoris-Rips*** de S de radio r al complejo simplicial abstracto

$$\mathrm{VR}(r) = \{\sigma \subseteq S \mid \operatorname{diam} \sigma \leq 2r\}$$

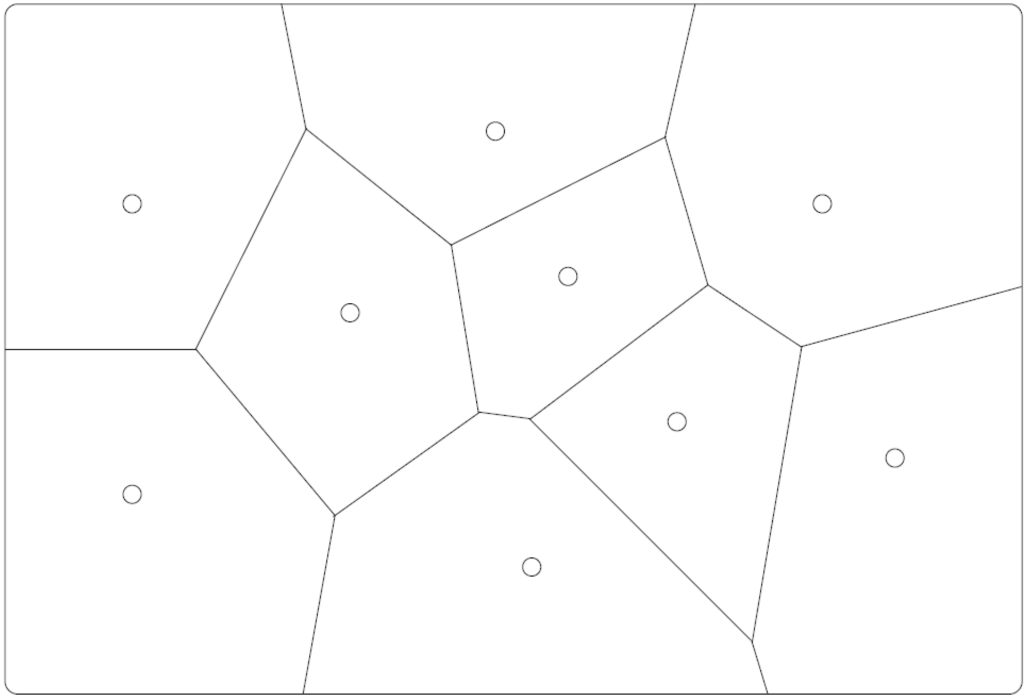
donde $\operatorname{diam} \sigma$ denota el diámetro del subconjunto σ.



Fuente: [1]

Celdas de Voronoi

- Def: La ***celda de Voronoi***, V_u, de un punto $u \in S$ es la intersección de los semiespacios de puntos al menos tan cerca de u como de v para todos los puntos $v \in S$.



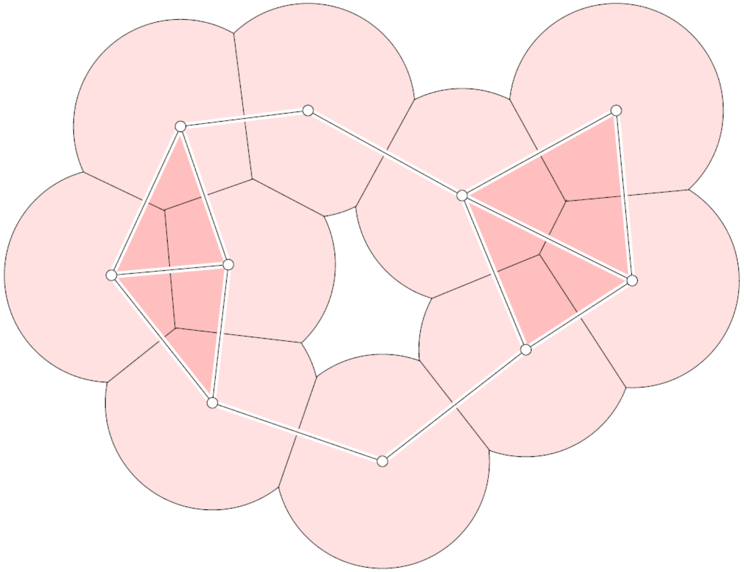
Fuente: [1]

Alfa complejo

- <u>Def:</u> Sea $S \subset \mathbb{R}^d$ un conjunto finito de puntos. Llamamos ***alfa complejo*** de radio r asociado a S como el complejo simplicial abstracto

$$\mathrm{Alpha}(r) = \{ \sigma \in S \mid \cap_{u\in\sigma} R_u(r) \neq \emptyset \}$$

donde $R_u(r) = \overline{B}_r(u) \cap V_u$.



