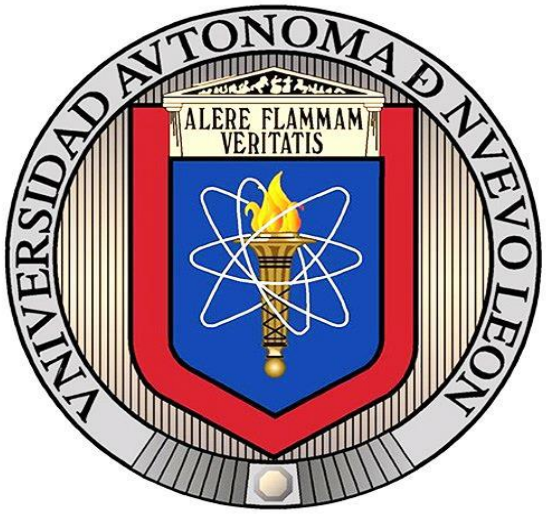
# UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
## FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS




**TESIS**

# ANÁLISIS DE PERSISTENCIA DE HOMOLOGÍA Y KRIGEADO APLICADO A DATOS DE PRECIPITACIÓN EN EL NORESTE DE MÉXICO

**POR**
**JUAN CARLOS SIFUENTES MONTAÑEZ**

**EN OPCIÓN AL GRADO DE MAESTRÍA EN CIENCIAS**
**CON ORIENTACIÓN EN MATEMÁTICAS**

**MARZO, 2020**

# UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

## FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

## CENTRO DE INVESTIGACIÓN EN CIENCIAS FÍSICO MATEMÁTICAS


**TESIS**

# ANÁLISIS DE PERSISTENCIA DE HOMOLOGÍA Y KRIGEADO APLICADO A DATOS DE PRECIPITACIÓN EN EL NORESTE DE MÉXICO

**POR**

**JUAN CARLOS SIFUENTES MONTAÑEZ**

**EN OPCIÓN AL GRADO DE MAESTRÍA EN CIENCIAS CON ORIENTACIÓN EN MATEMÁTICAS**

**SAN NICOLÁS DE LOS GARZA, NUEVO LEÓN, MÉXICO MARZO 2020**

# Universidad Autónoma de Nuevo León
# Facultad de Ciencias Físico-Matemáticas
# Centro de Investigación en Ciencias Físico Matemáticas

Los miembros del Comité de Tesis recomendamos que la Tesis “Análisis de persistencia de homología y krigeado aplicado a datos de precipitación en el noreste de México", realizada por el alumno Juan Carlos Sifuentes Montañez, con número de matrícula 1612858, sea aceptada para su defensa como opción al grado de Maestría en Ciencias con Orientación en Matemáticas.

El Comité de Tesis

Dr. Francisco Hernández Cabrera
Director Asesor

Dra. Lilia Alanís López
Co-Asesor

Dr. Francisco Javier Almaguer Martínez
Co-Asesor

Vo. Bo.

Dr. Omar Jorge Ibarra Rojas
Coordinador del Posgrado en Ciencias con Orientación en Matemáticas

San Nicolás de los Garza, Nuevo León, México 2 de Marzo de 2020

# Agradecimientos

Agradezco a mi familia, especialmente a: mi mamá Oralia Montañez de la Torre, que me ha apoyado más que nadie en toda mi vida y que gracias a ella estoy aquí; mi hermano, Luis Alberto Sifuentes Montañez; mi papá, Carlos Sifuentes Cardenas; mi tia, Blanca Ivon Montañez de la Torre y a mis abuelos que ya no están aquí. También quisiera agradecer a mis amigos, en especial a: Ashley Alitzel Garza Gatica, por ser de las mejores personas que conozco y una amiga muy querida; Ximena Dorely Medrano Gómez, que me ayudó a ampliar un poco mi visión del mundo; Jorge Iván Morales Ramírez, Adrian Mendoza Almaguer, Eduardo José Torres Gonzales, les agradezco que esten ahi cuando necesito un amigo; y a Evelyn Berlinda Garza Gatica, por el tiempo que pasamos juntos. Por último agradezco a la Universidad Autónoma de Nuevo León y al CONACYT.

# Índice general

Capítulo 1

# Introducción

En la actualidad con la ayuda de las computadoras, dispositivos móviles y satélites, se registra una gran cantidad de información. Con un teléfono celular se puede consultar la temperatura ambiental, la probabilidad de lluvia y mas datos atmosfericos por hora, con esto puede tenerse la impresión de tener toda la información de las variables en una región espacial. Sin embargo el medio ambiente es continuo y sus propiedades no pueden ser medidas en todos los puntos, además, las mediciones pueden contener ruido y esto se debe a que nuestras observaciones son el resultado de interacciones tan complejas que parecen aleatorias. Lo que podemos hacer es tratar de tener una muestra suficientemente grande y representativa de la realidad para poder aproximarla mediante estimaciones. Esto puede darnos la idea de usar algún método estadístico, pero en estadística, en general, se tiene una lista de datos medidos de una poblacion, donde se busca encontrar características que describan a la población y lo que nos interesa es conocer valores en puntos donde no tenemos información, para esto se usan métodos geoestadísticos. Entonces debemos utilizar métodos geoestadísticos para estimar los valores de las variables atmosféricas en cualquier punto del espacio.

En geoestadística se trata a los valores observados como si fueran una realización de una variable aleatoria y uno de sus métodos es el kriging ordinario. Este método toma un conjunto de puntos en un espacio, dode a cada punto se le asocia una variable y ademas se considera una métrica. el método vincula la variación observada en diferentes puntos con la separación entre estos y realiza estimación de valores no medidos que pertenecen a una vecindad de los puntos medidos y además nos da información de que tan bien se realiza la interpolación.

En general las bases de datos encontradas en geoestadistica presentan diversidad en las variables; un ejemplo pueden ser una región donde se registra la concentración de 4 minerales diferentes utilizando un muestreo de 100 puntos observados donde un análisis estadístico nos podria dar la concentración promedio en la zona de un mineral. Un análisis geoestadístico proporciona una estimación de la concentración en puntos no medidos y además nos dice cual seria la varianza de dicha estimación, pero en ambos casos

se utilizan pocos datos. Si tomamos las coordenadas del punto como variables espaciales junto con las mediciones de concentración de minerales, entonces podemos ver los datos como un conjunto de 100 puntos observados de un espacio de 6 dimensiones, 2 espaciales y 4 representadas por las concentraciones observadas. Así podemos analizar este conjunto de datos como elementos de un espacio topológico y con una distancia adecuada en este espacio, tenemos un espacio métrico, con esto podemos realizar una análisis que nos dará una clasificación de las propiedades invariantes topológicas, esto es llevado a cabo mediante el método de Análisis Topológico de Datos (TDA).

En este trabajo se utilizan dos métodos para el análisis de datos; el kriging ordinario y el análisis de persistencia de homología aplicados a datos atmosféricos en una región en el noreste de México.

## 1.1 Contexto

La geoestadística surgió en la industria minera, y fue formalizada por Matheron en [10], informacion de esta y sus métodos se puede consultar en los libros de [11] y el método base de la geoestadística, el kriging ordinario, puede consultarse en [12], además implementaciones de estos métodos se pueden encontrar en [13] y una guia [15].

Además del kriging ordinario existen diferentes metodos geoestadísticos de interpolación, en [6] se estudian el Kriging ordinario, el cokriging y el kriging con drift externo. Otros métodos de interpolación espacial que no son propios de la geoestadística son; regresión de componentes principales con corrección con residuos, regresion lineal múltiple y del inverso de la distancia [5] y en [7] se realiza un review de diversos métodos de interpolación espacial, entre ellos el kriging ordianrio y el método de las distancias inversas, en [8] y en [9] se realiza comparación entre diferentes métodos de interpolación espacial pero aún así, el kriging ordinario es el más utilizado. En las últimas décadas el Kriging ordinario se ha implementado para estimar y mapear la ubicación de sustancias potencialmente dañinas para el medio ambiente e identificar las fuentes como en [4].

El poder de cómputo de las plataformas tecnológicas actuales nos proporcionan gran ventaja para realizar análisis de datos a gran escala, por lo cual es necesario implementar algoritmos eficientes para resolver desafíos de bases de datos espaciales y lluvia [1]. se presentan algunos de estos aplicado al analisis de lluvia. Uno de los problemas a los que nos enfrentamos es que en general este tipo de base de datos no estaban planeadas y su elaboración e implementación, requiere la colaboración de entidades que se encuentran en diferentes estados o países, debido a esto es un problema encontrar bases de datos confiables. En [2] se implementa una base de datos relacional para la cuenca de México y

en 1997 la misión Tropical Rainfall Measuring de la nasa se lanzó para atacar el problema de la falta de bases de datos de lluvia, en [40] y [41], se puede encontrar más información en [39] .

Respecto al análisis temporal de lluvia en regiones de México se tiene [3] donde se utiliza el método de Kriging, tomando en cuenta la precipitación máxima y mínima para representar la precipitación diaria.

Por otra parte el análisis topológico de datos es un área reciente de las matemáticas aplicadas, tiene sus bases en la topología algebraica y geometría computacional, estos temas se pueden consultar en [25] y en [26] respectivamente y además algunas aplicaciones de la topología computacional se pueden encontrar en [20].

Una introducción a lo que es el análisis topológico y un poco de su historia se puede encontrar en [14] y en [16] además de eso, una guía más detallada se encuentra en [17] el cual cuenta con un artículo donde se presentan algunas implementaciones de los métodos.

El análisis topológico de datos se ha utilizado en varias áreas de investigación como en redes cooperativas en [21], en sistemas dinámicos en [22]. Otros ejemplos de aplicación del análisis topológico de datos se pueden encontrar en [23] y en [27]. En [24] se presenta como se realizan análisis estadísticos en el análisis topológico de datos.

## 1.2 Problemática

En la vida diaria dependiendo de la región donde nos encontremos, una de las acciones que es conveniente realizar antes de salir de casa, es revisar el pronóstico del tiempo, más con el acceso que nos dan a este los teléfonos celulares, esto es debido a que los fenómenos atmosféricos pueden causar problemas de tránsito, inundaciones, o podemos encontrarnos con cambios de temperatura bruscos. Además de esto las regiones dependen de estos datos para recolectar agua para beber o para la agricultura. Por esto es importante tener buenas bases de datos que registren los fenómenos atmosféricos y además poder analizar correctamente la información que estos datos nos presentan. Con esto en mente, el problema que busca enfrentar este trabajo es el siguiente:

**Problema:** Conocer los cambios en las variables climáticas del norte de México durante periodos establecidos. Se tomará como zona de estudio la región comprendida entre las longitudes: -104 y - 98.5 y las latitudes: 22 y 27, y el periodo de tiempo: enero del 2015 a enero del 2016, debido a que en este periodo de tiempo ocurrió el huracán patricia que afectó zonas de el área elegida para estudio.

## 1.3 Hipótesis

Uno de los métodos base en la geoestadístico es el método del kriging ordinario, este nos proporciona una interpolación de datos espaciales pero no toma en cuenta múltiples instantes de tiempo, por otro lado el análisis de persistencia de homología es un método más nuevo para analizar datos, el cual puede trabajar sobre un espacio de n dimensiones y nos presenta datos del espacio del que proviene la muestra, además nos da métodos para medir distancia entre estos espacios, aunque no da información tan puntual como los métodos geoestadísticos. Esto nos lleva a lo siguiente:

**Hipótesis:** El análisis por kriging ordinario y la persistencia de homología de estructuras topológicas de datos atmosféricos pueden establecer los cambios climáticos espacio-temporales.

## 1.4 Objetivos

En este trabajo se estudian métodos de geoestadística y TDA para aplicarlos a datos atmosféricos en el norte de México, en particular se estudiarán los datos de precipitación. Esto nos da:

**Objetivo General** identificar las variaciones espacio-temporales en la precipitación en el norte de México utilizando geoestadística y TDA entre las longitudes: -104 y - 98.5 y las latitudes entre: 22° y 27° en el periodo de tiempo: enero del 2015 a enero del 2016

Para esto se tienen los siguientes objetivos particulares:

- Búsqueda de bases de datos de precipitación de la región entre las longitudes: -104 y - 98.5 y las latitudes: 22 y 27 y el periodo de tiempo: enero del 2015 a enero del 2016 y realizar preprocesamiento de los datos de ser necesario.
- Aplicar el kriging ordinario para establecer la función discreta que vincula los puntos de la región que se estudia con la precipitación en dichos puntos.
- Utilizar la homología persistente para caracterizar las estructuras topológicas de la función discreta que vincula un punto en el área que se estudia con la cantidad de lluvia registrada en ese punto.
- Observar los cambios en las propiedades topológicas del espacio generado por el conjunto de puntos con la forma (longitud, latitud, cantidad de lluvia registrada)

en los meses de enero del 2015 a enero del 2016 mediante un análisis de persistencia de homología.

## 1.5 Contenido

En el capítulo 2 se exponen algunas bases de la geoestadística y se desarrollan el kriging ordinario y la semivarianza, despues se da una descripción del tipo de datos a los que se aplican los métodos de geoestadística. Además, se proporciona una breve introducción del análisis topológico de datos; complejos topológicos, homología persistente y los métodos más usuales para construir complejos simpliciales, los cuales serán la estructura del espacio.

En el capítulo 3 se presentan la metodología con la que se abordara el problema y se explica el proceso realizado.

En el capítulo 4 se presentan los resultados obtenidos en forma de mapas de calor para el kriging ordinario, de diagramas de vida y muerte para el análisis, persistencia de homología y se indica la distancia entre las gráficas de vida y muerte para el análisis temporal.

En el capítulo 5 se encuentran las conclusiones obtenidas y se plantea un posible trabajo futuro.

Capítulo 2

# Marco Teórico

## 2.1 Geoestadística

Geoestadística es un conjunto de técnicas estadísticas basadas en la teoría de los procesos estocásticos, establecida en las ciencias de la tierra como la teoría de las variables regionalizadas, debido a Matheron(1963, 1965) siguiendo el trabajo empírico de Krige(1951).

La geoestadística se desarrolló gracias a la industria minera. Su primera aplicación fue en minas de oro por Daniel Krige, donde la idea era saber si la extracción de material era justificada en un punto del planeta dado que se conoce la concentración de dicho material en puntos cercanos. Entonces, se busca conocer la concentración de un mineral en puntos cercanos donde no se han realizado observaciones y con base en esto se tiene la posibilidad de hacer mapas que describen propiedades de interés en ciertas áreas donde solo se tiene una muestra de valores.

Existen otros métodos de interpolación espacial, como el método de las distancias inversas, pero ningún modelo puede describir por completo la realidad y cualquier técnica de interpolación producirá resultados con cierto grado de error, he aquí la ventaja de los métodos geoestadísticos, donde junto con el valor estimado, también nos da el error de las estimaciones y además, estos métodos son en principio, no sesgados.

En el enfoque clásico el muestreo aleatorio suele usarse para asegurar que los estimadores sean no sesgados y con varianza conocida. En el caso de variables atmosféricas se requiere de un gran número de puntos espaciales, sin embargo, cuando se utilizan métodos geoestadísticos es suficiente con un número reducido de puntos. Además los métodos geoestadísticos están enfocados en estimar propiedades en puntos no muestreados; en lugar de estimar parámetros de la población.

La geoestadística puede ser definida como el estudio de fenómenos regionalizados, los cuales, son descritos por variables regionalizadas.

## 2.2 Variables regionalizadas

El término variables regionalizadas fue concebido por G. Matheron 1963 para enfatizar el hecho de que las variables geológicas presentan variación en pequeños intervalos, pero a gran escala presentan tendencia. Una forma de imaginar esto, es como pequeños picos que representan la variación sobre una superficie fija. Dado que no tenemos esta superficie, representamos la muestra como si se tratara de un proceso aleatorio descrito por la ecuación:

$$Z(\mathbf{x}) = \mu + \epsilon(\mathbf{x}) \tag{2.1}$$

donde se tiene la siguiente notación:

- $\mu$ es la media del proceso que representa la parte determinística
- $\epsilon(\mathbf{x})$ es una variable aleatoria con media cero y cuya covarianza $C(x, x+h)$ solo depende de la distancia entre puntos $\mathbf{h}$

entonces tenemos:

$$C(h) = C(x, x+h) = E\left[\epsilon(\mathbf{x})\epsilon(\mathbf{x}+\mathbf{h})\right] \tag{2.2}$$

donde $\mathbf{h}$ es llamada lag.

Ahora, dado que tenemos una muestra de valores observados en diferentes puntos de una superficie, a la posición de cada valor observado se le representara como $x_i$ y a su valor asociado se le denotará por $z(x_i)$ que es una realización de una variable regionalizada $Z(x)$, donde el promedio $m(x_i)$ en el punto $x_i$ es llamado drift.

En estadística es común asumir que los procesos que se estudian son estacionarios. Esto es, que las distribuciones de los procesos son invariantes bajo traslaciones. Estacionario en sentido estricto requiere que todos los momentos sean invariantes bajo traslaciones. Se requerirá solo que los dos primeros momentos sean invariantes bajo traslaciones. Esto es llamado estacionariedad en sentido débil o estacionariedad de segundo orden. En el contexto espacial, requerimos:

1. que el valor esperado de la función $Z(x)$ sea constante para todos los puntos x, $E[Z(x)] = m(x) = m$

2. que la covarianza entre los puntos x, x + h sólo depende de h.

En la realidad las condiciones anteriores no siempre se cumplen, como cuando hay una tendencia en la superficie, en este caso el valor de $m$ no se puede asumir como constante. Para esto usaremos la hipótesis de que la función aleatoria es intrínseca. Una función aleatoria se llama **intrínseca** si la esperanza matemática existe y no depende del punto, esto es:

- $E(z) = m$
- para cualquier vector h, el incremento $Z(x+h) - Z(x)$ tiene varianza finita y es independiente del punto x, entonces, esta diferencia sólo depende de h y se puede expresar $var(Z(x+h) - Z(x)) = 2\gamma(h)$.

La función $\gamma(h)$ es llamado **semivariograma** o por brevedad, variograma. La correlación espacial de variables aleatorias intrínsecas es caracterizada por la función semivariograma, esta fue definida por Matheron 1967 en [10] como:

$$\gamma(h) = \frac{1}{2}Var(Z(x+h) - Z(x)) \tag{2.3}$$

pero como la media de $Z(x+h) - Z(x)$ es cero entonces tenemos la siguiente simplificación:

$$\gamma(h) = \frac{1}{2}E(Z(x+h) - Z(x))^2$$

$$\hat{\gamma}(h) = \frac{1}{2N}\sum_{i=1}^{N}(Z(x_i+h) - Z(x_i))^2$$

La razón a la que crece $\gamma(h)$ es un indicador de la razón a la que decrece la influencia de la muestra al incrementarse la distancia, esto es, al alejarnos de un punto donde conocemos su variable asociada, su influencia decrece a la razón a la que crece $\gamma(h)$. Tenemos un valor que limita la influencia, el cual llamamos sill, que es la varianza de la población, ya que, en el caso en el que no existe correlación entre $Z(x+h)$ y $Z(x)$ tenemos:

$$\gamma(h) = \frac{1}{2}Var(Z(x+h) - Z(x))$$

$$\gamma(h) = \frac{1}{2}(Var(Z(x+h)) + Var(Z(x))) = \frac{2\sigma^2}{2} = \sigma^2$$

Estas definiciones se pueden encontrar en [11]

## 2.3 Estacionariedad

Estacionariedad es una suposición que nos permite tratar los datos como si tuvieran el mismo grado de variación sobre una región de interés.

**Segundo orden estacionariedad** La función aleatoria $\{Z(x), x \in X\}$ Es dicho que es estacionario de segundo orden o débilmente estacionario o simplemente estacionario en amplio sentido, si tiene segundo momento y verifica que:

- La esperanza existe y es constante, no depende de la x:

$$E(Z(x)) = \mu(x) = \mu$$

- la covarianza existe para cada par de valores aleatorios. $Z(x)$ y $Z(x+h)$ y solo depende de el vector h que une los puntos x y $x+h$

$$C(Z(x), Z(x+h)) = C(h)$$

Como la función de covarianza C(h) de una función aleatoria estacionaria de segundo orden es una función de h entonces la varianza de la función aleatoria, existe y es constante:

$$V(Z(x)) = C(0) = \sigma^2$$

Y esta se relaciona con el semivariograma de la siguiente forma:

$$\begin{aligned}
\gamma(h) &= \frac{1}{2}V(Z(x+h) - Z(x)) \\
&= \frac{1}{2}(V(Z(x+h)) + V(Z(x)) + 2C(Z(x+h), Z(x))) \\
&= \frac{1}{2}(C(0) + C(0) + 2C(h)) \\
&= C(0) - C(h)
\end{aligned}$$

La **covarianza** de una función aleatoria es una función no aleatoria de $x_i$ y $x_j$, tal que esta coincide con la covarianza entre la función aleatoria en esos puntos $(Z(x_i), Z(x_j))$:

$$C(x_i, x_j) = C(Z(x_i), Z(x_j)) = E((Z(x_i) - \mu(x_i))(Z(x_j) - \mu(x_j)))$$

Estacionariedad indica cierto grado de homogeneidad. Para hacer inferencias de forma consistente, necesitamos muchas realizaciones. Para resolver esto se adopta la hipótesis de estacionariedad o de homogeneidad espacial, la idea es reemplazar las repeticiones de la realización de la función aleatoria con repeticiones en el espacio, entonces, las observaciones de valores en diferentes ubicaciones tienen características y entonces pueden ser consideradas como realizaciones de la misma función aleatoria, esta hipótesis significa que la probabilidad de la función aleatoria es invariante bajo traslaciones.

## 2.4 Kriging

El krigeado es una técnica de interpolación espacial, es llamado kriging en honor del ingeniero sudafricano Daniel G. Krige. La idea del kriging es que, dado un conjunto de puntos $x_1, x_2, \cdots, x_n$ en un espacio, donde cada uno de estos puntos tiene un valor asociado $z(x_1), z(x_2), \cdots, z(x_n)$, y en base a esto podemos estimar el valor $\hat{z}(x_0)$ en un punto $x_0$ como una suma ponderada de los valores $z(x_i)$ dados. Es decir, este método de interpolación espacial es usado cuando tenemos:

- un espacio donde a cada punto $x_i$ corresponde una valor $z(x_i)$.
- $z(x_i)$ es una variable aleatoria que tiene media constante.
- la varianza de $z(X)$ es aleatoria y depende solamente de la distancia entre los puntos.

Una vez que tenemos esto, buscamos estimar el valor de $z(x_0)$ mediante una combinación lineal de los valores $z(x_i)$, en base a esta idea tenemos:

$$\hat{z}(x_0) = \lambda_1 z(x_1) + \lambda_2 z(x_2) + \cdots + \lambda_n z(x_n) \tag{2.4}$$

Un ejemplo de estas ideas se presenta en la figura 2.1 para el caso de una dimensión, donde nuestro interés es estimar $\hat{z}(x_0)$ con los demás puntos y en la figura 2.2 se presenta esta idea para el caso de dos dimensiones.

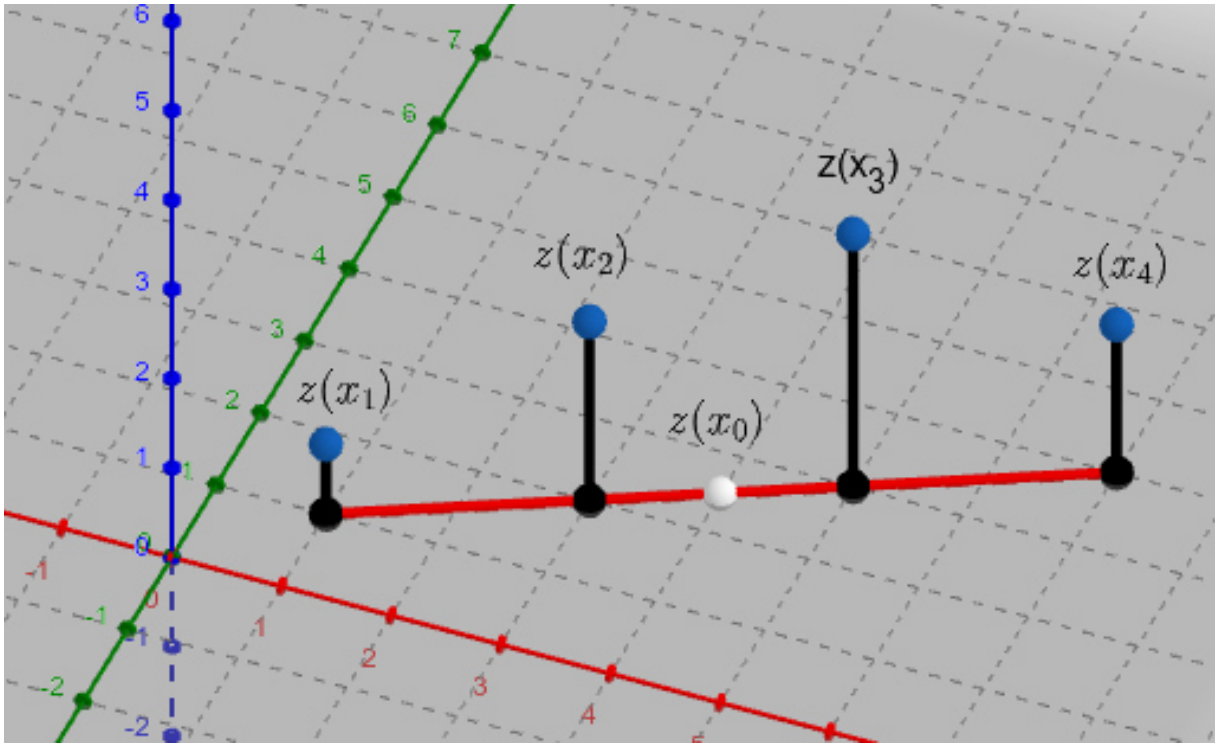



Figura 2.1: Ejemplo de interpolación por kriging en una dimensión espacial

Como se puede observar en la ecuación 2.4 lo que ocupamos para estimar $\hat{z}(x_0)$ son los $\lambda_i$ , estos tienen que ser tales se cumpla lo siguiente:

1. El estimador es insesgado, esto es $E\left[\hat{z}(x_0)\right] = E\left[z(x_0)\right]$
2. La varianza es mínima

Que el estimador sea insesgado, junto con la suposición de que $E\left[\hat{z}(x_i)\right] = \mu$ nos da:

$$\begin{aligned} E\left[\hat{z}(x_0)\right] &= E\left[z(x_0)\right] \\ E\left[\sum_{i=0}^{n} \lambda_i z(x_i)\right] &= \mu \\ \sum_{i=0}^{n} \lambda_i E\left[z(x_i)\right] &= \mu \\ \sum_{i=0}^{n} \lambda_i \mu &= \mu \\ \sum_{i=0}^{n} \lambda_i &= 1 \end{aligned}$$

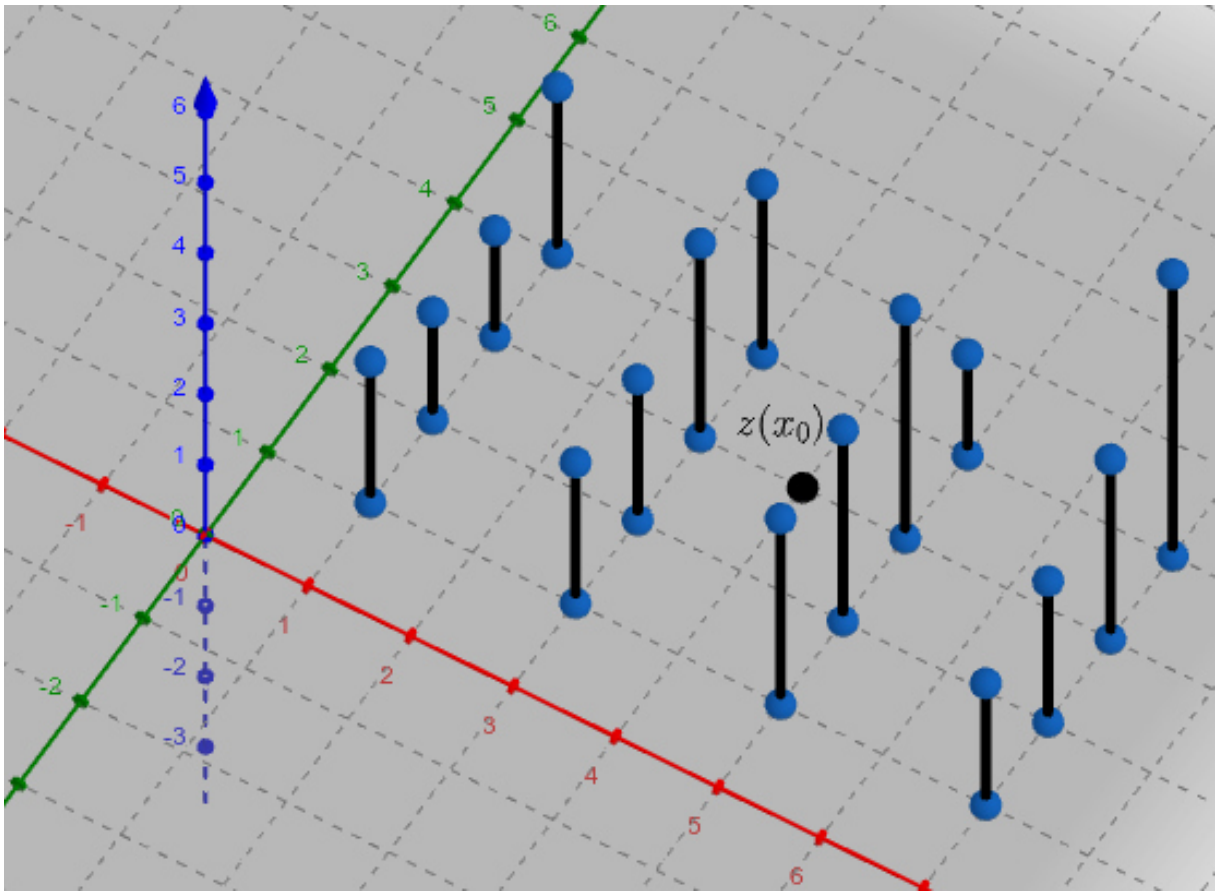



Figura 2.2: Ejemplo de interpolación por kriging en dos dimensiónes espaciales

Entonces tenemos que los $\lambda_i$ tienen que cumplir $\sum_{i=0}^{n} \lambda_i = 1$

Como el kriging es el mejor estimador insesgado con varianza mínima, primero expresamos la varianza como:

$$\begin{aligned} V(\hat{Z}(x_0) - Z(x_0)) &= E(\hat{Z}(x_0) - Z(x_0)^2) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j E(Z(x_i)Z(x_j)) - 2\sum_{i=1}^{n} \lambda_i E(Z(x_i)Z(x_0)) \\ &\quad + E(Z(x_0)^2) \end{aligned}$$

donde, como tenemos que:

$$C(x_i - x_j) = E(Z(x_i)Z(x_j)) + E(Z(x_i))E(Z(x_j)) = E(Z(x_i)Z(x_j)) + \mu^2$$

ya que $E(Z(x)) = \mu$, tenemos lo siguiente:

$$\begin{aligned} V(\hat{Z}(x_0) - Z(x_0)) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j C(x_i - x_j) - 2\sum_{i=1}^{n} \lambda_i C(x_i - x_0) + C(0) \\ &\quad + \mu^2\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j - 2\sum_{i=1}^{n} \lambda_i + 1\right) \end{aligned}$$

Por la condición de que el estimador es insesgado ( $\sum_{i=1}^{n} \lambda_i = 1$ ) tenemos:

$$V(\hat{Z}(x_0) - Z(x_0)) = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j C(x_i - x_j) - 2\sum_{i=1}^{n} \lambda_i C(x_i - x_0) + C(0) \quad (2.5)$$

Entonces, ahora la idea es resolver el problema de optimización:

$$\begin{aligned} min \quad & V(\hat{Z}(x_0) - Z(x_0)) \\ s.a. \quad & \sum_{i=1}^{n} \lambda_i = 1 \end{aligned}$$

Esto es resuelto por el método de los multiplicadores de lagrange, donde buscamos minimizar:

$$V(\hat{Z}(x_0) - Z(x_0)) - \alpha(\sum_{i=i}^{n} \lambda_i - 1)$$

Lo cual es:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j C(x_i - x_j) - 2\sum_{i=1}^{n} \lambda_i C(x_i - x_0) + C(0) - \alpha(\sum_{i=i}^{n} \lambda_i - 1)$$

Derivando con respecto a $\lambda_i$ tenemos:

$$\sum_{j=1}^{n} \lambda_j C(x_i - x_j) - 2\sum_{i=1}^{n} \lambda_i C(x_i - x_0) + C(0) - \alpha(\sum_{i=i}^{n} \lambda_i - 1)$$

$$\begin{aligned}
&\lambda_1\lambda_1 C(x_1 - x_1) + \lambda_1\lambda_2 C(x_1 - x_2) + \cdots + \lambda_1\lambda_n C(x_1 - x_n)\\
&+\lambda_2\lambda_1 C(x_2 - x_1) + \lambda_2\lambda_2 C(x_2 - x_2) + \cdots + \lambda_2\lambda_n C(x_2 - x_n)\\
&+\vdots\\
&+\lambda_n\lambda_1 C(x_n - x_1) + \lambda_n\lambda_2 C(x_n - x_2) + \cdots + \lambda_n\lambda_n C(x_n - x_n)\\
&-2\lambda_1 C(x_1 - x_0) + -2\lambda_2 C(x_2 - x_0) + \cdots + -2\lambda_n C(x_n - x_0) + C(0)\\
&+\alpha(\sum_{i=i}^{n} \lambda_i - 1)
\end{aligned}$$

$$\begin{aligned}
&2\lambda_1 C(x_1 - x_1) + \lambda_2 C(x_1 - x_2) + \cdots + \lambda_n C(x_1 - x_n)\\
&+\lambda_1 C(x_2 - x_1) + 0 + \cdots + 0\\
&+\vdots\\
&+\lambda_n C(x_n - x_1) + 0 + \cdots + 0\\
&-2C(x_1 - x_0) + 0 + \cdots + 0 + 0\\
&+\alpha
\end{aligned}$$

esto lo igualamos a cero:

$$2\lambda_1 C(x_1 - x_1) + 2\lambda_2 C(x_1 - x_2) + \cdots + 2\lambda_n C(x_1 - x_n) + -2C(x_1 - x_0) + \alpha = 0$$

entonces tenemos:

$$\sum_{j=1}^{n} \lambda_j C(x_1 - x_j) - \alpha = C(x_1 - x_0)$$

esto lo repetimos para cada $\lambda_i$ con $i \in \{1, 2, \cdots, n\}$ además derivamos con respecto a $\alpha$ e igualando a cero tenemos:

$$\sum_{i=i}^{n} \lambda_i = 1$$

lo que nos da el sistema de ecuaciones:

$$\begin{aligned}
\sum_{j=1}^{n} \lambda_j C(x_1 - x_j) - \alpha &= C(x_1 - x_0) \\
\sum_{j=1}^{n} \lambda_j C(x_2 - x_j) - \alpha &= C(x_2 - x_0) \\
&\vdots \\
\sum_{j=1}^{n} \lambda_j C(x_n - x_j) - \alpha &= C(x_i - x_0) \\
\sum_{i=i}^{n} \lambda_i &= 1
\end{aligned}$$

y en el caso de estacionariedad de segundo orden tenemos que:

$$C(x_i - x_j) = C(0) - \gamma(x_i - x_j) \tag{2.6}$$

entonces $\sum_{j=1}^{n} \lambda_j C(x_1 - x_j) - \alpha = C(x_1 - x_0)$ se hace:

$$\begin{aligned}
\sum_{j=1}^{n} \lambda_j (C(0) - \gamma(x_1 - x_j)) - \alpha &= C(0) - \gamma(x_1 - x_0) \\
C(0) \sum_{j=1}^{n} \lambda_j - \sum_{j=1}^{n} \lambda_j \gamma(x_1 - x_j) - \alpha &= C(0) - \gamma(x_1 - x_0)
\end{aligned}$$

como $\sum_{j=1}^{n} \lambda_j = 1$ tenemos:

$$C(0) - \sum_{j=1}^{n} \lambda_j \gamma(x_1 - x_j) - \alpha = C(0) - \gamma(x_1 - x_0) \tag{2.7}$$

$$\sum_{j=1}^{n} \lambda_j \gamma(x_1 - x_j) + \alpha = \gamma(x_1 - x_0) \tag{2.8}$$

Con esto tenemos:

$$\begin{aligned}
\sum_{j=1}^{n} \lambda_j \gamma(x_1 - x_j) + \alpha &= \gamma(x_1 - x_0) \\
\sum_{j=1}^{n} \lambda_j \gamma(x_2 - x_j) + \alpha &= \gamma(x_2 - x_0) \\
&\vdots \\
\sum_{j=1}^{n} \lambda_j \gamma(x_n - x_j) + \alpha &= \gamma(x_i - x_0) \\
\sum_{i=i}^{n} \lambda_i &= 1
\end{aligned}$$

reescribiendo $\gamma(x_i - x_j)$ como $\gamma_{ij}$ ya que solo depende de la distancia entre $x_i$ y $x_j$, reescribiendo esto de forma matricial tenemos:

$$\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n} & 1 \\ \gamma_{21} & \gamma_{22} & \ddots & \gamma_{2n} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{nn} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \alpha \end{bmatrix} = \begin{bmatrix} \gamma_{01} \\ \gamma_{02} \\ \vdots \\ \gamma_{0n} \\ 1 \end{bmatrix} \tag{2.9}$$

donde los $\gamma_{ij}$ representa el valor del variograma entre los puntos $i$ y $j$ y $\alpha$ es un valor agregado mediante el método de los multiplicadores de lagrange

Ahora si continuamos aplicando la ecuación 2.4 en la ecuación 2.5 tenemos:

$$
\begin{aligned}
V(\hat{Z}(x_0) - Z(x_0)) &= \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j(C(0) - \gamma(x_i - x_j)) - 2\sum_{i=1}^{n}\lambda_i(C(0) - \gamma(x_i - x_0)) \\
&\quad +C(0) \\
&= \sum_{i=1}^{n}\lambda_i(\sum_{j=1}^{n}\lambda_j C(0) - \sum_{j=1}^{n}\lambda_j\gamma(x_i - x_j)) - 2\sum_{i=1}^{n}\lambda_i C(0) \\
&\quad +2\sum_{i=1}^{n}\lambda_i\gamma(x_i - x_0)) + C(0) \\
&= \sum_{i=1}^{n}\lambda_i C(0) - \sum_{i=1}^{n}\lambda_i\sum_{j=1}^{n}\lambda_j\gamma(x_i - x_j)) - 2C(0) \\
&\quad +2\sum_{i=1}^{n}\lambda_i\gamma(x_i - x_0)) + C(0) \\
&= C(0) - \sum_{i=1}^{n}\lambda_i\sum_{j=1}^{n}\lambda_j\gamma(x_i - x_j)) - 2C(0) \\
&\quad +2\sum_{i=1}^{n}\lambda_i\gamma(x_i - x_0)) + C(0) \\
&= -\sum_{i=1}^{n}\lambda_i\sum_{j=1}^{n}\lambda_j\gamma(x_i - x_j)) + 2\sum_{i=1}^{n}\lambda_i\gamma(x_i - x_0))
\end{aligned}
$$

ahora, aplicando 2.8 tenemos:

$$
\begin{aligned}
V(\hat{Z}(x_0) - Z(x_0)) &= -\sum_{i=1}^{n}\lambda_i(\gamma(x_i - x_0) - \alpha) + 2\sum_{i=1}^{n}\lambda_i\gamma(x_i - x_0)) \\
&= -\sum_{i=1}^{n}\lambda_i\gamma(x_i - x_0) + \sum_{i=1}^{n}\lambda_i\alpha + 2\sum_{i=1}^{n}\lambda_i\gamma(x_i - x_0) \\
&= -\sum_{i=1}^{n}\lambda_i\gamma(x_i - x_0) + \alpha + 2\sum_{i=1}^{n}\lambda_i\gamma(x_i - x_0) \\
&= \alpha + \sum_{i=1}^{n}\lambda_i\gamma(x_i - x_0)
\end{aligned}
$$

Entonces, la ecuación del kriging nos da el estimador insesgado de varianza mínima, pero además nos da el valor de esa varianza, este lo reescribimos como $\sigma^2_{ok}(s_0)$ esto nos da la ecuación:

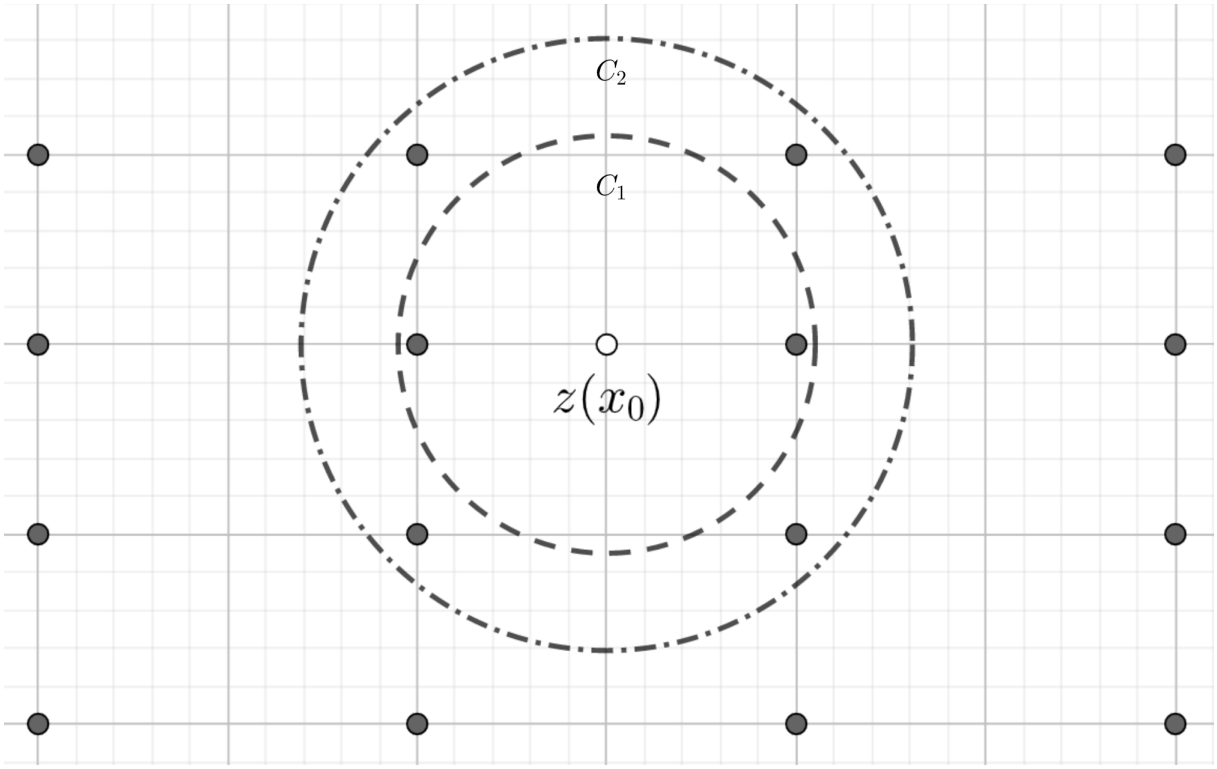



Figura 2.3: Efecto según distancia

$$\sigma^2_{ok}(s_0) = \sum_{i=1}^{n} \lambda_i \gamma(x_i - x_0) + \alpha \tag{2.10}$$

Ahora, para calcular esto solo ocupamos los valores de $\gamma$, que son los valores de la semivarianza

## 2.5 Semivarianza

El variograma o variograma teórico, es una función que como entrada tiene la distancia $h$ entre 2 puntos del espacio y como salida tiene la varianza que se espera observar entre todos los puntos que se encuentran a la distancia $h$. Se suele llamar también variograma a la gráfica del variograma (ver figura 2.8) y que describe la variación con respecto a la cercanía de los puntos, esto se ve en la figura 2.3, los puntos dentro de $C_1$ afectan de forma diferente a $z(x_0)$ que los que están en $C_2$ pero no en $C_1$.