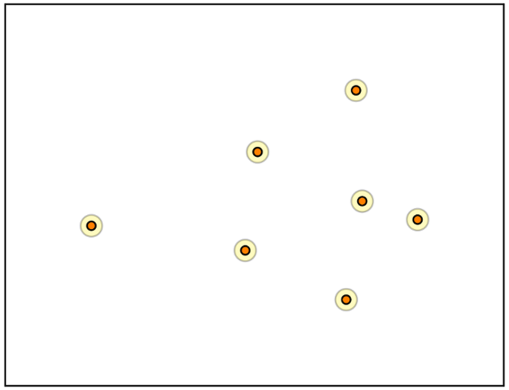
Fuente: [1]

Homología simplicial

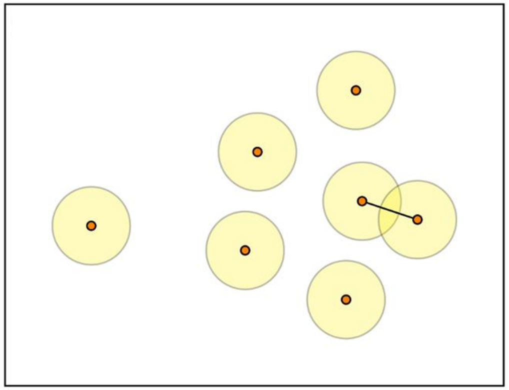
- Formalismo algebraico que nos permitirá contar:
  - Componentes conexas.
  - Agujeros.
  - Cavidades.
  - Etc.

- Def: Sea un objeto geométrico X, definimos $\beta_i(X)$, el i-ésimo ***número de Betti*** de X, como el número de agujeros i-dimensionales de X.

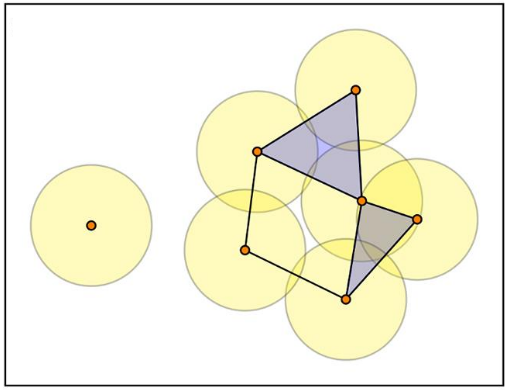
- Nos va a permitir calcular los números de Betti de un complejo simplicial haciendo uso del álgebra lineal.



- $\beta_0(K) = 7$ (Comp. Conexas)
- $\beta_1(K) = 0$ (Huecos)
- $\beta_2(K) = 0$ (Cavidades)



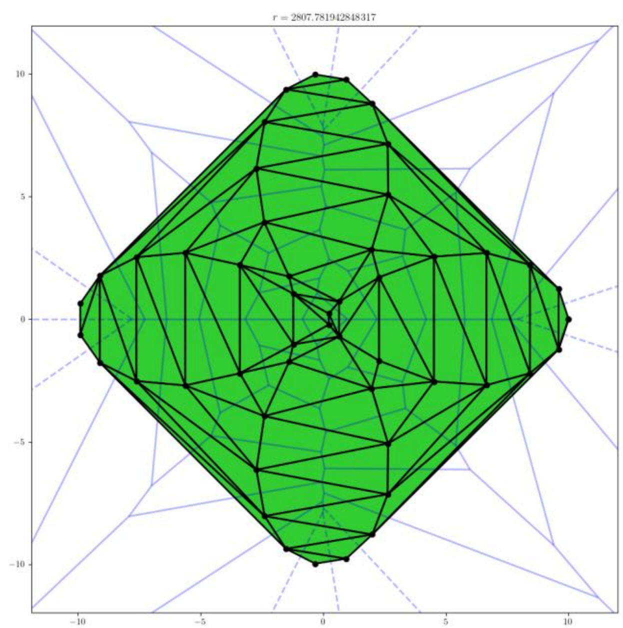
- $\beta_0(K) = 6$ (Comp. Conexas)
- $\beta_1(K) = 0$ (Huecos)
- $\beta_2(K) = 0$ (Cavidades)

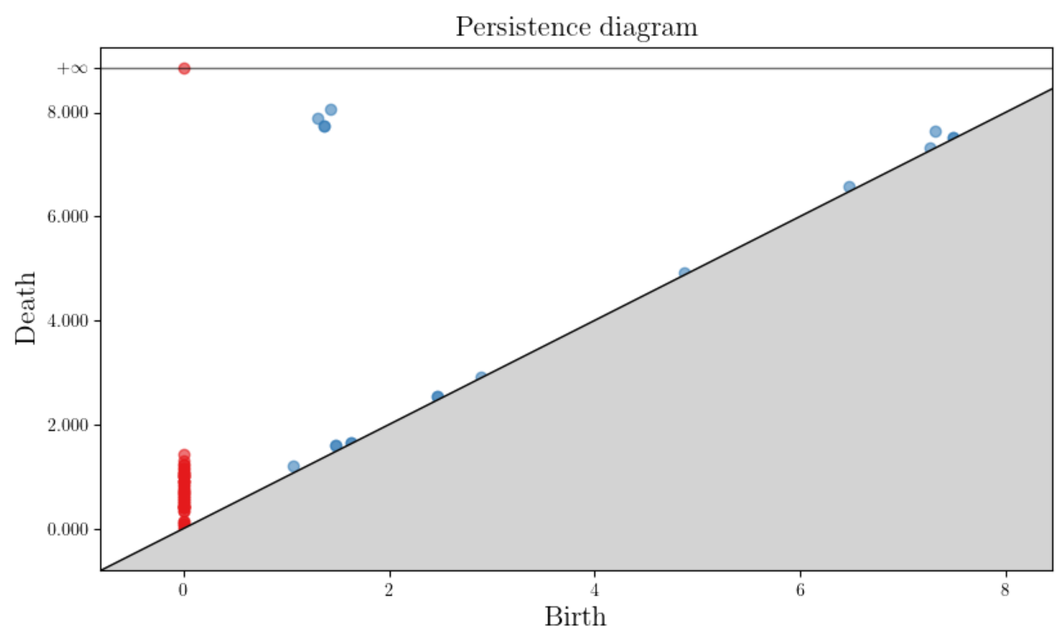


- $\beta_0(K) = 2$ (Comp. Conexas)
- $\beta_1(K) = 1$ (Huecos)
- $\beta_2(K) = 0$ (Cavidades)

Homología persistente

- Def: El ***diagrama de persistencia*** $\mathrm{Dgm}_l(K) \subset \overline{\mathbb{R}}^2$ de K es el multiconjunto de puntos (a_i, a_j) con multiplicidad μ_i^j para todo $0 \le i < j \le n+1,$ unión los puntos de la diagonal, con multiplicidad infinito.
  - Hay μ_i^j agujeros l-dimensionales que nacen en el "instante" a_i y mueren en a_j.