Para obtener el variograma teórico, primero creamos un variograma experimental, este se hace tomando a los puntos que se encuentran a una distancia h y se realiza la

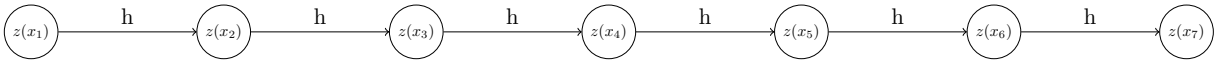



Figura 2.4: Puntos a distancia h

| **Inicio** | **Final** | **Diferencia** |
|---|---|---|
| $z(x_1)$ | $z(x_2)$ | $z(x_2) - z(x_1)$ |
| $z(x_2)$ | $z(x_3)$ | $z(x_3) - z(x_2)$ |
| $z(x_3)$ | $z(x_4)$ | $z(x_4) - z(x_3)$ |
| $z(x_4)$ | $z(x_5)$ | $z(x_5) - z(x_4)$ |
| $z(x_5)$ | $z(x_6)$ | $z(x_6) - z(x_5)$ |
| $z(x_6)$ | $z(x_7)$ | $z(x_7) - z(x_6)$ |

Tabla 2.1: Tabla para puntos a distancia h

tabla 2.1 donde en la última columna se toma como si fuera una muestra y se calcula la varianza, esto nos dará el valor de $\gamma_h$.

Esto se repite para los puntos que están a una distancia $2h$, esta idea se resume en la ecuación de Matheron:

$$\gamma(h) = \frac{1}{2N}\sum_{i=1}^{N}(z(x_i + h) - z(x_i))^2 \tag{2.11}$$

Y así obtenemos el valor para el variograma en $2h$, de la misma forma, calculamos el valor para todas las varianzas a intervalos de distancia h.

Y así obtenemos una gráfica como 2.7, que es el variograma experimental.

Con base a esta gráfica obtenemos un modelo téorico, este modelo en general se busca que sea simple y los modelos usuales suelen incluir los siguientes parámetros

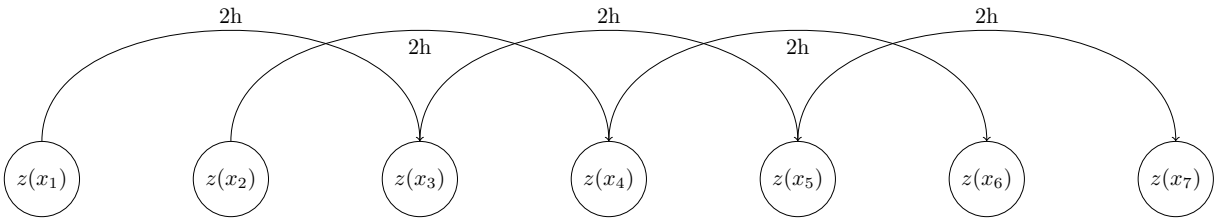



Figura 2.5: Puntos a distancia 2h

| **Inicio** | **Final** | **Diferencia** |
|---|---|---|
| $z(x_1)$ | $z(x_3)$ | $z(x_3) - z(x_1)$ |
| $z(x_2)$ | $z(x_3)$ | $z(x_4) - z(x_2)$ |
| $z(x_3)$ | $z(x_4)$ | $z(x_5) - z(x_3)$ |
| $z(x_4)$ | $z(x_5)$ | $z(x_6) - z(x_4)$ |
| $z(x_5)$ | $z(x_6)$ | $z(x_7) - z(x_5)$ |

Tabla 2.2: Tabla para puntos a distancia 2h

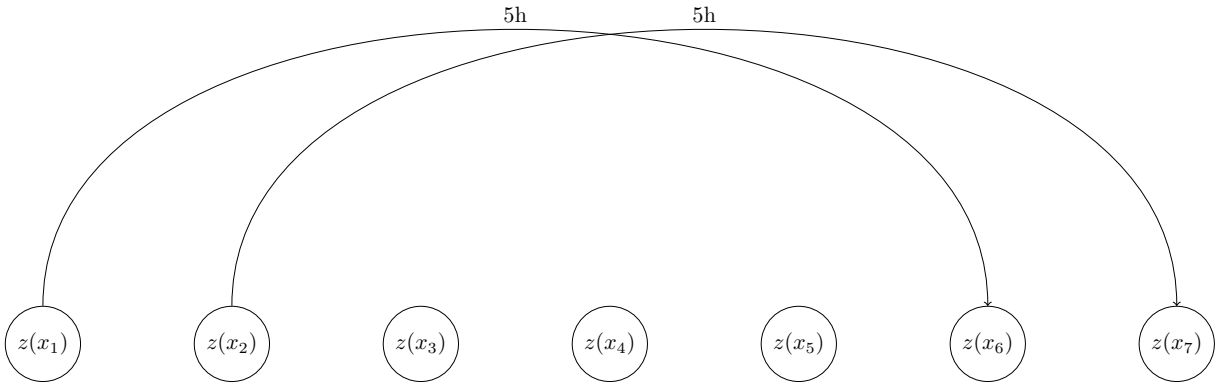



Figura 2.6: Puntos a distancia 5h

| **Inicio** | **Final** | **Diferencia** |
|---|---|---|
| $z(x_1)$ | $z(x_6)$ | $z(x_6) - z(x_1)$ |
| $z(x_2)$ | $z(x_7)$ | $z(x_7) - z(x_2)$ |

Tabla 2.3: Tabla para puntos a distancia 5h

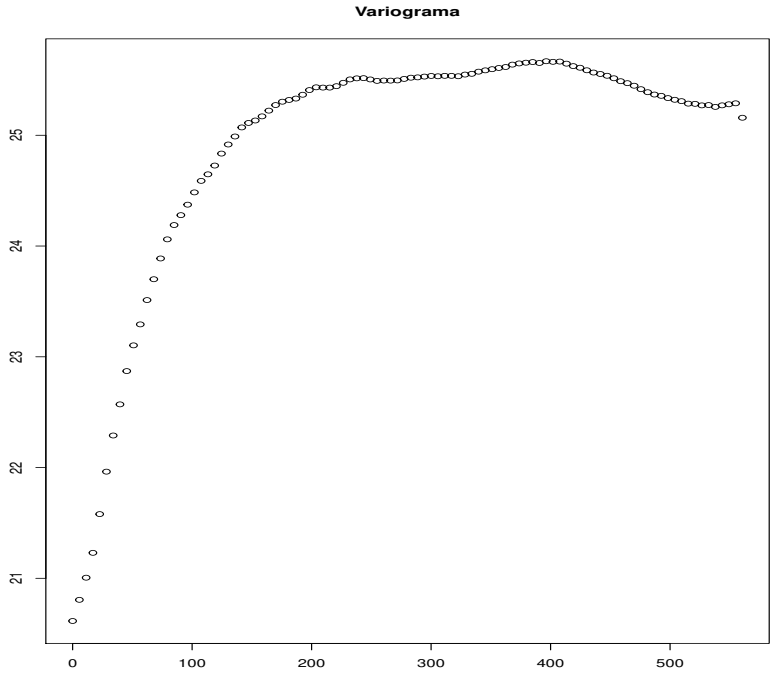



Figura 2.7: Variograma experimental

**nugget ($c_0$).** Es el valor del variograma cuando $h = 0$ esto es el valor cuando tomamos distancias muy pequeñas entre mediciones, esto suele ser por errores de medición o en los instrumentos de medida.

**sill (c).** Es el valor de la varianza al tomar todos los datos en consideración y es la mayor variación, es la altura a la que se estabiliza la curva.

**rango (a).** es el valor donde el modelo se estabiliza y este se calcula de mediante prueba y error, en general es el valor en el cual la gráfica alcanza el sill.

Estos parámetros se muestran en la figura 2.8. Con base en estos se suelen definir los modelos más comunes del krigiado, estos se muestran a continuación:

**Modelo lineal**

$$\gamma(h) = \begin{cases} c_0 + c\frac{h}{a} & si \quad 0 < h \leq a \\ c_0 + c & si \quad h > a \\ 0 & si \quad h = 0 \end{cases} \tag{2.12}$$

**Modelo esférico.**

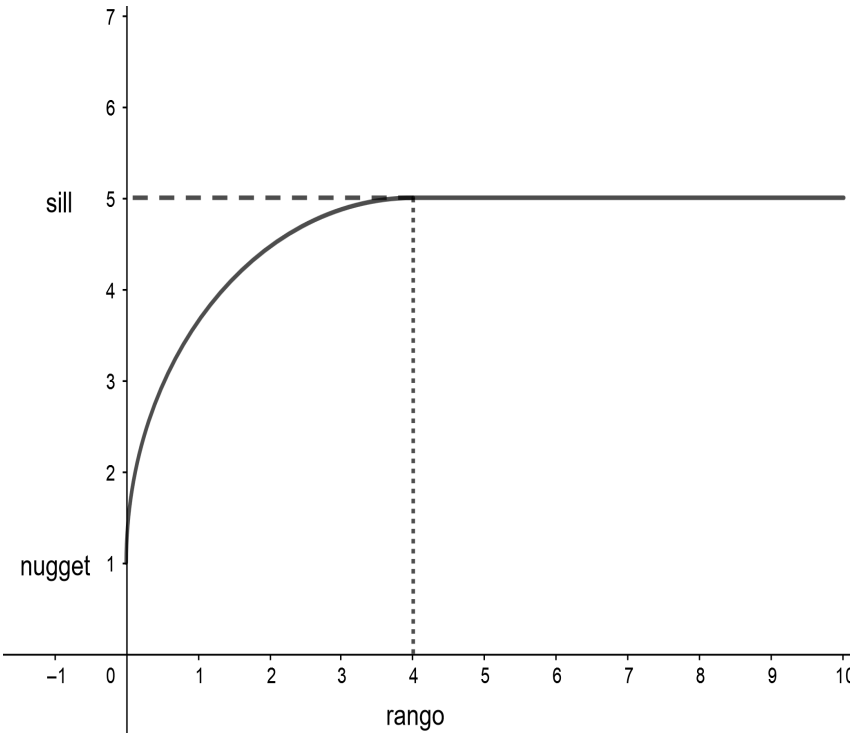



Figura 2.8: Partes del variograma

$$\gamma(h) = \begin{cases} c_0 + c(\frac{3h}{2a} - \frac{1}{2}\frac{h}{a}^3) & si \quad 0 < h \leq a \\ c_0 + c & si \quad h > a \\ 0 & si \quad h = 0 \end{cases} \tag{2.13}$$

**Modelo exponencial.**

$$\gamma(h) = \begin{cases} c_0 + c(1 - \exp^{\frac{h}{a}}) & si \quad h > 0 \\ 0 & si \quad h = 0 \end{cases} \tag{2.14}$$

**Modelo gaussiano.**

$$\gamma(h) = \begin{cases} c_0 + c(1 - \exp^{\frac{h}{a}^2}) & si \quad h > 0 \\ 0 & si \quad h = 0 \end{cases} \tag{2.15}$$

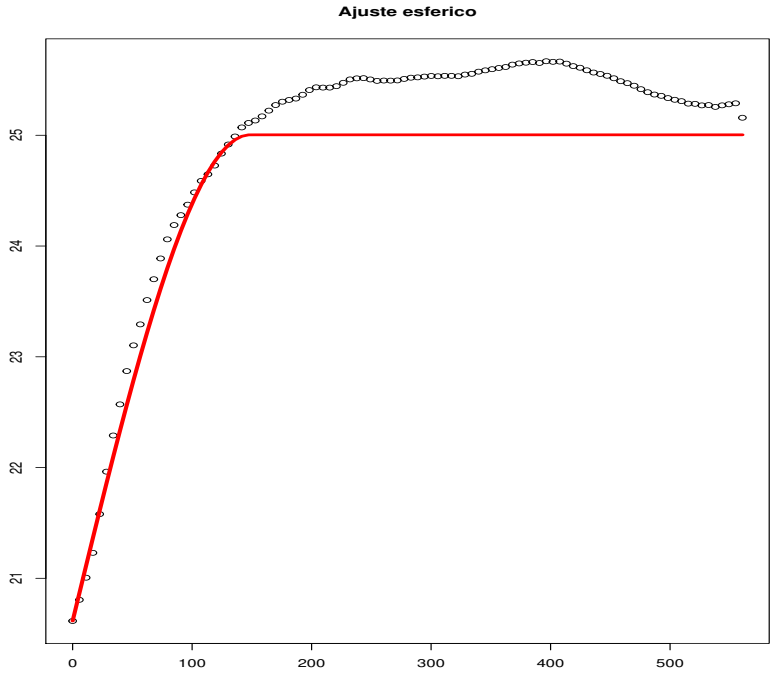



Figura 2.9: Variograma teórico

Ahora que se tienen los parámetros, estos se sustituyen en los modelos y para decidir cual modelo usar, se puede probar con métodos como el método de cuadrados medios del error y así comparar los errores y elegir el modelo que mejor ajuste, una vez hecho esto tendremos un gráfico como el de 2.9.

Con ayuda del variograma teórico obtenido, calculamos los valores $\gamma_{ij}$ (que usamos para abreviar $\gamma(d(x_i, x_j))$, donde d es una métrica en el espacio que se está trabajando). y con esto podemos calcular los pesos $\lambda_i$ y así estimar los valores en puntos no muestreados.

### 2.5.1 puntos prácticos del variograma

En esta sección se presentan algunos datos del análisis geoestadísticos

- Una línea horizontal en el variograma teórico indica que los valores no están correlacionados para ningún valor de h.
- Es deseable que el conjunto de puntos, en los cuales se basa el variograma sea construido por al menos 50 muestras y de preferencia que las muestras se encuentren a intervalos regulares.

| x | y | $z_0(x,y)$ | $\cdots$ | $z_m(x,y)$ |
|---|---|---|---|---|
| $x_1$ | $y_1$ | $z_{11}$ | $\cdots$ | $z_{m1}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $x_n$ | $y_n$ | $z_{1n}$ | $\cdots$ | $z_mn$ |

Tabla 2.4: Datos

- en la literatura del kriging h es llamada lag y $\gamma(h)$ es llamada semivarianza porque es la mitad del valor esperado del cuadrado de las diferencias.
- Para calcular $\gamma(h)$ , necesitamos encontrar todos los pares de observaciones que se encuentran a una distancia h. En la realidad, es raro que estas se encuentren a una distancia h. Para solucionar esto, se toman pares de puntos, donde la distancia entre estos se encuentran entre $(h-\epsilon, h+\epsilon)$. Evidentemente si cambiamos el valor de $\epsilon$ cambiaremos los valores del conjunto de semivarianzas.
- La covarianza empírca tiende a decrecer al incrementarse la distancia lag h. La semivarianza y la covarianza tienen tendencias contrarias y si el fenómeno tiene media y varianza que son espacialmente constantes, podemos establecer la relación entre ambas como:

$$C(0) = E[(z(x)-\mu)^2] = var[z(x)] \tag{2.16}$$

$$\gamma(h) = C(0) - C(h) \tag{2.17}$$

  Con esto, podemos ver que cuando h es suficientemente grande $\gamma(h)$ tiende al sill y $C(h)$ tiende a cero, entonces podemos poner el valor del sill como $C(0)$

## 2.6 Tipo de datos

Lo datos analizadas mediante el kriging ordinario en general provienen de bases de datos con una estructura similar a la de la tabla 2.4 donde $x, y$ indican las coordenadas del punto y $z_{ki}$ representa la observación en el punto $x_i, y_i$ de la variable $k$, con $k \in 0, \cdots, n$ , cada variable observada es presentada mediante funciones de la forma $z_n : \mathbb{R}^2 \to \mathbb{R}$ . Pero al usar el kriging ordinario solo se toma en cuenta una variable asociada a cada punto, entonces, los datos que usa el kriging tendran la forma presentada en la tabla 2.5. Descartamos el resto de las variables, con la idea de interpolar sobre la variable que no descartamos, pero se observa que al hacer esto se pierde información en el proceso.

| x | y | $z_k(x,y)$ |
|---|---|---|
| $x_1$ | $y_1$ | $z_{k1}$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $x_n$ | $y_n$ | $z_{kn}$ |

Tabla 2.5: Datos entrada kriging ordinario

Notando que si tomamos cada fila de 2.4 como un punto en un espacio de dimensión $m+2 \in \mathbb{N}$ tenemos $n \in \mathbb{N}$ puntos de un espacio de $n$ dimensiones. Dado un espacio de n dimensiones, el análisis topológico de datos nos da un método para realizar una análisis descriptivo de este tipo de espacio, además de que si tenemos bases de datos para los mismos puntos pero en diferentes instantes en el tiempo, el análisis topológico de datos nos da la posibilidad de comparar los espacios topológicos en diferentes instantes.

## 2.7 Introducción al análisis topológico de datos

En las bases de datos meteorológicas para cada punto espacial (longitud, latitud) se asocian diferentes valores de variables (precipitación, temperatura, presión atmosférica, humedad relativa,...). En general se tiene asociados a un punto en el planeta diferentes variables un ejemplo de uno de estos datos tendría la forma:

$$(\text{longitud}, \text{latitud}, \text{variable}_1, \text{variable}_2, \cdots, \text{variable}_n)$$

como ejemplo de $\text{variable}_i$ pueden tenerse variables como las siguientes: temperatura, precipitación, altura, presión, etc.

En general al realizar análisis estadístico sobre un conjunto de datos como estos se suele tomar una variable en una región y se ignora el resto de las variables y el resto de los puntos, cuando se realiza un análisis geoestadístico se toman en cuenta los puntos como realizaciones de una variable aleatoria sobre esos puntos por lo cual se toma más información de el conjunto de los datos, pero aun así descartamos las otras variables. Finalmente al realizar un análisis topológico podemos tomar en cuenta tantos datos como queramos y tomar en cuenta todas las variables de forma simultánea.

La topología es la parte de las matemáticas dedicada al estudio de las propiedades de los cuerpos geométricos inalterables por deformaciones que no corten o peguen al objeto. Esto es, un objeto es equivalente al resultado de una transformación que pueda doblarlo,

estirarlo, ampliarlo, etc, mientras esta transformación no separe al objeto en componentes disconexos, ni una partes separadas.

La esencia de la topología es darle una estructura a un conjunto de puntos, este conjunto se representa como X, dicha estructura nos ayuda a representar la cercanía y convergencia, así podemos definir:

**Topologia.** Dado un conjunto X, una topología sobre dicho conjunto, es una colección $\mathcal{T}$ de subconjuntos de X tal que las siguientes propiedades se cumplen:

1. $\emptyset$ y X estan en $\mathcal{T}$
2. La unión de cualquier subcolección de $\mathcal{T}$ esta en $\mathcal{T}$
3. La intersección de cualquier subcolección finita de elementos de $\mathcal{T}$ están en $\mathcal{T}$

A los elementos de la topología les llamamos conjuntos abiertos y a sus complementos se les llama conjuntos cerrados y a la pareja (X, $\mathcal{T}$) le llamamos **espacio topológico**.

Con base en la definición se busca analizar características cualitativas del espacio X en el que existen los puntos S. Estas características son llamadas invariantes topológicas, las cuales son las propiedades del espacio topológico invariantes bajo homeomorfismos, esto es, si un espacio X posee una propiedad, todo espacio homeomorfo a X posee la misma propiedad. Así, objetos equivalentes en la topología deben tener el mismo número de trozos, huecos, intersecciones, etc $\cdots$

Como se puede apreciar la comparación y clasificación son unos de los objetivos de la topología, con la idea de poder realizar esto, se introducen las siguientes definiciones:

Una **cubierta abierta** de un espacio topológico $X$ es una coleccion $U = (U_i)_{i\in I}$ de conjuntos abiertos de $X$, donde $I$ es un conjunto tal que $X = \cup_{i\in i} U_i$.

**Función continua.** Es una función f de X a Y donde X,Y son espacios topológicos y para todo elemento V de la topología en Y la preimagen de V es un abierto en X.

**Homeomorfismo.** Es una función f biyectiva entre 2 espacios topológicos X, Y donde f y $f^{-1}$ son continuas.

Cuando existe un homeomorfismo entre 2 espacios se dice que estos son homeomorfos y esto es que tienen la misma estructura. Esto se puede consultar con más detalle en [28], en [29] y [30]

Con esto podemos analizar espacios, encontrando espacios homeomorfos que podamos representar de manera más sencilla, pues al ser homeomorfos compartirán las mismas cualidades.

Invariantes topológicos

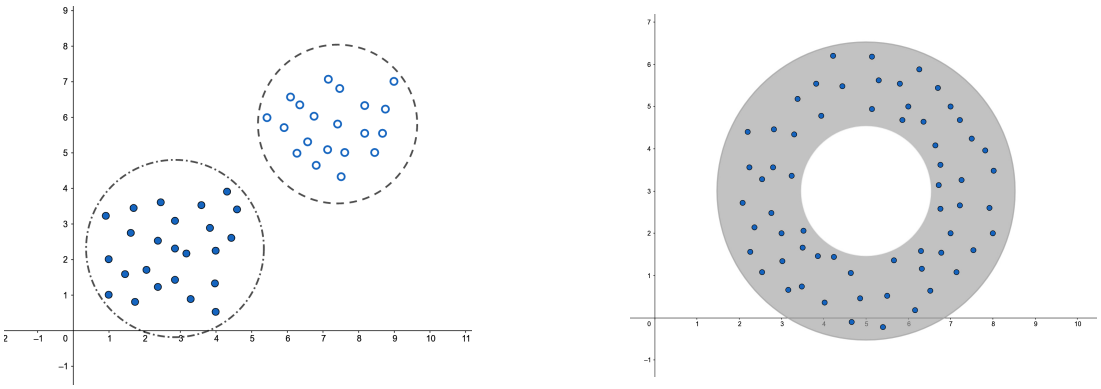
1. cardinalidad
2. número de componentes conexos
3. compacidad
4. metrizabilidad
5. separacion
6. grupo de homología

Entre las principales opciones que nos presenta el análisis topológico de datos se encuentran:

- Algoritmo mapper
- Calculo de Euler
- Cellular sheaves
- homología persistente

En este trabajo nos concentramos en el análisis de homología persistente sobre datos atmosféricos. Este tipo de análisis, se basa en la idea de que podemos descomponer un espacio X y reconstruirlo como una combinación de bloques finitos llamados simplejos, los cuales son un conjunto finito de elementos del espacio (Ejemplo de dos componentes en figura) y al conjunto que contiene a todos los simplejos se le llama complejo simplicial, el cual representaremos mediante $K$, El cual nos servirá como esqueleto para representar el espacio X y el cual tiene la ventaja de que puede verse como un objeto topológico o como un objeto combinatorio. Como nuestra intención es que este espacio represente al espacio X, este debe mantener las propiedades importantes del X. Estas propiedades se estudian mediante los grupos de homología, los cuales se denotan mediante $H_0, H_1, \cdots, H_n, \cdots$. La información que estos grupos nos presentan, son el número de ciclos de dimensión n en el espacio, la cual está dada por $B_n(K) = dim(H_n)$, como ejemplo, en el caso de $H_0$, $B_0(K) = dim(H_0)$ nos informa el número de componentes conexas en el espacio (Ejemplo

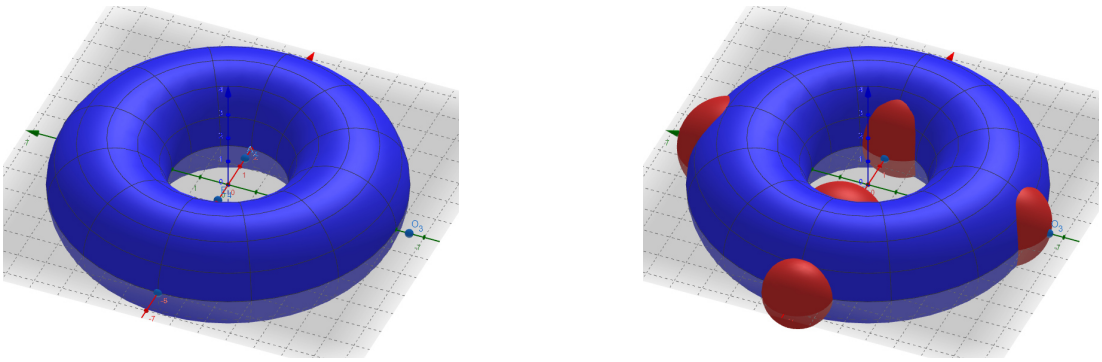
de un ciclo de dimensión 0 en figura 2.10 a) y $B_1(K) = dim(H_1)$ nos dice el número de huecos (ejemplo en figura 2.10 b). Como otro ejemplo, en el inciso a de la figura 2.11, el número de componentes conexos es 1 y el número de huecos es 1. Este tipo de análisis tiene la ventaja de poder aplicarse cuando los datos presentan ruido ya que las propiedades topológicas se mantienen, como se observa en la figura 2.11 b. Además de esto se puede trabajar con los datos cuando estos se encuentran en un espacio de alta dimensión.



a b

Figura 2.10: Ejemplo de huecos de dimencion 0 y 1

Más acerca de los complejos simpliciales se pueden encontrar en [19].



a) Toro sin ruido b) Toro con ruido

Figura 2.11:

En qué tipos de datos se puede aplicar:

- espacios métricos finitos
- imagenes

- redes

La intención tras el análisis de homología persistente es presentar de forma resumida algunas características del espacio del cual provienen los datos, un ejemplo de esto son los diagramas de vida y muerte, como el de 2.12

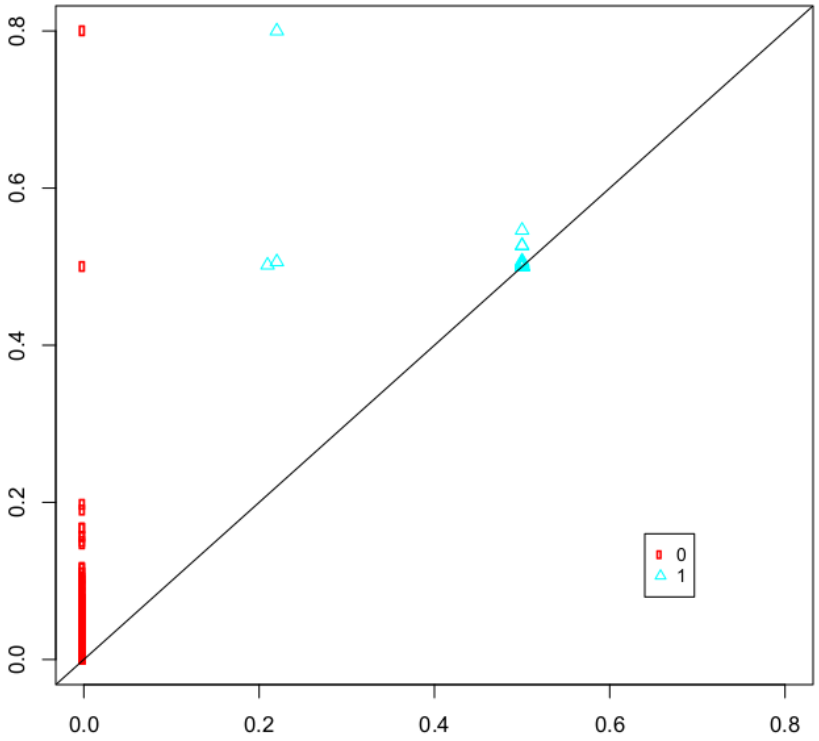



Figura 2.12: Gráfico de vida y muerte

El cual es un conjunto de puntos $(x, y)$ donde $x$ representa en qué momento aparece una característica de dimensión k y donde $y$ representa en qué momento desaparece la carcterística, como se puede observar $x < y$ entonces todos los puntos aparecen sobre la línea $y = x$ y entre los puntos están más cerca de esta línea significa que las características tienen una vida corta.

Más detalles se pueden consultar en [23] y en [16].

## 2.8 Topología algebraica

Se representara el espacio topológico como un complejo simplicial que cuente con las mismas invariantes topológicas, por lo tanto primero se definirá lo que es un complejo simplicial.

Dado un conjunto $S = \{s_1, s_2, \cdots, s_n\}$ un complejo simplicial abstracto es un conjunto de subconjuntos de S, $K = \{\sigma \subset X\}$ tal que $\forall \beta \subset \sigma$ con $\sigma \in K$, entonces $\beta \in K$. a los elementos de K se les llama simplejos y dado $\sigma \in K$ donde $\sigma$ cuenta con n elementos, se dice que es un n-1 simplejo. Dado K, $dim(K)$ es la dimensión del mayor simplejo en K

Si tenemos un complejo simplicial K que representa al espacio X, lo siguiente es generar $C_i(K; F)$ con $i \in \{0, 1, \cdots, dim(K)\}$ que son los espacios vectoriales generados con los simplejos de orden k sobre el campo F.

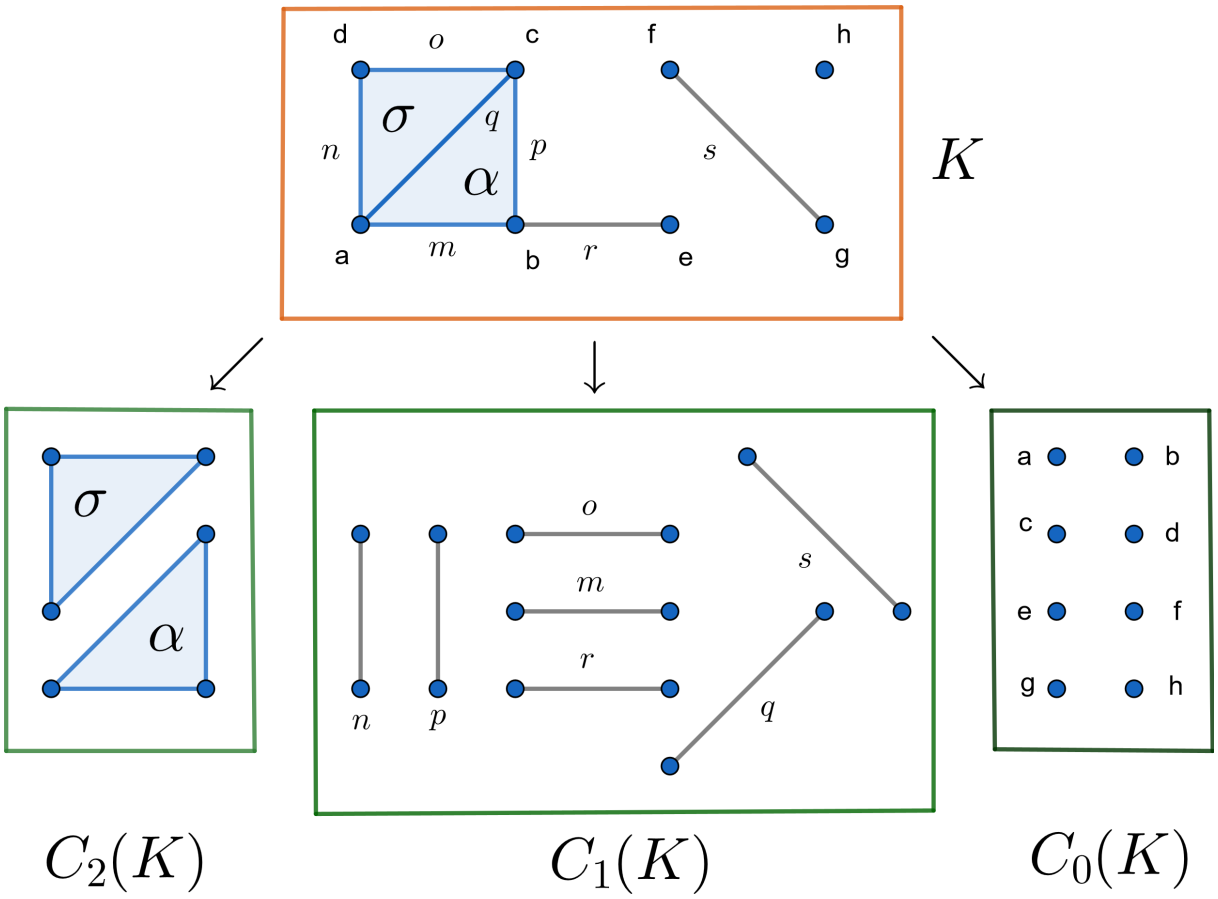



Figura 2.13: Ejemplo de descomposición de K en $C_i(K)$

En adelante se usará $\mathbb{F}_2$ como el campo y se denotará $C_i(K) = C_i(K; \mathbb{F}_2)$. Con esto se tiene $\alpha_i + \alpha_j$ si $i \neq j$ y 0 si $i = j$

Lo siguiente es que necesitamos un conjunto de mapeos líneales líneal de $C_i(K)$ a $C_{i-1}(K)$ el cual llamamos mapeo frontera y está dado por:

$$\partial_i : C_i(K) \rightarrow C_{i-1}(K)$$

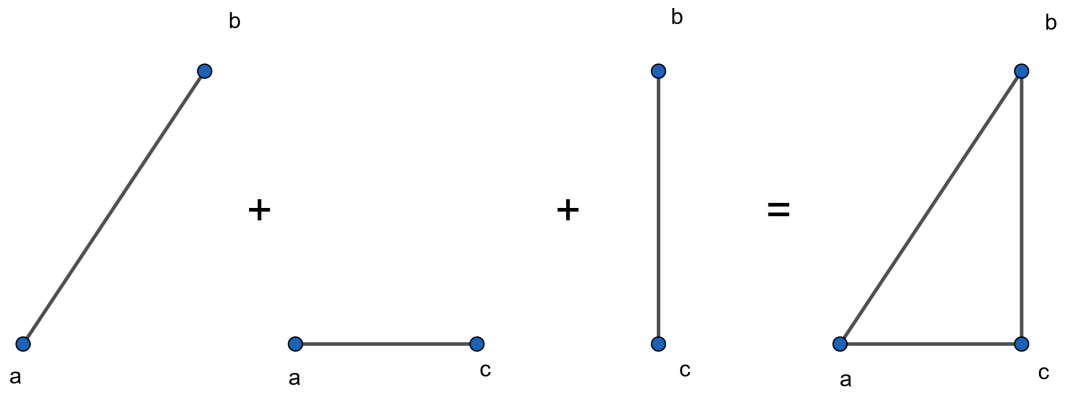



Figura 2.14: Ejemplo de suma en $C_1(K)$ con todos los elementos distintos

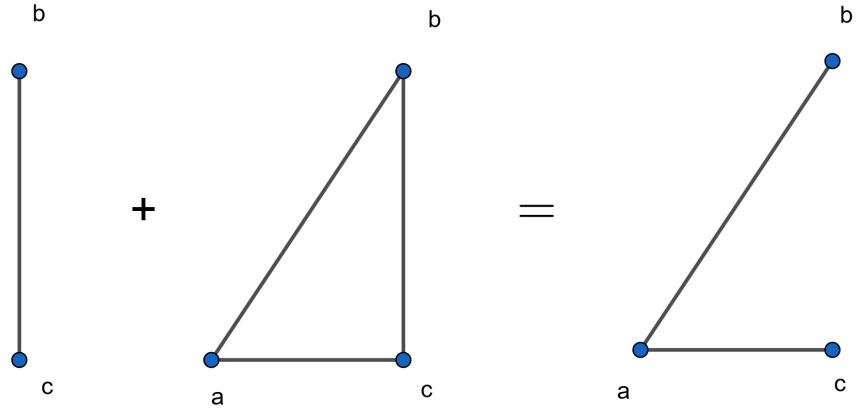


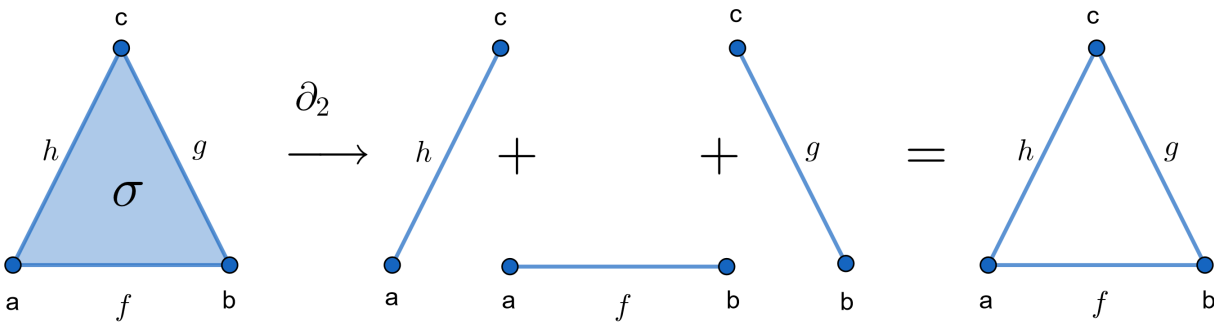

Figura 2.15: Ejemplo de suma en $C_1(K)$ con un elemento repetido

$$\partial_i(\sigma) \mapsto \sum_{\alpha \subset \sigma, \alpha \in C_{i-1}(K)} \alpha$$

Este mapeo lo que hace es que descompone $\sigma$ en su frontera, que es la suma de todos los elementos que están en *sigma* y que son de dimensión $i-1$ un ejemplo se ve en la figura 2.8

Estos mapeos tienen la siguiente propiedad: $\forall \sigma \in C_{i+1}(K), \partial_i \circ \partial_{i+1}(\sigma) = 0$ , lo que significa $img(\partial_{i+1}) \subset ker(\partial_i)$. Con esto generamos el espacio vectorial cociente

$$H_k(K) = \frac{ker(\partial_i)}{img(\partial_{i+1})}$$

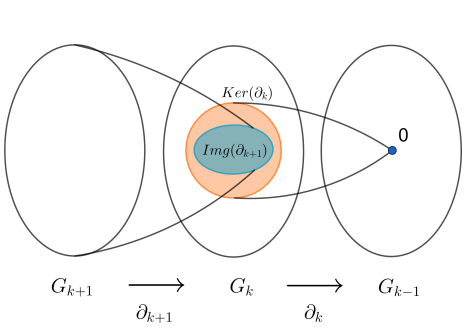


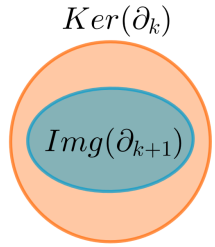

$$\partial_2(\sigma) \quad \longmapsto \quad f + g + h$$

Figura 2.16: Ejemplo de mapeo frontera

Los espacios cocientes pueden consultarse en [31], en [32] y en [34].

Esta construcción se representa en la figura 2.8. A los elementos de $ker(\partial_i)$ son llamados p-ciclos y los elementos de $img(\partial_{i+1})$ son p-fronteras. Entonces lo que nos dice $B_p(K) = dim(ker(\partial_p)) - dim(Img(\partial_{p+1}))$ es que los p-ciclos que no sean p-fronteras, son los huecos de dimensión p que presentan el espacio.




$$H_k = \frac{Ker(\partial_k)}{Img(\partial_{k+1})}$$

Figura 2.17: Construcción de grupos de homología

## 2.9 Homología persistente

Para tener grupos de homología que representen al espacio se debe cumplir que el complejo simplicial posea características invariantes semejantes, pero cuando se tiene una base de datos dada, entonces nos quedan dos problemas:

- Construir complejos simpliciales a partir de los datos
- Que estos complejos tengan las características del espacio del cual provienen

A continuación se presentarán algunas de las principales formas de construir complejos simpliciales a partir de datos. Para realizar estas construcciones necesitamos que el espacio topológico con el cual se está trabajando sea un espacio métrico, lo cual implica que tenemos un función distancia $d : X \to \mathbb{R}$, que cumple las siguientes propiedades:

- $\forall x, y \in X, d(x, y) \geq 0$
- $d(x, y) = 0 \leftrightarrow x = 0$
- $\forall x, y, x \in X$ se tiene que $d(x, z) \leq d(x, y) + d(y, z)$

Al conjunto X junto con una métrica d asociada a el, le llamamos espacio métrico, esto se puede representar mediante $(X, d)$.

Más sobre métricas se puede consultar en [33] y en [34].

Esta métrica nos dice cómo construir los conjuntos abiertos del espacio topológico, los cuales tendrán la forma $B(x, \epsilon) = \{y \in X | d(x, y) < \epsilon\}$ y serán usados para construir los complejos simpliciales.