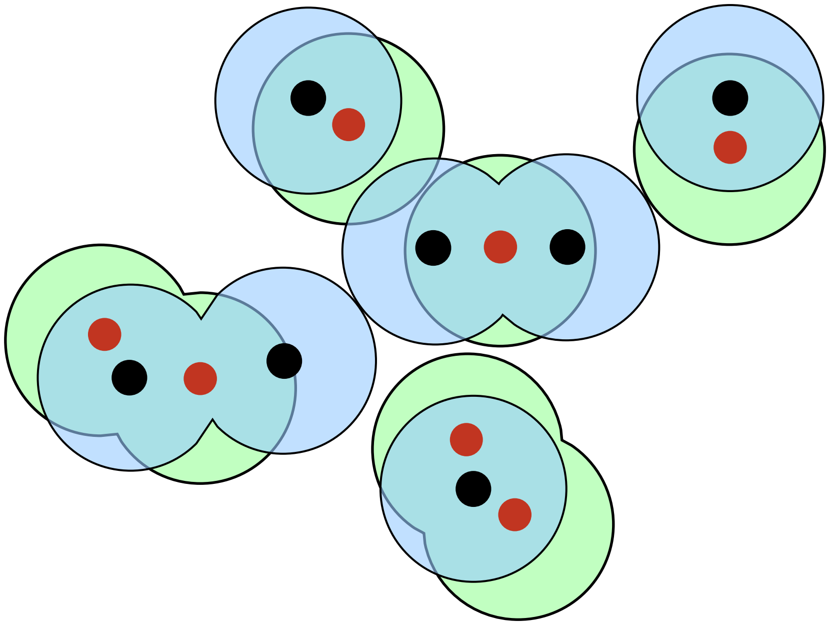


Teorema de estabilidad

Distancia Hausdorff

Sean X e Y dos conjuntos finitos. Entonces, la ***distancia Hausdorff*** entre X e Y es

$$H(X,Y) = \text{máx}\left\{\sup_{x\in X}\ \inf_{y\in Y} \|x-y\|_\infty, \sup_{y\in Y}\ \inf_{x\in X} \|y-x\|_\infty\right\}$$

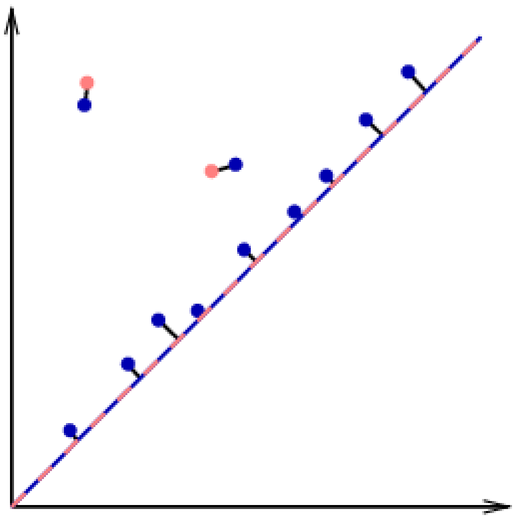


Distancia Bottleneck

Sean X e Y dos multiconjuntos. Entonces, la ***distancia bottleneck*** entre X e Y es

$$W_\infty(X,Y) = \inf_{\eta: X \to Y} \sup_{x \in X} \|x - \eta(x)\|_\infty$$

siendo η: $X \to Y$ las biyecciones de X a Y.



Fuente: [2]

Teorema de estabilidad

Sean A y B subconjuntos finitos de $\mathbb{R}^d$. Entonces, para cada k

$$W_\infty\big(\mathrm{Dgm}_k(A), \mathrm{Dgm}_k(B)\big) \leq H(A, B).$$

Nubes de puntos "cercanas" dan lugar a diagramas de persistencia "cercanos".

Este resultado se puede generalizar de forma que podemos garantizar la robustez de los diagramas de persistencia asociados a una función real definida en un espacio topológico, bajo ciertas hipótesis leves [2].

Implementación

Cálculo de la distancia Hausdorff

$$H(X,Y) = \max\left\{ \underbrace{\sup_{x\in X}\inf_{y\in Y} \|x-y\|_\infty}_{\breve{H}(X,Y)}, \underbrace{\sup_{y\in Y}\inf_{x\in X} \|y-x\|_\infty}_{\breve{H}(Y,X)} \right\}$$

Cálculo de las distancias Hausdorff directas:

Algoritmo NAIVEHDD [3] del orden de $\mathcal{O}(n*m)$, donde $m = |X|$ y $n = |Y|$.