Una clasificación de los tipos de complejos simpliciales sería la siguiente:

- Nervios
- Witness
- Flag complex o clique complex

Esto se puede consultar en [23].

### 2.9.1 Nervios

Los nervios en este contexto son un tipo de complejos generados a partir de una buena cubierta V de un espacio topológico X. En este caso, una buena cubierta es una cubierta finita, $\forall M \in V$ se tiene que M es contractible y además se tiene que $\forall M, N \in V$ con $M \cap N \neq \emptyset$ entonces $M \cap N$ es contractible, dadas estas condiciones el complejo se forma:

- Se toman los elementos de V como los 0-simplejos
- Se forman los n-simplejos cuando tenemos un $\sigma \subset V$ donde $\sigma$ tiene n + 1 elementos y además:

$$\cap_{M \in \sigma} M \neq \emptyset$$

Entre este tipo de complejos se encuentran los siguientes:

- Cech
- Delaunay
- Alpha

### 2.9.2 Čech complex

El complejo de Čech a escala $\epsilon$ del espacio $S \subset X$ se construye de la siguiente forma:

- Para cada $s \in S$ se generan las vecindades $V_\epsilon(s) = \{x \in X | d(s, x) \leq \epsilon\}$
- La colección $V = \{V_\epsilon(s) \subset X | s \in S\}$ es una cubierta de $\cup V \cap X$
- Se calcula $N(V)$ que es el nervio de la cubierta V

### 2.9.3 Delaunay complex

Considera a $S \subset X = \mathbb{R}^n$, entonces, $\forall s \in S$ se crea el conjuntos:

$$V_s = \{X \in \mathbb{R}^n | d(x, s) \leq d(x, s) \forall s' \in S\}$$

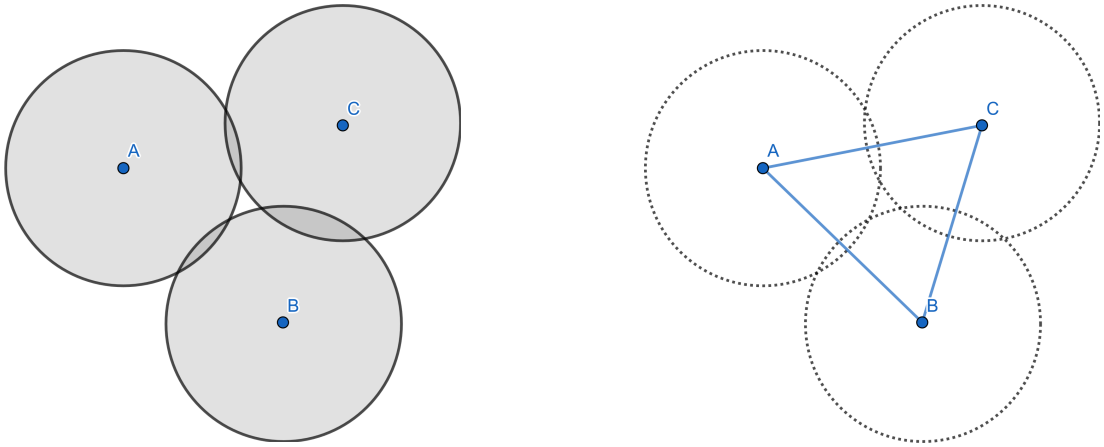



Figura 2.18: 1-simplejos de Čech

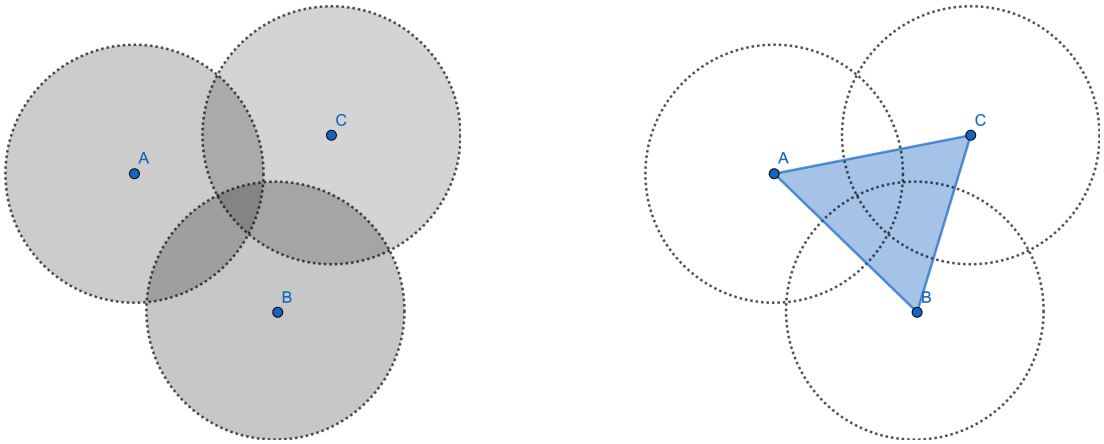



Figura 2.19: 2-simplejo de Čech

El cual es el conjunto de puntos de X que son más cercanos a s que a cualquier otro s' en S. El conjunto $V = \{V_s \subset \mathbb{R}^n \ s \in S\}$ es una cubierta de X, llamada descomposición de Voronoi de X con respecto a S

El nervio de esta cubierta es el complejo de Delaunay[1], se escribe $Del(S; \mathbb{R}^n)$

### 2.9.4 Alpha complex

Dado $S \subset X$, entonces formamos la descomposición de Voronoi $V = \cup_{s \in S} V_s$, esta es una cubierta abierta de X, al formar $N(V)$ obtenemos un nervio de X, entonces se buscará obtener una representación del espacio S, para esto, dado un $\epsilon$ tenemos para todo $s \in S$ tenemos $B(s, \epsilon)$. La unión de estas bolas abiertas $S_\epsilon = \cup_{s \in S} B(s, \epsilon)$ es un subespacio de X y

[1] también llamada triangulación de Delaunay

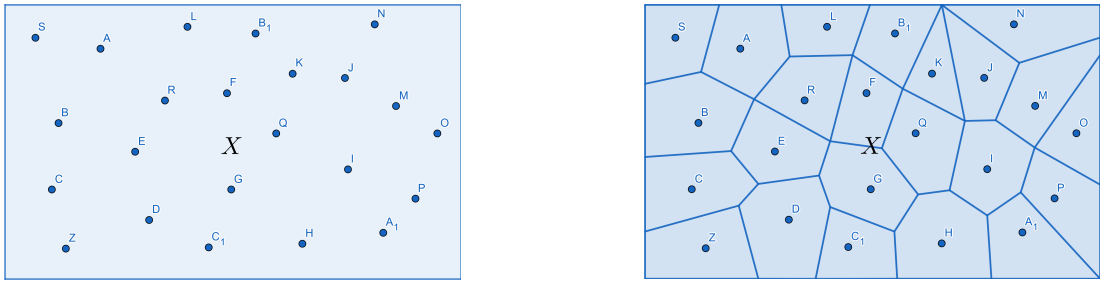



Figura 2.20: Complejos de Delaunay

al intersectar $B(s,\epsilon)\cap V_s$ para todo s en S, tenemos $A_\epsilon = \{U \subset X | U = B(s,\epsilon)\cap V_s \forall s \in S\}$ una cubierta abierta de $S_\epsilon$ y al obtener el nervio $N(A_\epsilon)$ obtenemos el Alpha complejo de resolución $\epsilon$

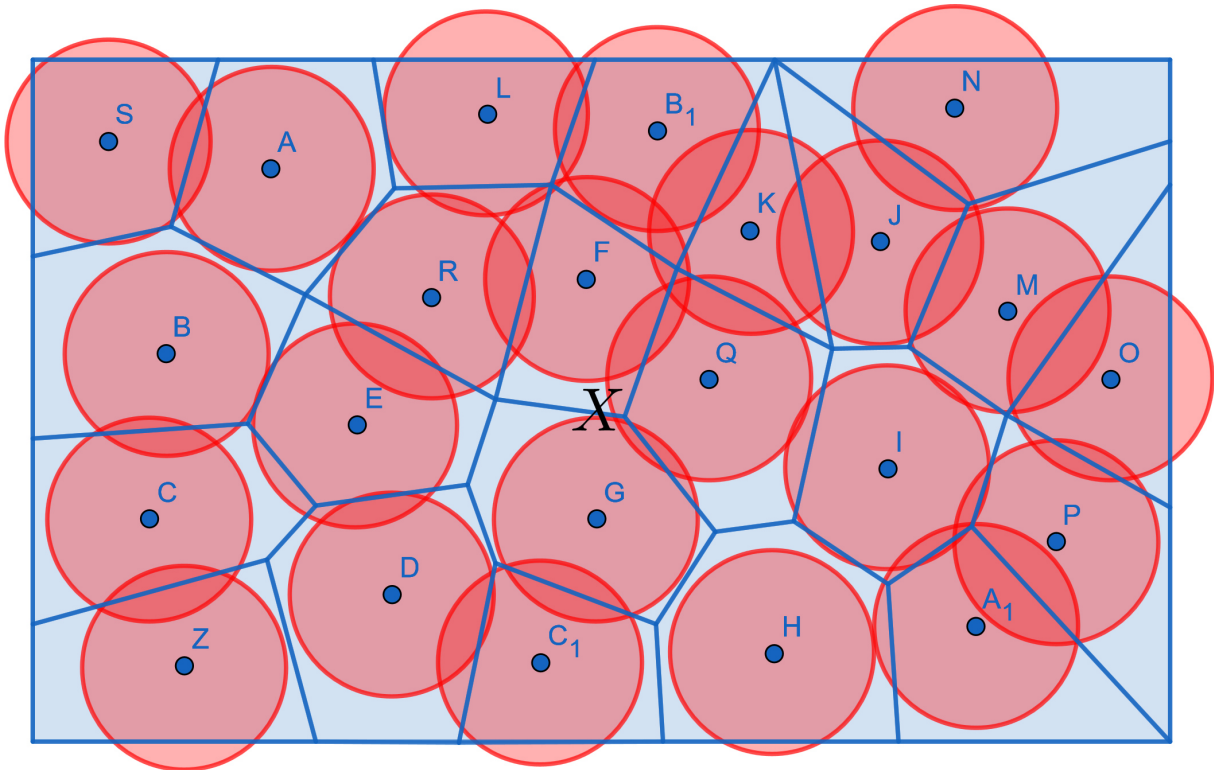



Figura 2.21: Base para los Alpha complex

### 2.9.5 Witness complex

Entre más puntos tengamos en el espacio S a analizar, se empezaran a crear complejos de mayor dimensión, hasta tener un complejo de dimensión $n = |S| = dim(S)$,

entonces, la idea es que puede bastar con un menor número de puntos para observar la forma de S. La construcción de los complejos witness es la siguiente:

Dado $L \subset S$, los elementos del conjunto L son llamados landmarks, estos serán los nodos y los puntos de S se les llamará witness, dado que los puntos witness son testigos de la conexión que existe entre los landmarks

Dado un subconjunto $\sigma \subset L, s \in S$, s es un testigo débil de $\sigma$ si:

$$d(s,a) \leq d(s,b) \forall a \in \sigma \forall b \in L - \sigma$$

Esto es, el punto s está más cerca de los elementos de $\sigma$ que de cualquier otro landmark

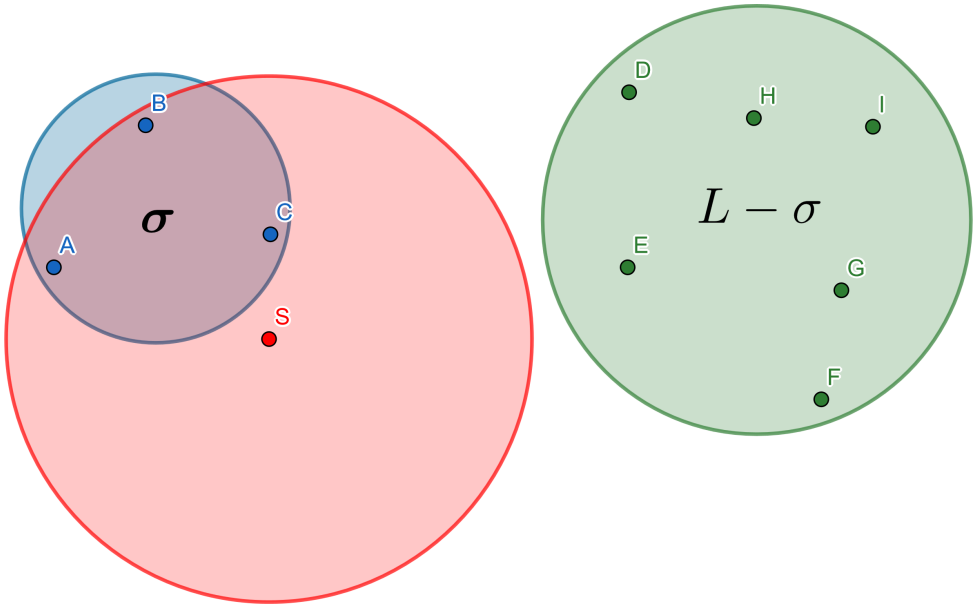



Figura 2.22: Complejos Witness

El complejo débil de Delaunay con respecto a L es el complejo $Del^W(L;S)$ que consta de todos los $\sigma$ que tienen un testigo débil.

Una forma de extender esto para usar distancias $\epsilon$ para tener diferentes resoluciones, es cambiar cómo obtenemos los testigos débiles. S es un $\epsilon$-testigo débil de $\sigma$ respecto a L si y sólo si

$$d(s,a) \leq d(s,b) + \epsilon \forall a \in \sigma \forall b \in L - \sigma$$

Ahora, el complejo simplicial débil de Delaunay a escala $\epsilon$, $Del^W(L;S,\epsilon)$ es el complejo simplicial con vértices en L y donde $\sigma \subset L$ es un simplejo si tiene un $\epsilon$-testigo débil, estos complejos son llamado weak witness complex

### 2.9.6 flag complex

Dado un grafo G = V,A y dado un conjunto de vértices $M = \{m_0, m_1, \cdots, m_n\} \subset V$ forma un n-simplejo si cada par de vértices está conectado por una arista, esto es:

$$\forall m_i, m_j \in M, i \neq j, \{m_i, m_j\} \in A$$

Al conjunto M se le llama clique

Entonces un flag complex es el que dado el conjunto de aristas A (el conjunto de 1-simplejos), se crean todos los clique's para generar el resto de los n-simplejos. Como en este tipo de complejos solo se revisa la distancia a pares, los clique complex también son llamados Lazy porque solo utilizan los 1-simplejos para generar los demás simplejos.

Entre los más conocidos de estos complejos son:

- Vietoris-Rips
- lazy witness

En general el proceso para generar estos complejos es el siguiente:

- Calcular todos los 0-simplejos
- Calcular los clique complejos

### 2.9.7 Vietoris-Rips

El complejo de Vietoris-Rips a escala $\epsilon$ es el conjunto:

$$VR_\epsilon(S) = \{\sigma \subset S : d(x, y) \leq 2\epsilon \forall x, y \in \sigma\}$$

Esto es el conjunto S se toma como los 0-simplejos y se crean los 1-simplejos $\{s_i, s_j\}$ donde $d(s_i, s_j) \leq 2\epsilon$ y con esto los n-simplejos son los conjunto de n+1 puntos que forman un clique

### 2.9.8 Lazy witness complex

Estos son una modificación de los weak witness complex, que consiste en:

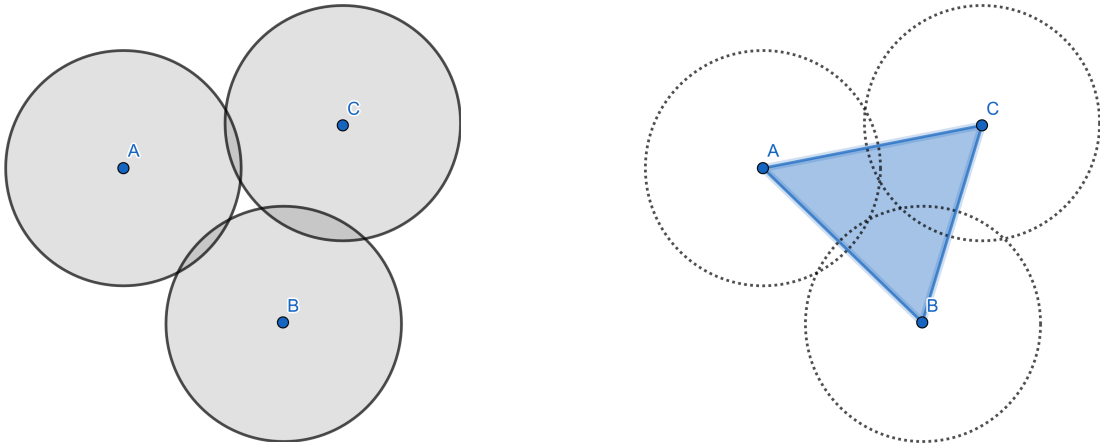



Figura 2.23: Complejos de Vietoris-Rips

- dado $n \in \mathbb{N}$, para todo $s \in S$ se tiene que $m_n(s)$ es la distancia de s al n landmark más cercano (para todo s $m_0(s) = 0$)
- Definimos a los elementos de L como los vértices
- se crean los 1-simplejos de la siguiente forma: $\{l_i, l_j\}$ son un 1-simplejo si existe un $s \in S$ tal que $max\{d(l_i, s), d(l_j, s)\} \leq m_v(s) + \epsilon$
- se agregan los clique

Este complejo simplicial se representa mediante $W_n(L; S, \epsilon)$

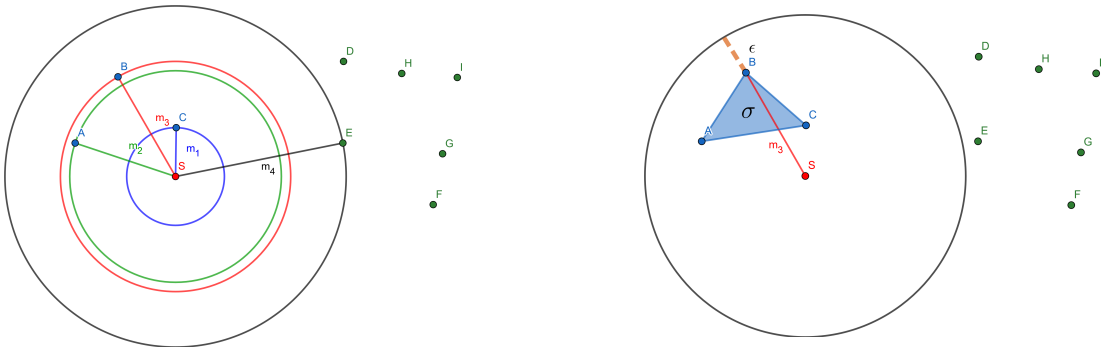



Figura 2.24: $m_3(s)$ 2-simplejo Lazy witness

En el caso de n = 0 tenemos que $W_0(L; S, \epsilon)$ aproxima $VR(L; \epsilon)$

### 2.9.9 Homología persistente

Ahora que tenemos una forma de construir complejos simpliciales, lo siguiente, es que necesitamos obtener las características del espacio del que proviene, para esto, tendríamos que tener un $\epsilon$ tal que el complejo simplicial generado represente el espacio del que provienen los datos. Dependiendo del $\epsilon$ se tendrán complejos simpliciales con propiedades diferentes, estos se representarán como $K_\epsilon$, para empezar, si $\epsilon = 0$, se tendrá que $B_0(K_0) = n$ donde n es el número de elementos con el que creamos el complejo simplicial, en cambio, si $\forall x, y \in X$ se tiene $m \geq d(x, y)$ se tendrá $B_0(K_m) = 1$. Entonces se toma el siguiente enfoque, se tomará $\epsilon$ como la resolución del espacio generado por el complejo simplicial generado. Con esto, en lugar de las características del espacio del que provienen los puntos, se observan cuáles características persisten más al variar la resolución.

Para observar las características que persisten se tomar los $K_\epsilon$ como una filtración, con esto se tendra:

$$K_0 \subset K_{\epsilon 1} \subset K_{\epsilon 2} \subset \cdots \subset K_{\epsilon m-1} \subset K_{\epsilon m}$$

Se denota a la filtración como $K = \{K_0, K_{\epsilon 1}, K_{\epsilon 2}, \cdots, K_{\epsilon m-1}, K_{\epsilon m}\}$

Sobre cualquiera de estos complejos simpliciales se pueden crear los grupos de homología, pero para la persistencia homológica tomaremos un mapeo entre las complejos simpliciales

un mapeo entre complejos simpliciales K , L abstractos, es un mapeo $f : K \to L$ donde $\forall \sigma_k \in K, \exists \sigma_l \in L$ tal que $f(\sigma_k) = \sigma_l$. Ya que generamos un espacio vectorial a partir de los complejos simpliciales, este mapeo $f$ induce un mapeo $\tilde{f}$ tal que el diagrama que aparece en la figura 2.9.9 es un diagrama conmutativo

$$\begin{array}{ccc} K & \xrightarrow{\quad f \quad} & L \\ \downarrow & & \downarrow \\ C_p(K) & \xrightarrow{\tilde{f}_p} & C_p(L) \end{array}$$

Figura 2.25:

entonces como tenemos una filtración $K$, si $K_i \subset K_j$ entonces el mapeo inclusión nos genera los siguientes mapeos

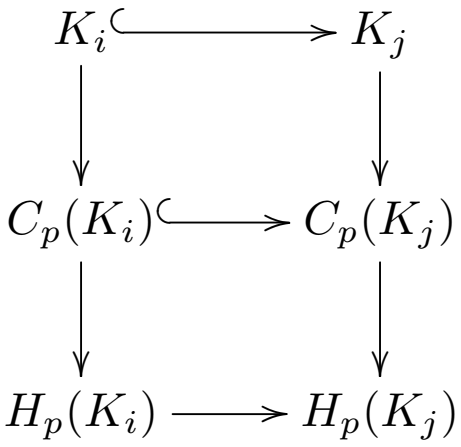



Figura 2.26:

Al aplicar estas propiedades a la filtración $K$ nos da la relación presentada en la figura 2.27, donde cada fila $\epsilon x$ son los grupos de homología del complejo simplicial $K_{\epsilon x}$.