Cálculo de la distancia Bottleneck

$$W_\infty(X,Y) = \inf_{\eta:X\to Y} \sup_{x\in X} \|x-\eta(x)\|_\infty$$

Búsqueda de la biyección η: $X \rightarrow Y$ de distancia mínima

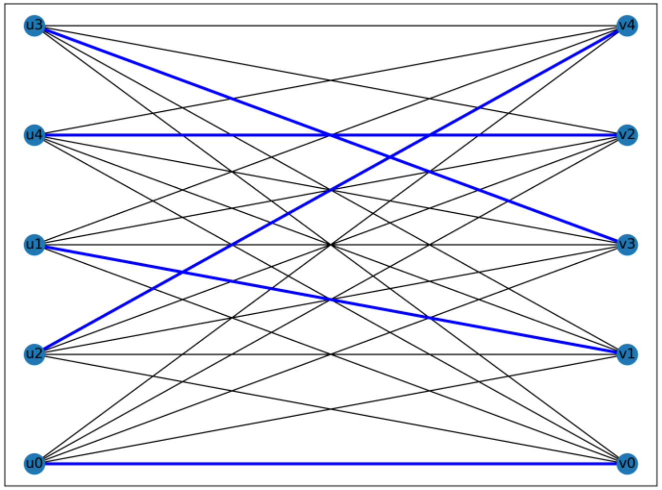
$\downarrow$ Modelización

Búsqueda de un emparejamiento óptimo de coste mínimo en un grafo bipartido

$\downarrow$ Resolución

Variante del método Húngaro, el cual se utiliza para resolver problemas de asignación

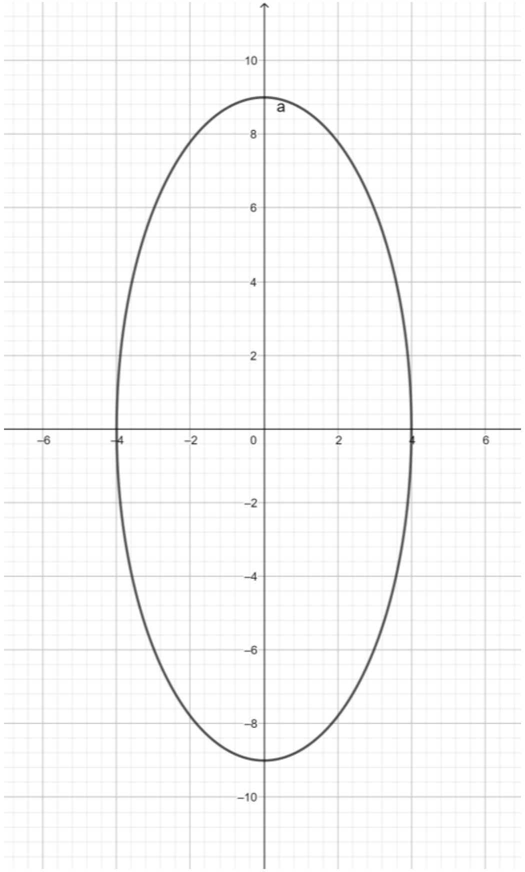
Algoritmo [1] del orden de $\mathcal{O}(n^3)$.