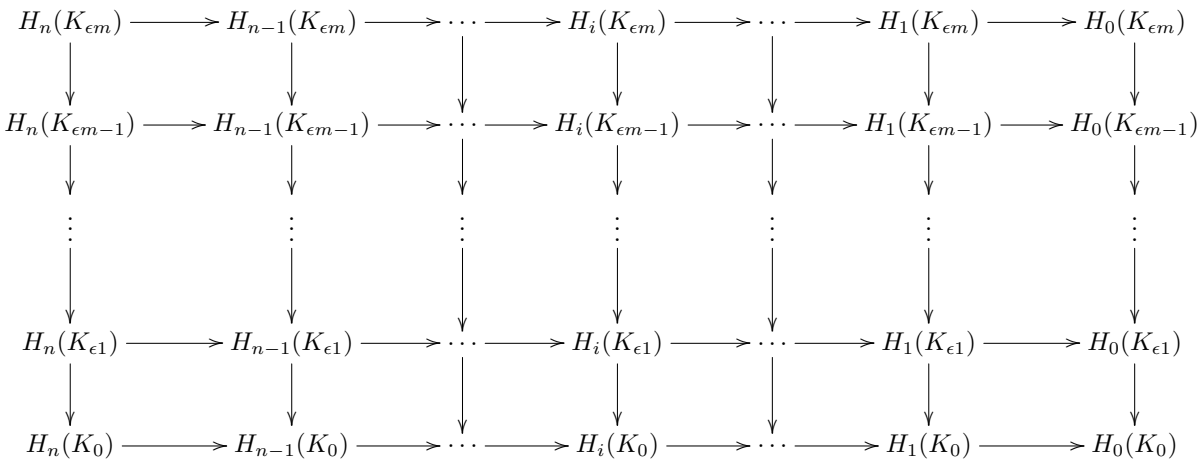



Figura 2.27: Mapeos que generan la homología persistente

Lo que nos interesa de la figura 2.27 no son las filas, si no las columnas, esto es porque la columna i nos indica cómo evolucionan las propiedades de dimensión i del complejo simplicial a diferentes valores de $\epsilon$, esto es nos interesan los mapeos de la forma:

$$H_i(K_0) \to H_i(K_{\epsilon 1}) \to H_i(K_{\epsilon 2}) \to \cdots \to H_i(K_{\epsilon j}) \to \cdots \to H_i(K_{\epsilon m-1}) \to H_i(K_{\epsilon m})$$

lo que se analiza aquí son los generadores de $H_i(K_{\epsilon j})$ y con estos se crea lo que es un diagrama de código de barras o un diagrama de vida y muerte, para el primero, se cuenta

el número de generadores, esto nos dice que en la resolución $\epsilon j$ se tienen $\alpha$ generadores, se tendrán $\alpha$ puntos sobre el eje x en el valor de $\epsilon j$ y se dibujara una línea horizontal hasta la recta $x = \epsilon j + 1$. Si en el instante $\epsilon j + 1$ se tiene $\beta$ generadores de $H_i(K_{\epsilon j+1})$, se dibujarán $\beta$ puntos en el valor del eje x igual a $\epsilon j + 1$, ahora se tendrán los siguientes casos:

1. $\alpha \leq \beta$, se conectarán cada punto sobre $\epsilon j$ con uno en $\epsilon j + 1$ y se dibujara una línea horizontal desde cada punto sobre $\epsilon j + 1$ hasta $\epsilon j + 2$

2. $\alpha > \beta$, se conectarán $\beta$ puntos de $\epsilon j$ con puntos de $\epsilon j + 1$ y se tendrá que $\alpha - \beta$ terminaran en la resolución $\epsilon j + 1$ después se dibujara una línea horizontal desde cada punto sobre $\epsilon j + 1$ hasta $\epsilon j + 2$

En el caso 1 se dice que $\beta - \alpha$ puntos nacieron en $\epsilon + 1$ y en el caso 2 se dice que $\alpha - \beta$ puntos murieron en la resolución $\epsilon + 1$

Este proceso se realiza para todos los elementos de la filtración y esto nos da el diagrama de código de barras de $H_i$, entre mayor longitud tenga la vida de una característica, esto es la diferencia entre el punto del eje x donde esta muere con el punto del eje x donde esta aparece, es más probable que sea una característica del espacio, esto nos da el intervalo de vida [$\epsilon i$ donde aparece,$\epsilon j$ donde muere ]. Para obtener el diagrama de vida y muerte simplemente se dibujan todos los intervalos de vida sobre el plano donde $\epsilon i$ donde aparece se toma como la coordenada x y $\epsilon j$ donde muere se toma como la coordenada y, aquí entre más alejado esté un punto de la diagonal más probable es que sea una característica real del espacio y entre más cerca de esta, más probable que sea ruido

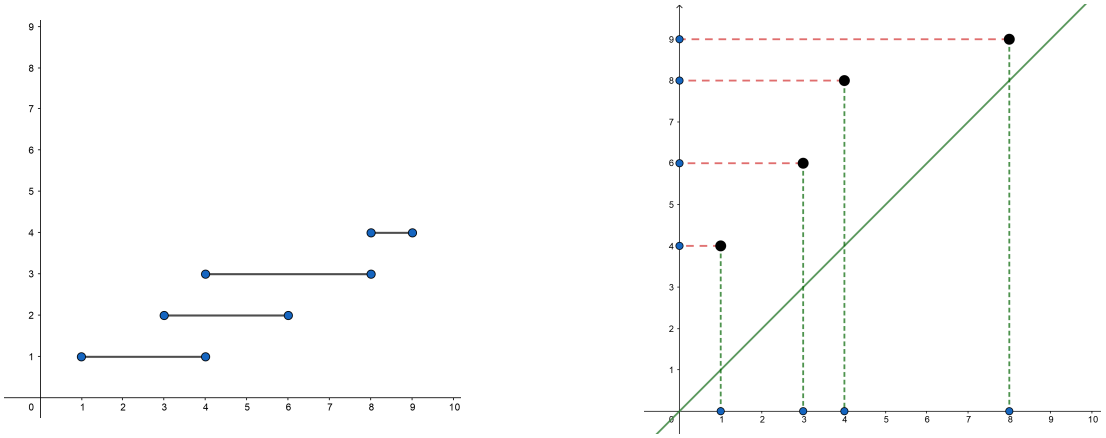


Figura 2.28:

Más detalles se pueden consultar en [18] y en [16].

## 2.10 ANÁLISIS DE LA PERSISTENCIA DE GRUPOS DE HOMOLOGÍA

Cuando se tiene un código de barras o un diagrama de vida y muerte nos interesa interpretarlo, para esto se tienen principalmente los siguientes enfoques:

- Estudiar espacios métricos cuyos puntos son diagramas de persistencia, cada diagrama es un punto en este caso
- Encontrar una forma de transformar la información obtenida a un espacio métrico donde se pueda realizar análisis de las características del espacio

Se verá un diagrama de vida y muerte como un multiconjunto de $\mathbb{R}^2$ junto con la line $y = x$ los cuales tienen multiplicidad infinita, esto es ya que si tenemos A, B dos diagramas de vida y muerte con diferente número de puntos, los demás puntos se mapean a la línea. Dado que tenemos esto, si A y B son dos diagramas de vida y muerte, sea F el conjunto de todas las biyecciones de A en B y $p \in \mathbb{N}$, entonces la p-distancia de Wasserstein está dada por:

$$W_p\left[d\right](A,B) = \inf_{f\in F}\left[\sum_{x\in A} d(x,f(x))\right]^{\frac{1}{p}}$$

En el caso que tengamos que $p = \infty$ tenemos la distancia de cuello de botella [2]:

$$W_p\left[d\right](A,B) = \inf_{f\in F}\sup_{x\in A} d(x,f(x))$$

Un ejemplo de estos mapeos se ve en la figura 2.29, ya que se tienen los puntos relacionados mediante f, d es una métrica en $\mathbb{R}^2$

Para esto se puede observar que lo que se busca es tener múltiples diagramas para poder realizar análisis estadístico entre los diagramas

Para estudiar las propiedades de un diagrama en forma independiente se buscará mapear el diagrama a un espacio donde se pueda hacer estadística o machine learning, entre los métodos que se tienen para hacer esto están los siguientes:

- persistence landscapes

[2] Bottleneck distance

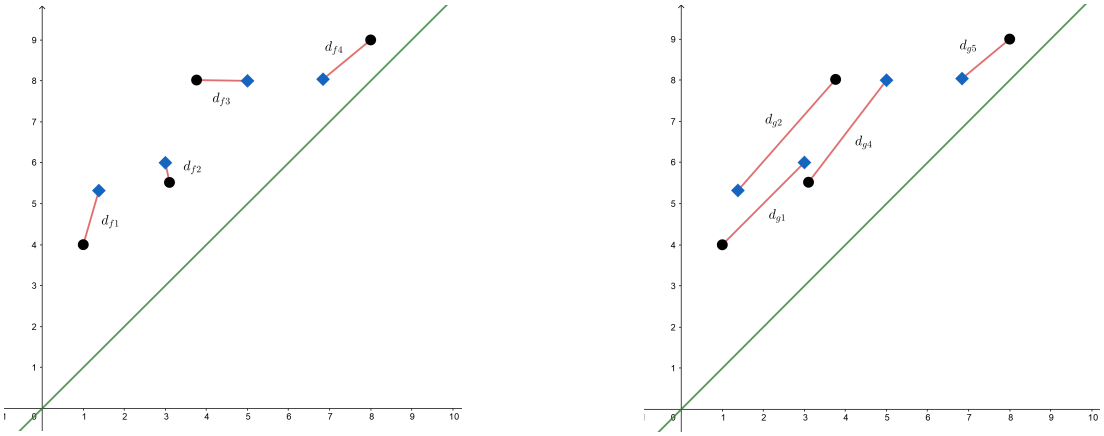



Figura 2.29:

- espacio de funciones algebraicas
- persistence images
- técnicas de kernelization

Capítulo 3

# Metodología

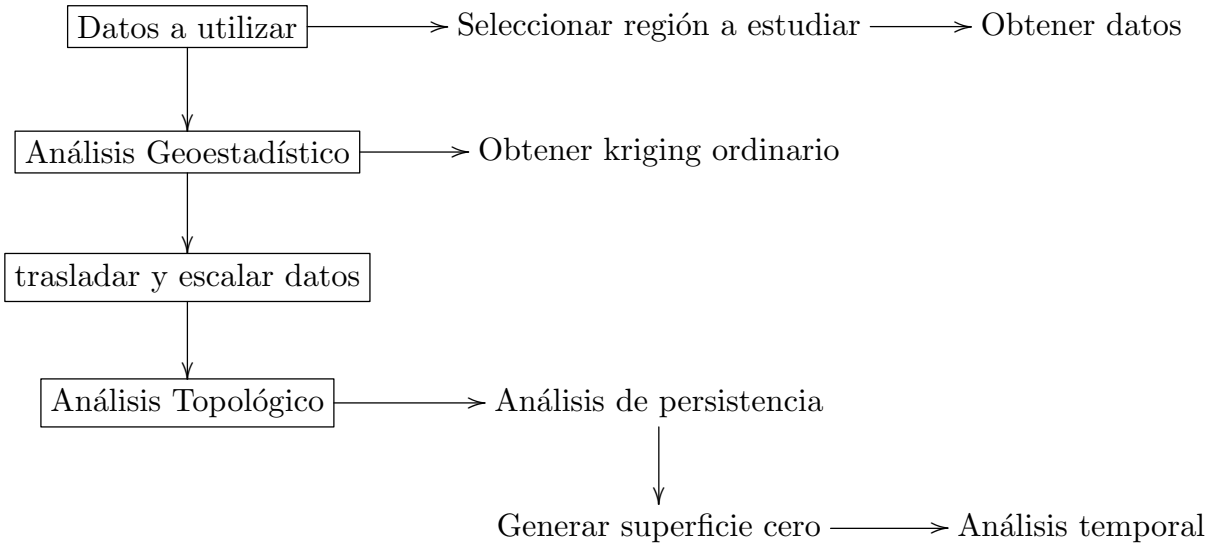



Figura 3.1:

## 3.1 Datos a Utilizar

Los datos que se utilizaron fueron descargados de la página NEO [1] [2] Donde se encuentra una gran cantidad de datos atmosféricos, en particular, los datos que se usaron fueron descargaron de la sección atmospher/rainfall con las siguientes características:

- Formato: CSV para excel
- Fechas: enero del 2015 - enero del 2016
- Resolucion: 1440x720

[1] nasa earth observations
[2] https://neo.sci.gsfc.nasa.gov/

- Separación entre puntos, longitud: 0.15 °
- Separación entre puntos, latitud: 0.15 °

Los datos utilizados en este trabajo fueron recolectados por la misión Tropical Rainfall Measuring Mission (TRMM), que es una mision conjunta entre la nasa y la agencia espacial japonesa (JAXA). Estas mediciones fueron obtenidas por un satélite que orbita cerca del ecuador, las cuales están tomadas en milímetros (un milímetro equivale a un litro por metro cuadrado), los datos registrados se encuentran entre las latitudes 35 norte y 35 sur, mas información de como se obtienen estos datos se puede consultar en [40] o en [41].

Los datos seleccionados para ser trabajados se encuentran entre las longitudes: -104 y - 98.5 y las latitudes: 22 y 27 esta región se presentan en la figura 3.1 y en 3.3

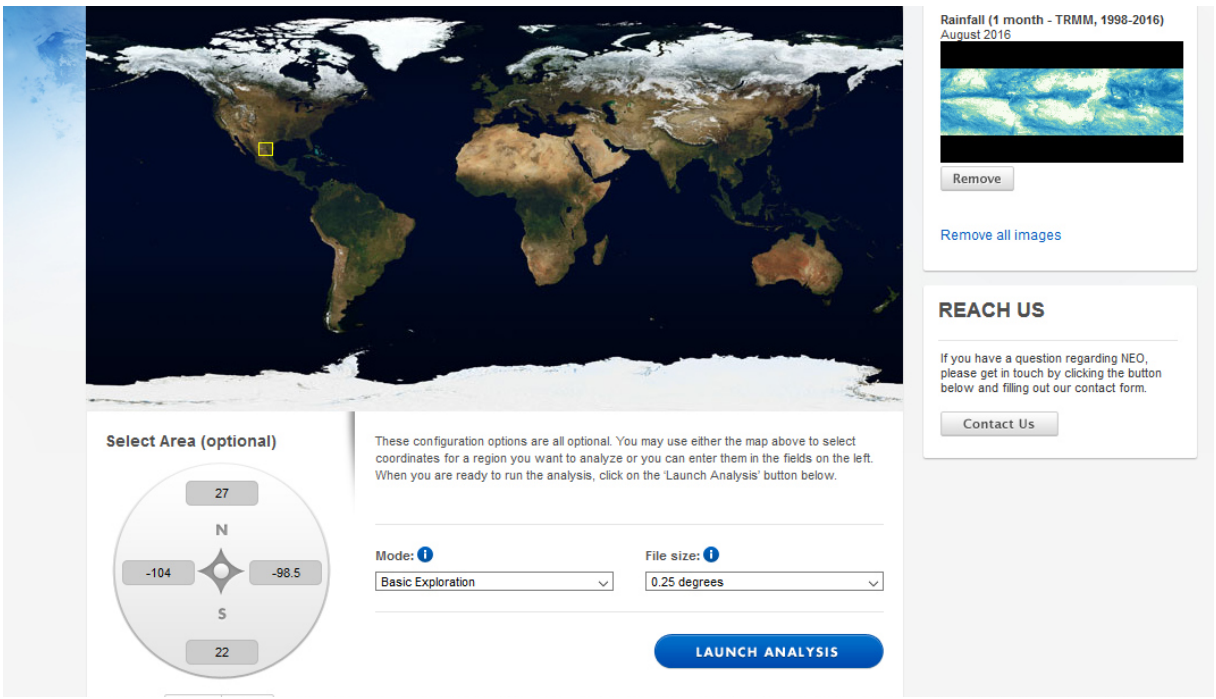



Figura 3.2: Portal NEO, área seleccionada

Un ejemplo del formato en el que se encuentran los datos se muestra en la figura 3.1. Estos son los datos del mes de enero del 2015. Aquí se ve que los datos de la parte superior tienen el valor 99999, esto indica que en estos puntos no se tiene medicion, lo cual es debido a que el satélite orbita cerca del ecuador. Solo se tiene valor para los datos que se encuentran entre las latitudes -49.875 y 49.875. Esto se ve en la figura 3.1


Figura 3.3: region estudiada

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | lat/lon | -179.875 | -179.625 | -179.375 | -179.125 | -178.875 | -178.625 | -178. |
| 2 | 89.875 | 99999 | 99999 | 99999 | 99999 | 99999 | 99999 | 99 |
| 3 | 89.625 | 99999 | 99999 | 99999 | 99999 | 99999 | 99999 | 99 |
| 4 | 89.375 | 99999 | 99999 | 99999 | 99999 | 99999 | 99999 | 99 |
| 5 | 89.125 | 99999 | 99999 | 99999 | 99999 | 99999 | 99999 | 99 |
| 6 | 88.875 | 99999 | 99999 | 99999 | 99999 | 99999 | 99999 | 99 |
| 7 | 88.625 | 99999 | 99999 | 99999 | 99999 | 99999 | 99999 | 99 |
| 8 | 88.375 | 99999 | 99999 | 99999 | 99999 | 99999 | 99999 | 99 |
| 9 | 88.125 | 99999 | 99999 | 99999 | 99999 | 99999 | 99999 | 99 |
| 10 | 87.875 | 99999 | 99999 | 99999 | 99999 | 99999 | 99999 | 99 |
| 11 | 87.625 | 99999 | 99999 | 99999 | 99999 | 99999 | 99999 | 99 |

Figura 3.4: Parte superior de los datos de enero 2015

Los datos descargados tienen formato de matriz, todas las obsevaciones se acomodaron como una tabla con 3 columnas; longitud x, latitud y, precipitación en xy, esto para facilitar su análisis. Una vez que se le dio este formato a los datos se extrajeron los que se encuentran entre las longitudes: -104 y - 98.5 y las latitudes: 22 y 27. Esto se realizó con el programa estadístico CRAN R en la versión 3.6.2 y este proceso se repitió para los datos de cada mes entre enero del 2015 y enero del 2016. Una gráfica de los datos seleccionados de enero del 2015 se presenta en la figura 3.1

Después de seleccionar los datos a usar se tienen 440 observaciones para cada mes, con 3 variables para cada observación, longitud, latitud y cantidad de lluvia.

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 160 | 50.375 | 99999 | 99999 | 99999 | 99999 | 99999 | 99999 | 999 |
| 161 | 50.125 | 99999 | 99999 | 99999 | 99999 | 99999 | 99999 | 999 |
| 162 | 49.875 | 12.22 | 10.84 | 8.26 | 5.25 | 4 | 4.66 | 5. |
| 163 | 49.625 | 17.03 | 15.1 | 8.77 | 5.58 | 4.38 | 4.13 | 4 |
| 164 | 49.375 | 19.22 | 13.79 | 10.2 | 6.49 | 5.58 | 5.75 | 14. |
| 165 | 49.125 | 19.22 | 16.53 | 10.2 | 6.49 | 7.1 | 14.21 | 18. |
| 166 | 48.875 | 19.81 | 17.03 | 12.22 | 8.51 | 10.51 | 12.6 | 11. |
| 167 | 48.625 | 24.46 | 21.04 | 15.1 | 12.6 | 21.68 | 18.65 | 5. |
| 168 | 48.375 | 25.21 | 21.04 | 17.03 | 15.1 | 24.46 | 19.22 | 22. |
| 169 | 48.125 | 27.6 | 20.41 | 19.81 | 13.38 | 16.04 | 16.53 | 34. |
| 170 | 47.875 | 33.08 | 21.04 | 16.53 | 16.04 | 16.53 | 15.56 | 20. |
| 171 | 47.625 | 25.21 | 17.03 | 11.51 | 12.99 | 17.56 | 19.81 | 23. |
| 172 | 47.375 | 27.6 | 19.22 | 11.51 | 12.22 | 21.68 | 24.46 | 30 |

Figura 3.5: Parte media de los datos de enero 2015

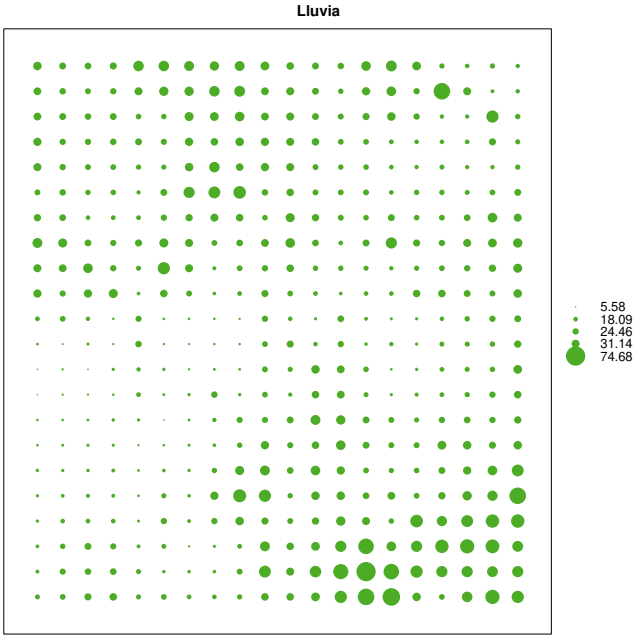



Figura 3.6: Datos seleccionados de enero 2015

## 3.2 Análisis Geoestadístico

Dados los datos de la región a trabajar, se realizó el proceso del kriging ordinario con los datos de cada mes y esto nos generó una aproximación a una superficie. La interpolación mediante el kriging se realizó con el paquete gstats 2.0-4 en el programa CRAN R. además de esto se obtuvierón gráficas de la varianza de las estimaciones. Estos procesos se pueden consultar en [36] y en [37]

## 3.3 Traslación y escalado de los datos

Una vez que se llevó a cabo el análisis geoestadístico para cada mes se realizó el análisis de persistencia de homología en los datos de cada mes, entonces se observó que el análisis tomaba mucho tiempo para generar resultados. Para solucionar este problema se realizó una transformación sobre los datos. En esta transformación, se escalaron los datos y se trasladaron, la transformación se realizó por columnas, donde cada dato de la columna j se transformó de la siguiente forma:

$$dato_{ij} = \frac{dato_{ij} - min_j}{max_j - min_j}$$

Aquí se usó la siguiente notación:

- $dato_{ij}$ es el valor en la columna j y en la fila i
- $min_j$ es el mínimo en la columna j
- $max_j$ es el máximo valor en la columna j

Un ejemplo de los datos antes de la transformación se presentan en la figura 3.3.

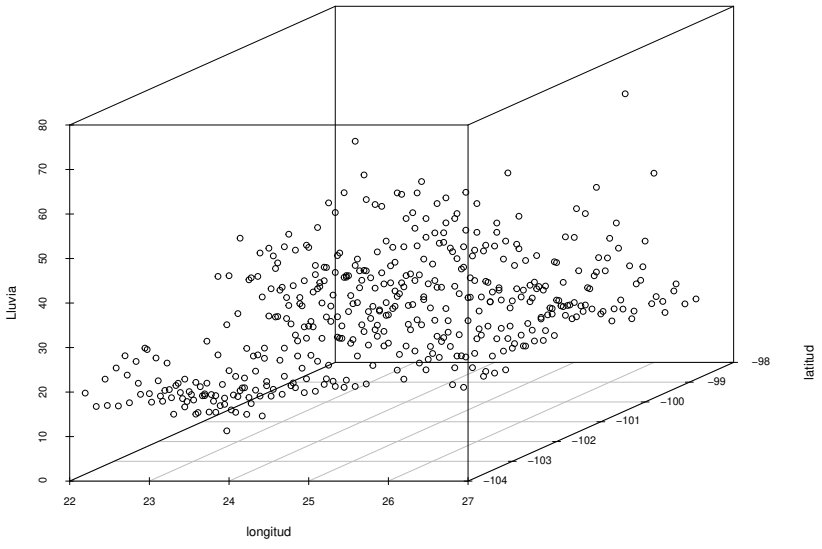



Figura 3.7: Datos del mes de enero del 2015

una vez que se realizó la transformación los datos mantienen la misma estructura general, como ejemplo, esto se observa en la figura 3.8

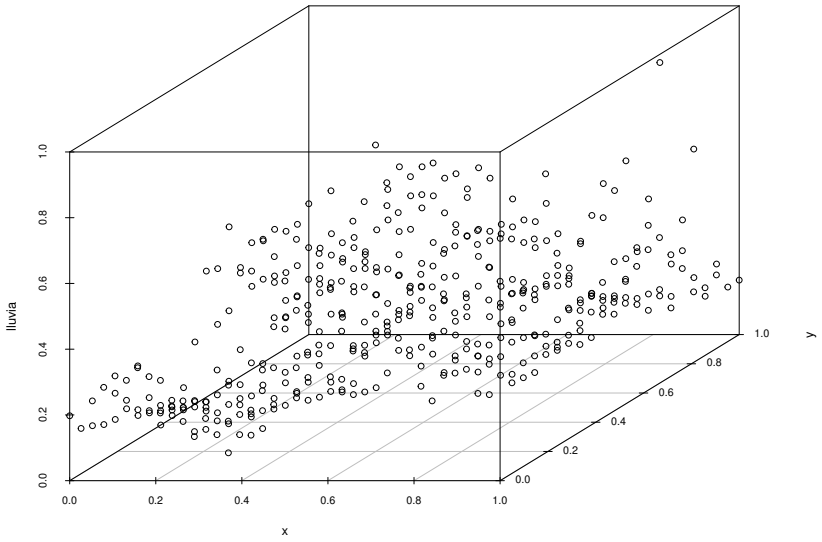



Figura 3.8: datos después de la transformación del mes de enero del 2015

## 3.4 Análisis Topológico de datos

Una vez que se tiene los datos transformados se llevó a cabo un análisis de persistencia de homología para cada mes, esto se realizó para las dimensiones 0, 1 y 2 pero se tiene como base una superficie en 3 dimensiones, entonces no se tienen huecos de dimensión 2. Entonces se obtuvo un código de barras y un diagrama de vida y muerte para cada mes del periodo trabajado, como ejemplo se tiene los gráficos 3.9 donde se presenta un código de barras para $H = 1$ y 3.10 donde se presenta un diagrama de vida y muerte para $H = 0$ y $H = 1$.

Lo anterior se hizo con la ayuda del paquete pHom de CRAN R en su versión 1.0.3, sus métodos se pueden consultar en [35]

Una vez que se obtuvieron los gráficos de vida y muerte se genero una superficie con altura constante y con los mismos valores en x, y que los datos después de la transformación, esta aparece en la figura 3.4.

A esta superficie con altura constante se le aplicó el análisis de persistencia de homología y se obtuvo su diagrama de vida y muerte. El diagrama de esta superficie sirvió como la clase de equivalencia del cero para los diagramas de vida y muerte en este espacio.

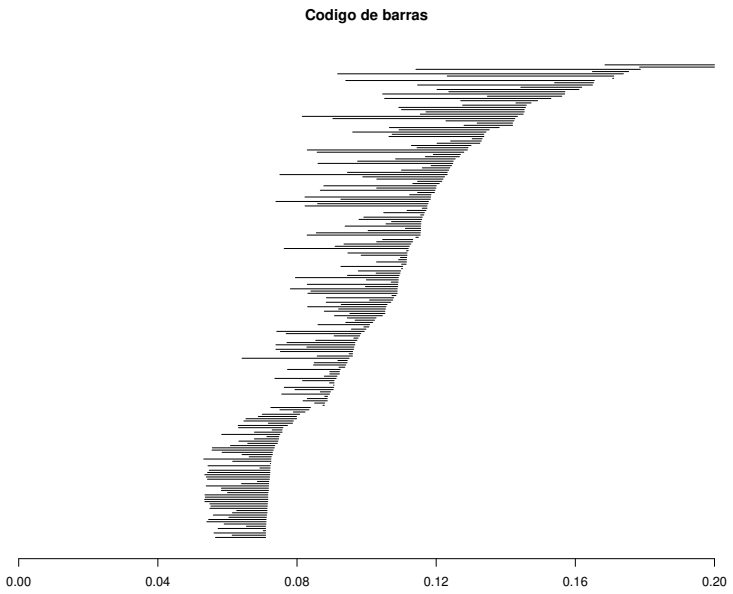



Figura 3.9: Código de barras del mes de enero del 2015

Con los gráficos de vida y muerte obtenidos se realizaron el análisis temporal de las siguientes 3 maneras:

- Se calculó la distancia de cuello de botella de meses contiguos, esto es, entre: enero - febrero, febrero - marzo, $\cdots$
- Se calculó la distancia de cuello de botella entre el diagrama cero y cada mes
- Se calculó el movimiento de la distancia entre cada mes, donde la distancia de un mes al siguiente es positiva si la distancia entre el siguiente mes y el mes cero es mayor que la distancia entre el mes actual y el mes cero, es negativa en caso contrario

El análisis temporal se realizó con el paquete TDA en su versión 1.6.9 de CRAN R. este se puede consultar en [38]

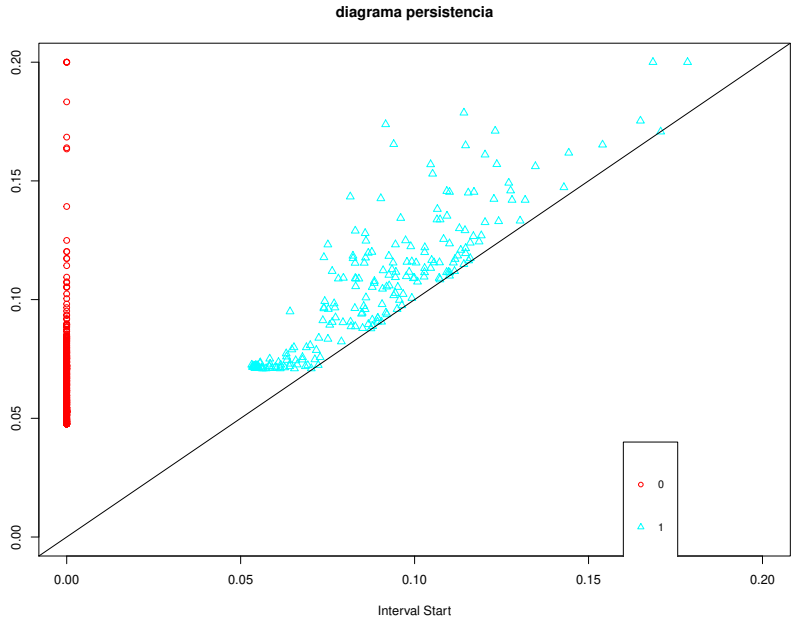



Figura 3.10: Gráfica de vida y muerte del mes de enero del 2015

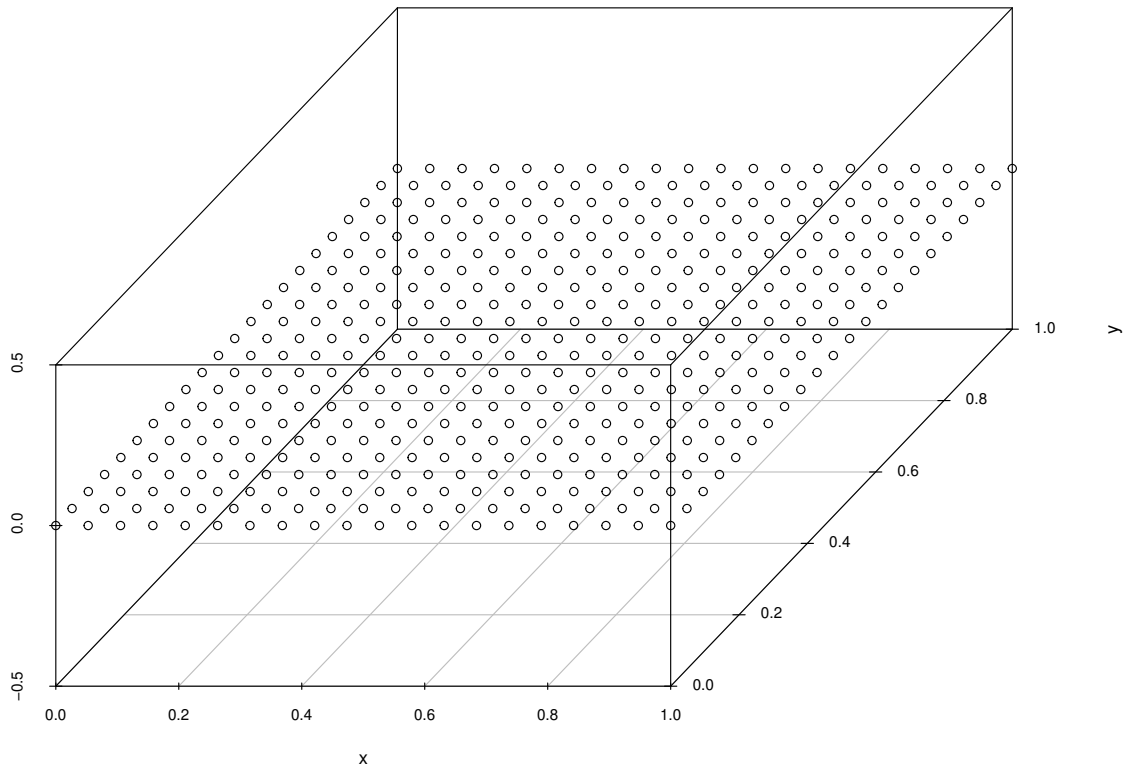


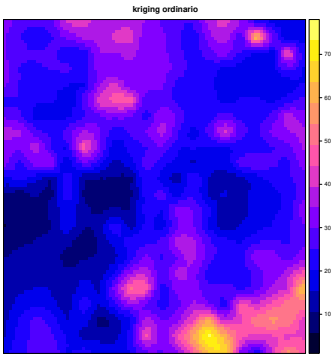

Figura 3.11: Superficie con altura constante

Capítulo 4

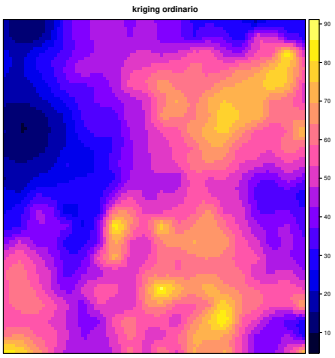
# Resultados

## 4.1 Análisis Geoestadístico

En esta sección se presentan los gráficos de calor obtenidos mediante el kriging ordinario




a. Enero 2015




b. Febrero 2015

Figura 4.1:

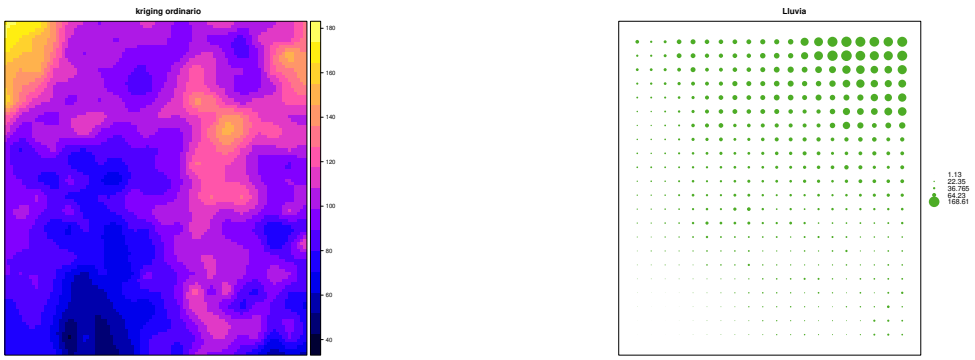
En los meses de abril y julio no se presenta el resultado del kriging ordinario debido a que el paquete arrojó los siguientes errores:

para Abril:

Error in seq.default(zrng[1], zrng[2], length.out = cuts + 2) : 'from' must be a finite number

para Julio:

Error in seq.default(zrng[1], zrng[2], length.out = cuts + 2) : 'from' must be a finite number




a. Marzo 2015 b. Abril 2015

Figura 4.2:

## 4.2 Análisis de persistencia

A continuación se presentan los diagramas de vida y muerte de los meses estudiados y del mapa cero

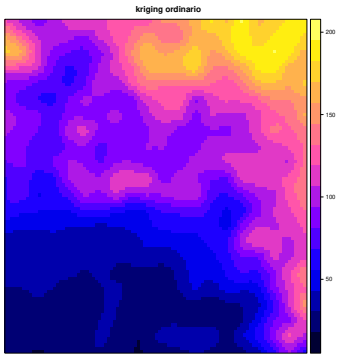
Aquí cabe resaltar que el análisis de persistencia de homología no tiene problema encontrando la gráfica de vida y muerte de los meses de abril y julio.

## 4.3 Análisis temporal

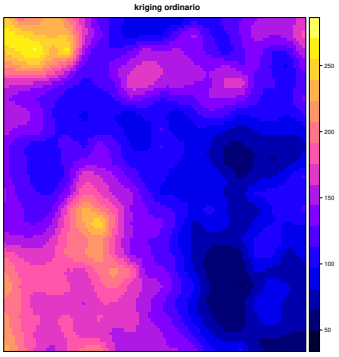
En esta sección se presenta el resultado de los análisis temporales, para los 3 tipos de análisis se utilizó la distancia de cuello de botella entre los diagramas de vida y muerte obtenidos para los datos de cada mes y del diagrama cero, se utilizo esta distancia por ser la más simple de implementar y por ser el default en el paquete con el que se trabajó, los resultados se presenta en la forma de 3 gráficas.

En la gráfica 4.15 El eje x indica la interacción entre meses contiguos, esto es, 1 indica enero-febrero, 2 febrero-marzo, etc. El eje y indica la distancia de cuello de botella entre los diagramas de vida y muerte de meses contiguos.

En la gráfica 4.16 el eje x indica la interacción entre un mes y la superficie con altura constante y el eje y indica la distancia de cuello de botella entre el diagrama de vida y muerte del mes indicado por el eje x y el diagrama de vida y muerte de la superficie constante.

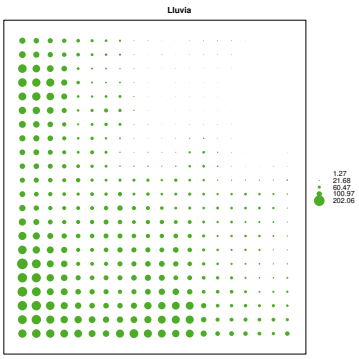



a. Mayo 2015

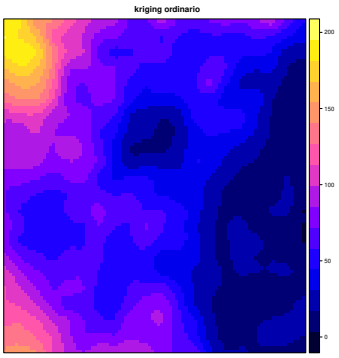



b. Junio 2015

Figura 4.3:




a. Julio 2015

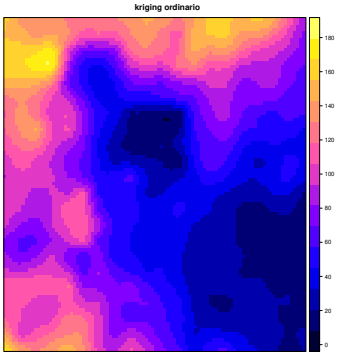



b. Agosto 2015

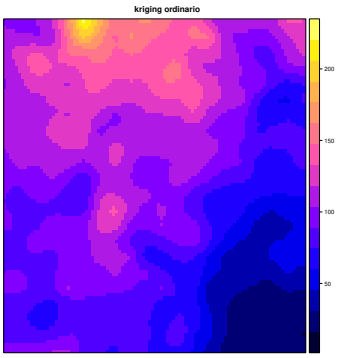
Figura 4.4:

En la gráfica 4.17 el eje x indica el mes, la altura en 1, es la distancia entre el diagrama de vida y muerte de la superficie constante, la altura en un valor de x es la altura en x - 1 y se le suma o se le resta la distancia entre los diagramas de vida y muerte de x con x-1 dependiendo de:

- se suma si la distancia entre x y la superficie constante es mayor o igual que la distancia entre x-1 y la superficie constante.
- se resta si la distancia entre x y la superficie constante es menor que la distancia entre x-1 y la superficie constante.