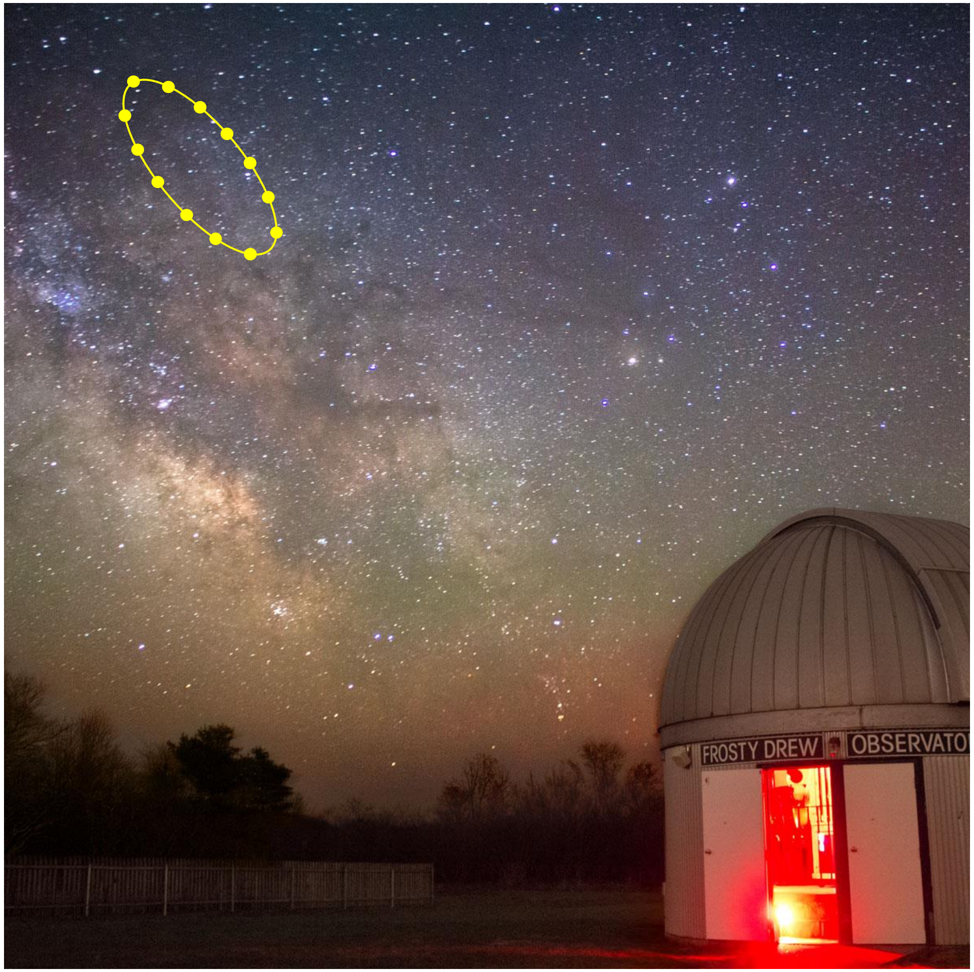


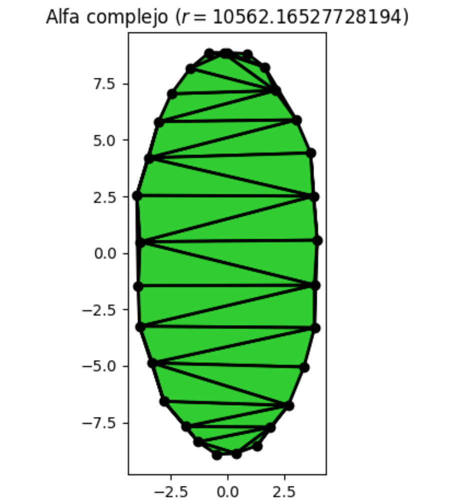
Ejemplos

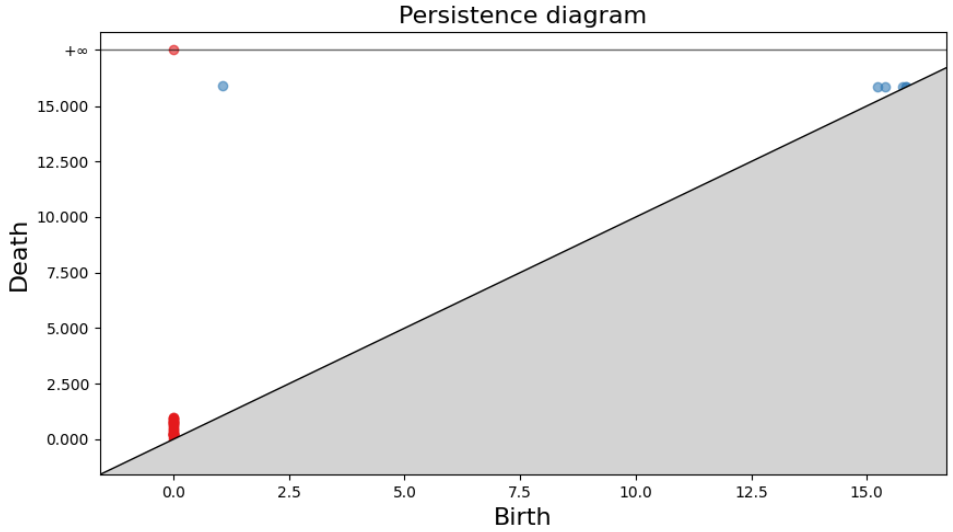
Ejemplo 1: Elipse

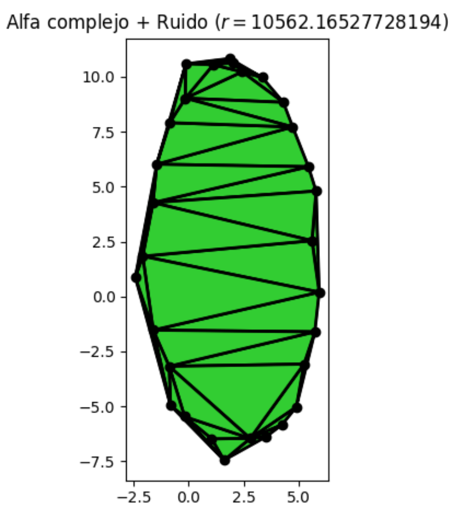
$\gamma(t) = (4\sin(t), 9\cos(t))$, con $t \in [0,2\pi]$

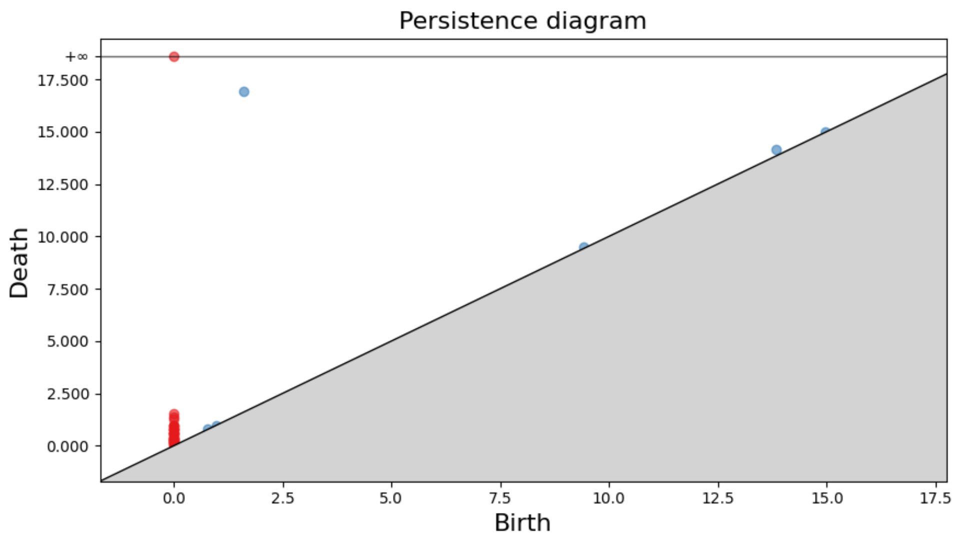












Ruido:

$N(2,0.09)$

Dist. Hausdorff:

2.4398

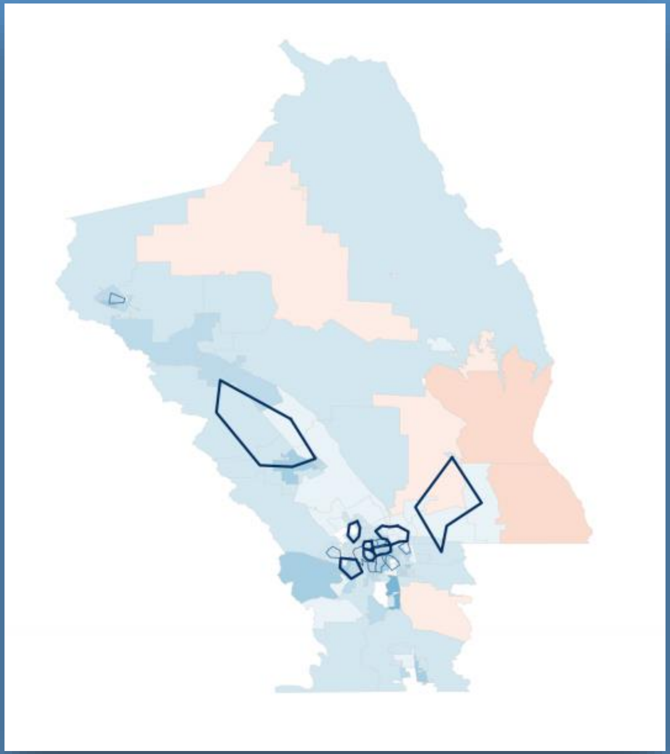
Dist. Bottleneck dim. 0: 0.55573

Dist. Bottleneck dim. 1: 0.98572

Ejemplo 2: Elecciones 2019

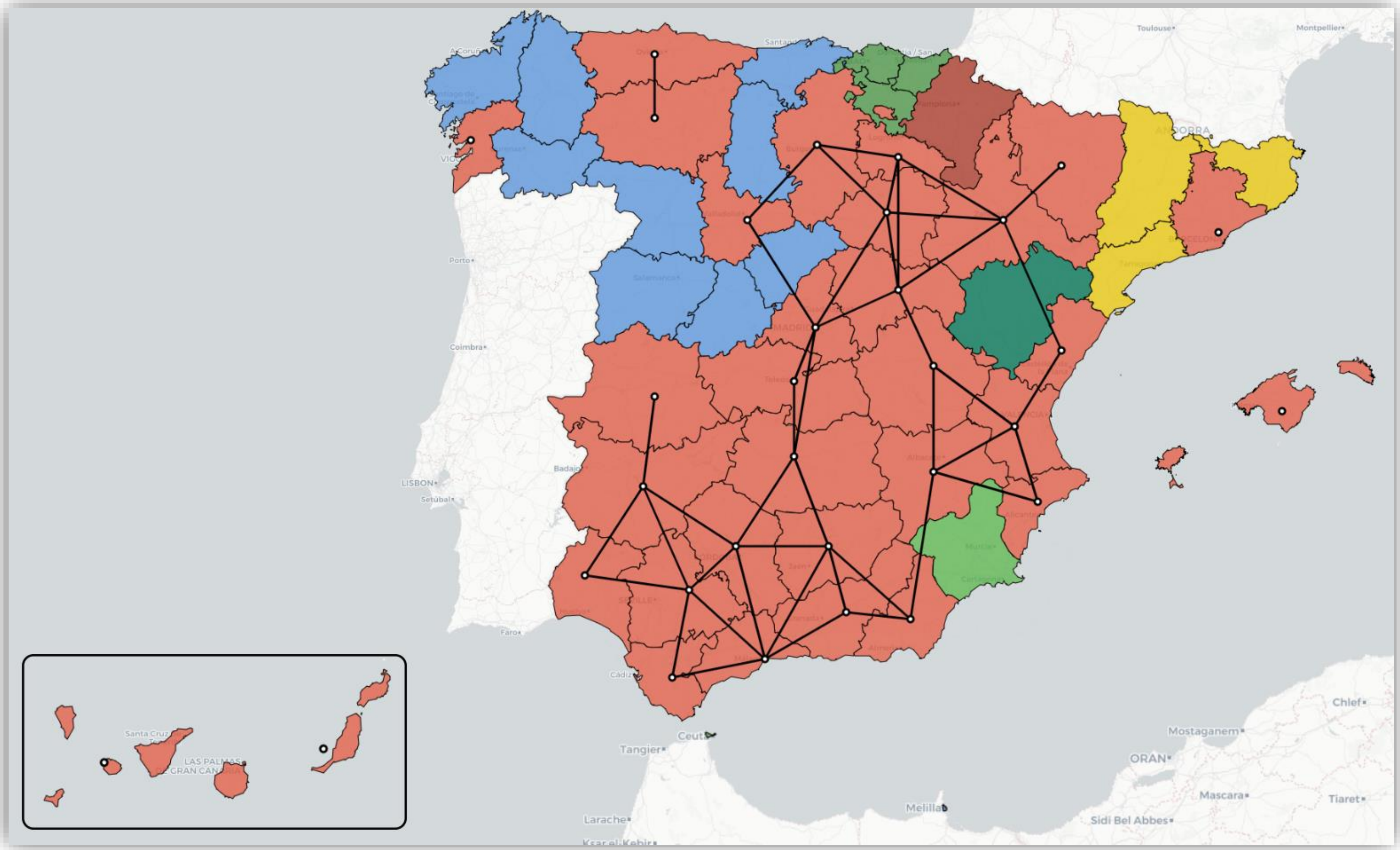
- Simulamos el procedimiento descrito en [4]
  - Usar homología persistente para determinar si hay distritos electorales donde se ha votado diferente respecto al resto de distritos que lo rodean.
  - Agujero → "Distrito isla".

- Resultados de las elecciones generales de noviembre de 2019:
  - Cada provincia con el color del partido con mayor número de votos
  - Coordenada representativa de cada provincia:
    - Wikipedia.
    - Random.
- Comprobar que los resultados obtenidos en [4] son robustos a la elección de la coordenada representativa del territorio.

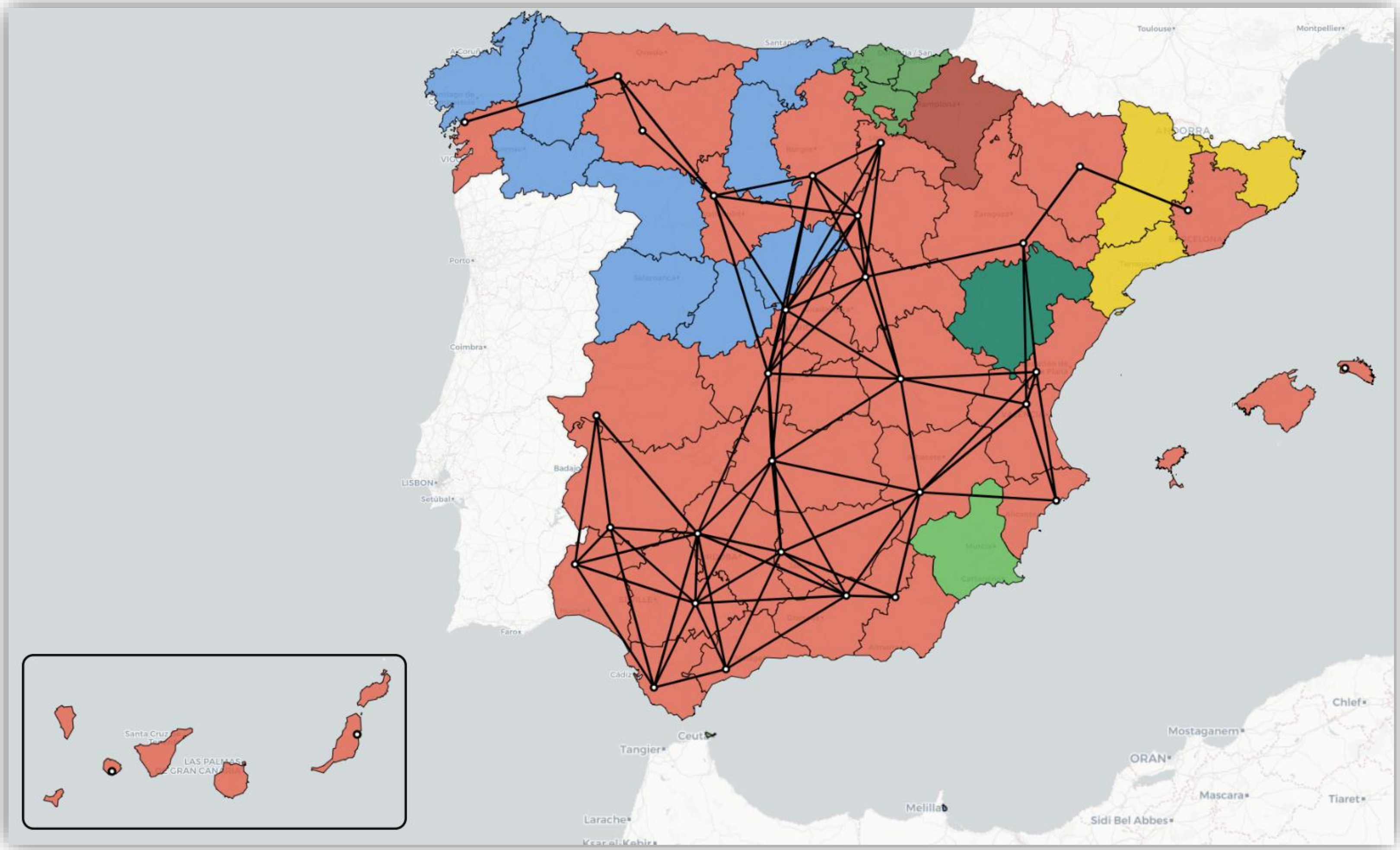


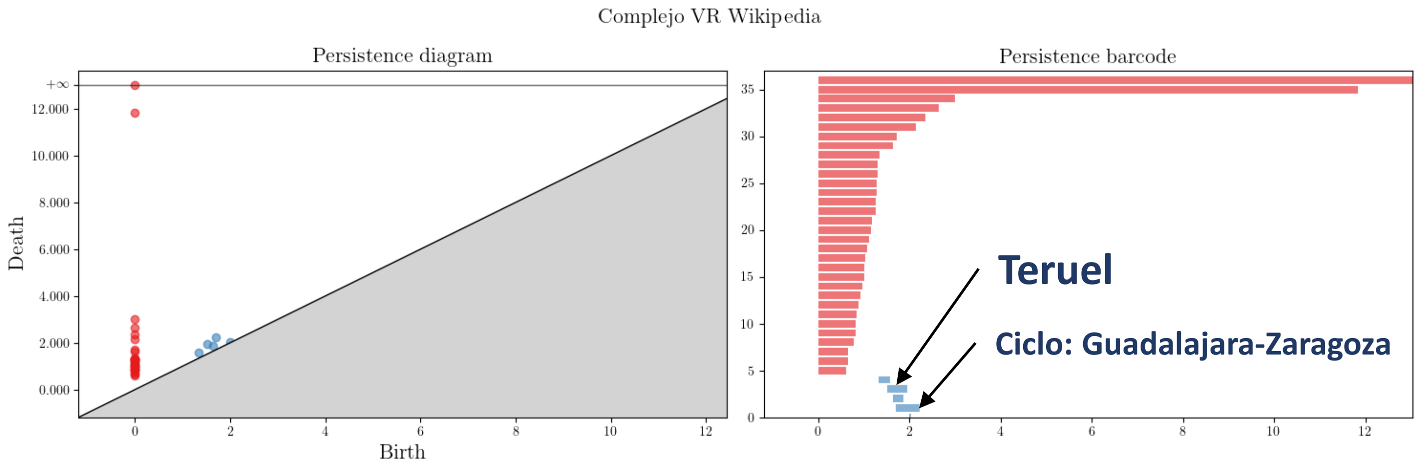
Condado de Napa, California
Fuente: [4]

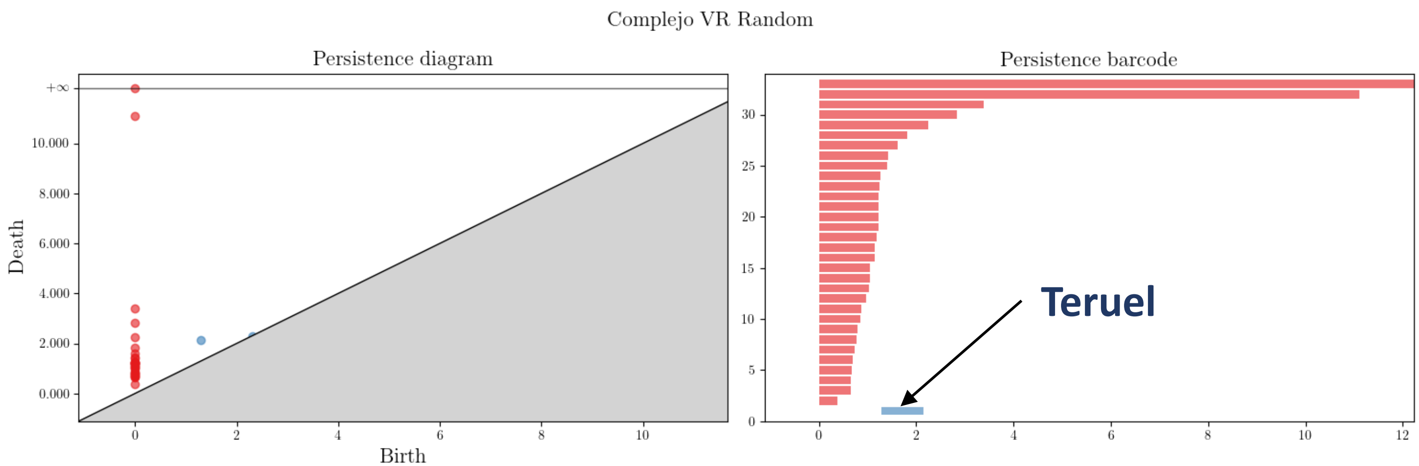
Complejo VR Wikipedia



Complejo VR Random







Dist. Hausdorff:

0.89583

Dist. Bottleneck

dim. 0: 0.7323

Dist. Bottleneck

dim. 1: 0.2682

Conclusiones

Conclusiones

- **Gran utilidad** de la homología persistente para el estudio cualitativo de los datos
  - Sin embargo, requiere interpretación de los resultados → **Posibilidad de inconsistencia en las interpretaciones.**
  - Dependen de la selección del complejo simplicial para obtener mejores resultados.
- **Amplias aplicaciones**:
  - Biología computacional: Estudio genético…
  - Neurociencia
  - Lingüística computacional.
  - Medicina: Oncología…
  - Machine Learning.
  - Etc.
- **Posibilidad de mejora** en el cálculo de ambas distancias:
  - Estudiar los algoritmos considerados *State of Art.*

Referencias

Referencias

- [1] H. Edelsbrunner and J. Harer, Computational Topology: An Introduction. American Mathematical Society, 01 2010.
- [2] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, "Stability of persistence diagrams," Discrete & Computational Geometry, vol. 37, no. 1, pp. 103–120, Jan 2007. [Online]. Available: https://doi.org/10.1007/s00454-006-1276-5
- [3] A. A. Taha and A. Hanbury, "An efficient algorithm for calculating the exact hausdorff distance," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 11, pp. 2153–2163, Nov. 2015. [Online]. Available: https://doi.org/10.1109/tpami.2015.2408351
- [4] M. Feng and M. A. Porter, "Persistent homology of geospatial data: A case study with voting," 2019.

GRACIAS

Alejandro García Castellanos, z17m008

Tutor: Héctor Barge Yañez