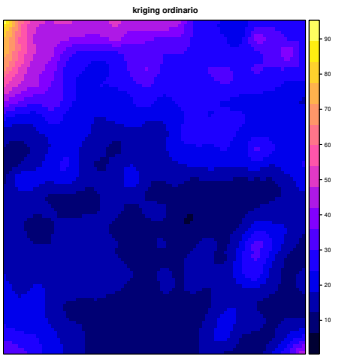



a. Septiembre 2015

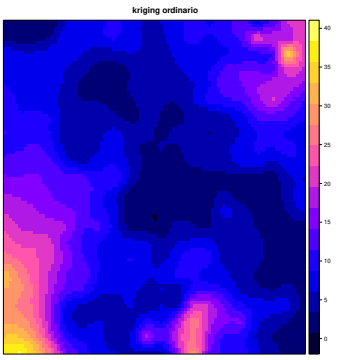



b. Octubre 2015

Figura 4.5:

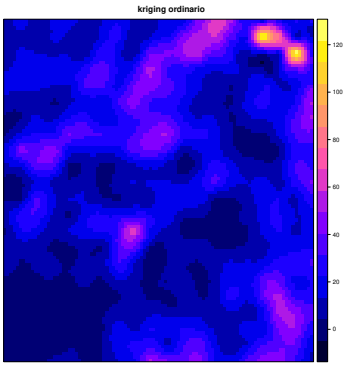



a. Noviembre 2015



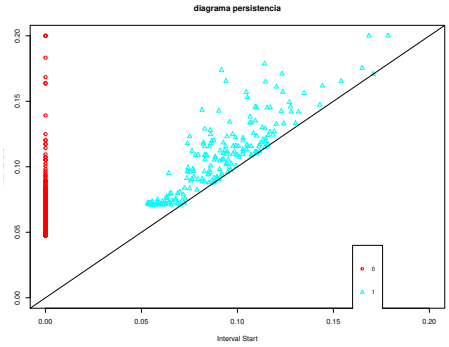

b. Diciembre 2015

Figura 4.6:

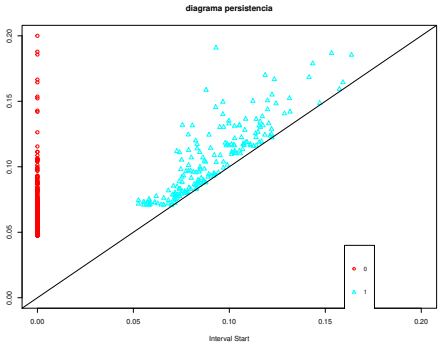



Enero 2016

Figura 4.7:

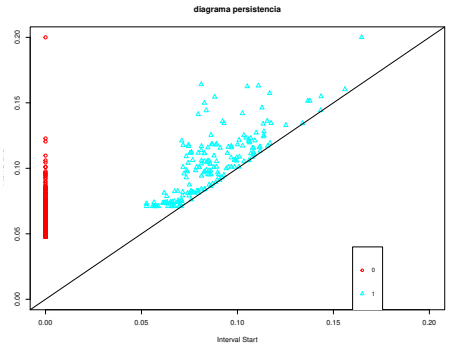



a. Enero 2015

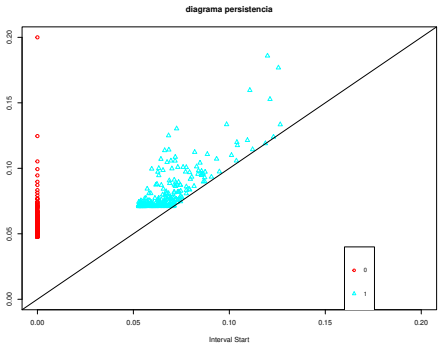



b. Febrero 2015

Figura 4.8:

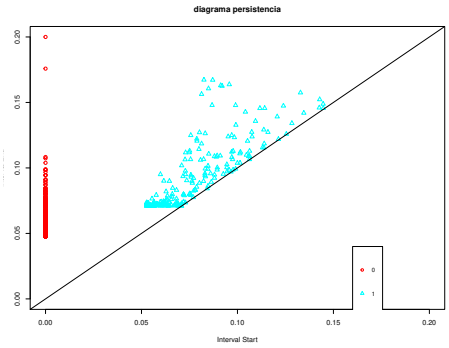



a. Marzo 2015

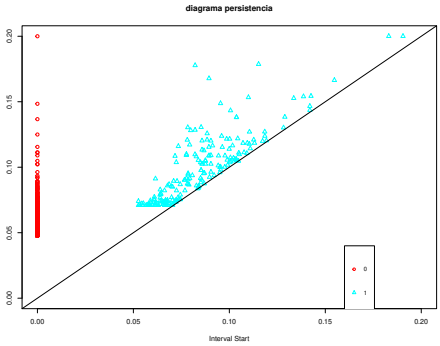



b. Abril 2015

Figura 4.9:

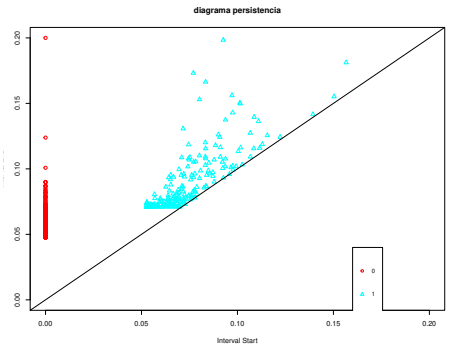



a. Mayo 2015

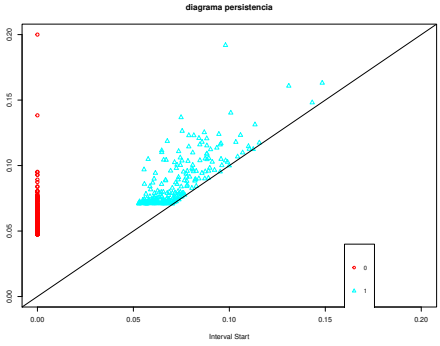



b. Junio 2015

Figura 4.10:

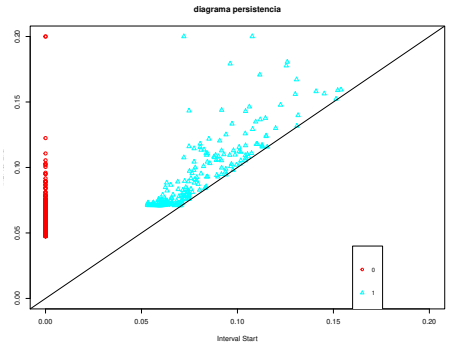



a. Julio 2015

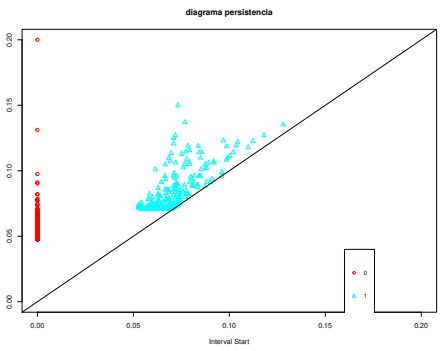



b. Agosto 2015

Figura 4.11:

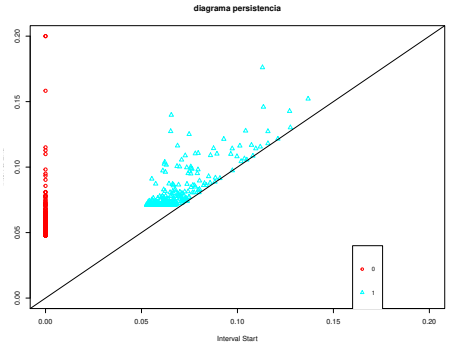



a. Septiembre 2015

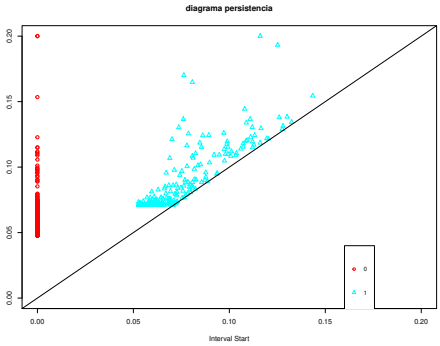



b. Octubre 2015

Figura 4.12:

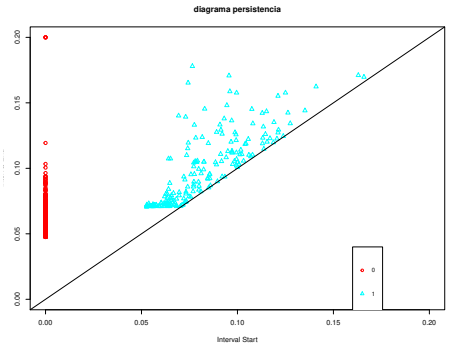



a. Noviembre 2015

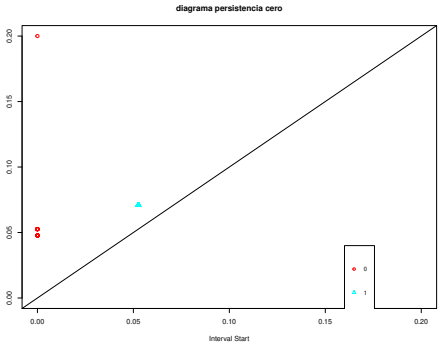



b. Diciembre 2015

Figura 4.13:




a. Enero 2016




b. Diagrama cero

Figura 4.14:

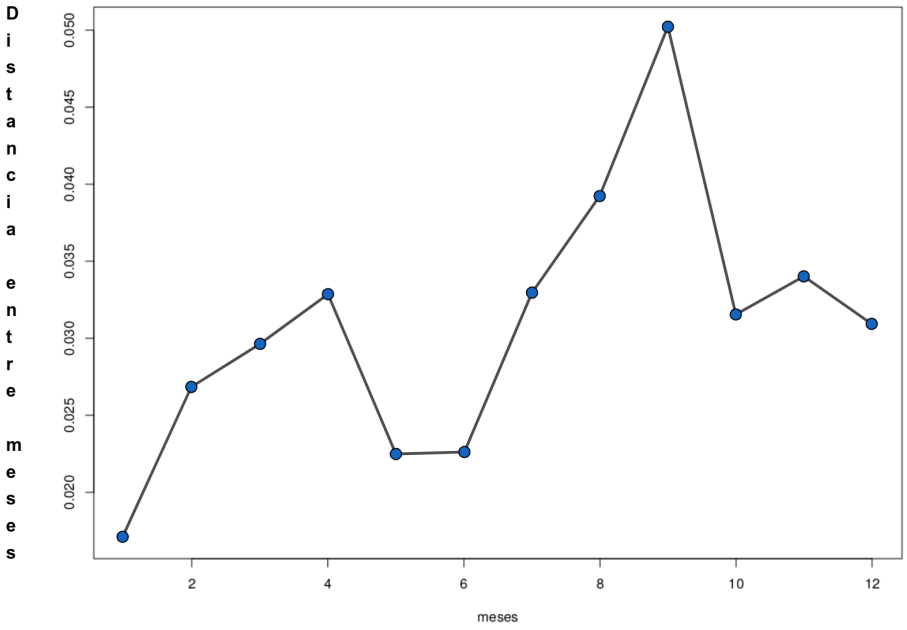



Figura 4.15: Distancia cuello de botella entre meses

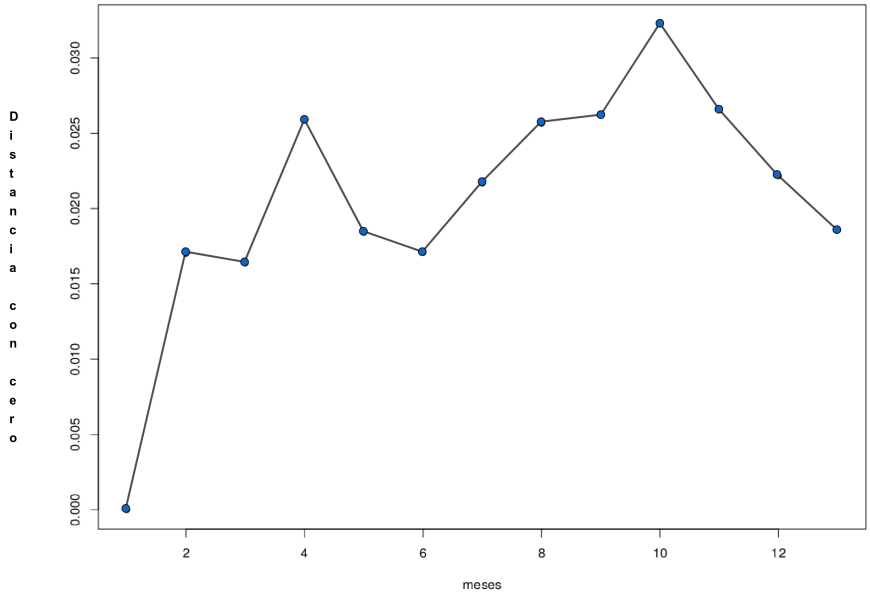



Figura 4.16: Distancia cuello de botella entre meses y superficie constante

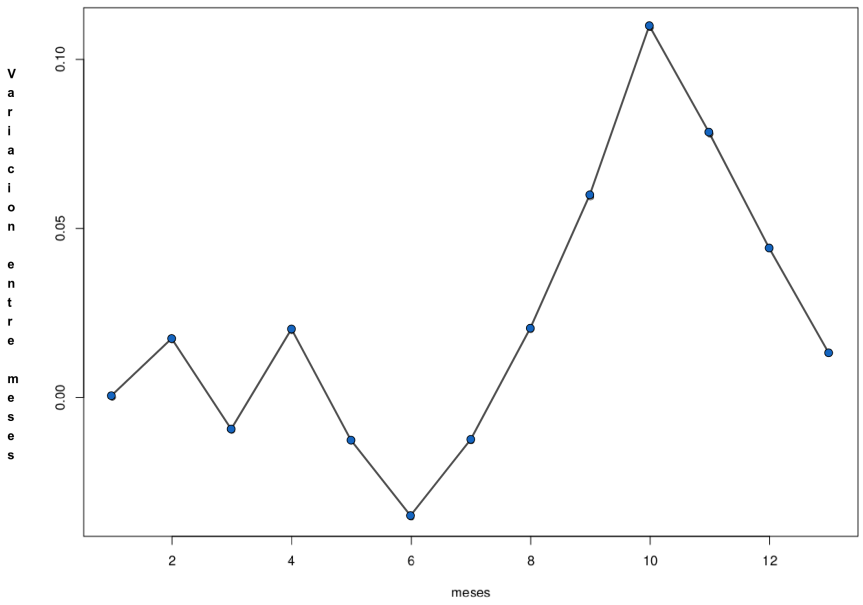



Figura 4.17: Variación entre meses y superficie constante

Capítulo 5

# Conclusiones

## 5.1 Análisis Geoestadístico y de Persistencia Homológica

Al implementar el método del kriging ordinario se observa en la sección 4.1 que en el mes de abril, gráfica 4.2 a y el mes de julio, gráfica 4.4 a, del 2015, el paquete standard de geoetadística de R no pudo aplicar el método a los datos correspondientes a estos meses, esto debido a que no se pudo resolver el sistema 2.4 correspondiente a estos meses.

En la sección 4.2 se observa que la persistencia de homología funciona en todos los datos de todos los meses, esto es, el método, persistencia de homología puede ser usado como opción al método geoestadístico del kriging ordinario aun cuando este no proporciona una solucion.

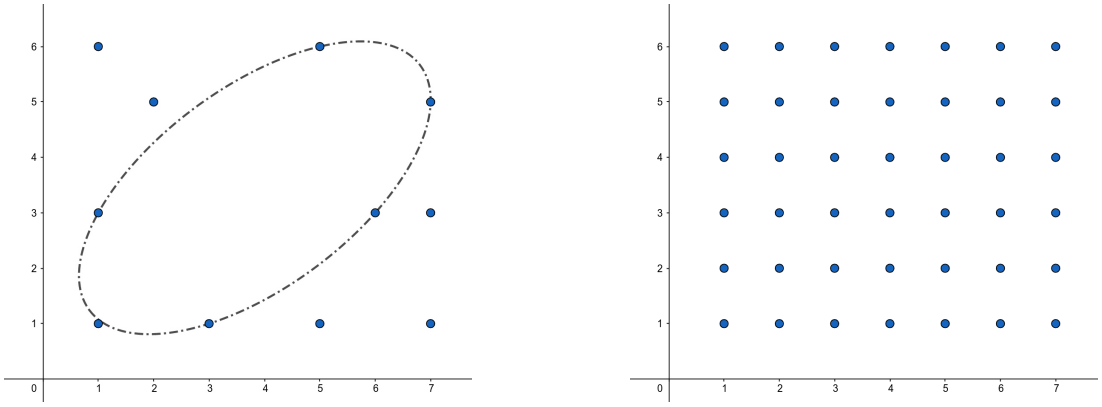
En el mes de octubre del 2015 la región estudiada fue afectada por el huracán patricia esto se ve reflejado en las gráficas de la sección 4.3 donde:

- En la gráfica 4.15 que muestra la distancia entre meses, se observa un pico en 9 que indica la distancia entre septiembre y octubre
- En la gráfica 4.16, distancia entre meses con cero, se observa un pico en 10 que es la distancia entre el gráfico de vida y muerte y el diagrama cero.
- En la gráfica 4.17 que muestra la variación entre meses se observa de nuevo un pico en el mes de octubre

## 5.2 Concluciones

Se observó que los métodos del kriging ordinario y el de persistencia de homología se pueden complementar de las siguientes maneras:

- En los casos trabajados se tiene que los datos están bien distribuidos espacialmente, en este caso, al realizar el análisis, persistencia de homología, la presencia de ciclos indica variaciones bruscas en la variable asociada a los puntos cercanos espacialmente, en este caso, la lluvia observada, entonces, en caso de que los puntos no tengan una distribución espacial adecuada, el análisis podría arrojar ciclos que indican zonas donde no se han tomado muestras (como en figura 5.1 a), para evitar esto, una opción es utilizar el kriging ordinario para que así las observaciones sean más regulares (como en la figura 5.1 b) y que los ciclos observados sean provocados por variaciones bruscas en la variable asociada a los puntos.

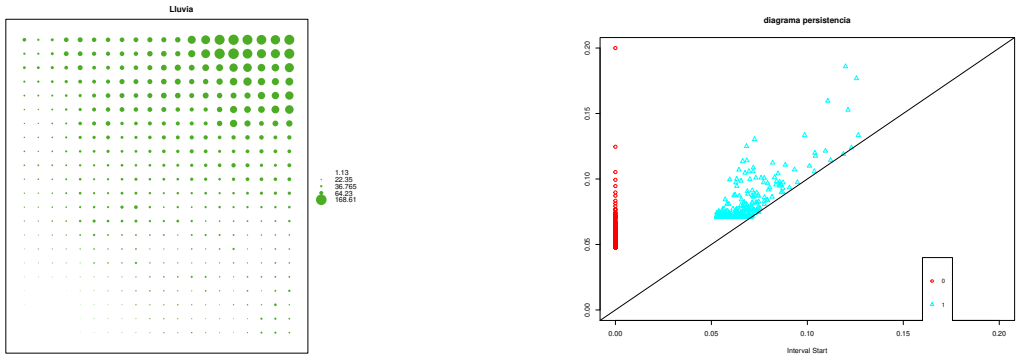


a b

Figura 5.1: Uso de kriging para tener datos regulares

- En el caso de que el kriging ordinario no nos arroje un resultado (como en figura 5.2 a) o que la varianza observada sea grande, se puede aplicar el análisis, persistencia de homología (figura 5.2 b), el cual nos indicará en cuantos puntos persiste una variación grande entre los valores asociados a los puntos. Esto se tiene que tratar con cuidado, pues si los puntos no están distribuidos a intervalos regulares, tenemos que descartar los ciclos que se presenten debido a la separación espacial.

- Debido a que el kriging ordinario es un análisis que se realiza en un instante en el tiempo, es una opción realizar un análisis de persistencia de homología para observar como cambian las condiciones del terreno en el tiempo (como se ve en las figuras 4.15, 4.16, 4.17).




a b

Figura 5.2: Resultados para Abril del 2015

## 5.3 Trabajo a futuro

El análisis topológico de datos es una herramienta reciente y unas de las posibles direcciones a las que se puede enfocar, en conjunto con el análisis espacial de datos son las siguientes:

- En las bases de datos atmosféricos se suelen encontrar múltiples cantidades de variables asociadas a un punto en el espacio, el análisis topológico puede tomar en cuenta todas estas variables como descripciones de un punto, faltaria buscar la interpretación de los resultados.
- Al tener más variables asociadas a los mismos puntos se puede realizar diferentes tipos de análisis geoestadísticos que toman en cuenta la correlación entre diferentes variables para realizar la estimación
- En el portal de datos neo, donde se encontraron los datos usados, también se encuentran los datos de otros años, con estos se puede investigar si existe algun patron que se presente por temporadas.
- El análisis de persistencia de homología, aplicado en este trabajo, nos indica en cuantas ubicaciones existen cambios bruscos con respecto a la lluvia observada en puntos cercanos, pero no nos indica donde suceden estas variaciones bruscas, con base a los $\epsilon$ donde aparecen estas variaciones, es posible reconstruir esos ciclos para identificar estas zonas.
- El análisis temporal se realizó utilizando la distancia de cuello de botella entre los diagramas de vida y muerte por ser la más común y la que está por default en el

paquete de R, es posible realizar esta comparación utilizando diferentes valores para la distancia de Wasserstein y analizar si alguna en entrega un valor que refleje mejor las variaciones observadas entre los meses.

# Bibliografía

[1] Hu, Q., Li, Z., Wang, L., Huang, Y., Wang, Y., & Li, L. (2019). Rainfall Spatial Estimations: a review from spatial interpolation to multi-source data merging. Water, 11(3), 579.

[2] Carrera-Hernández, J. J., & Gaskin, S. J. (2008). The Basin of Mexico Hydrogeological Database (BMHDB): Implementation, queries and interaction with open source software. Environmental Modelling & Software, 23(10-11), 1271-1279.

[3] Carrera-Hernández, J. J., & Gaskin, S. J. (2007). Spatio temporal analysis of daily precipitation and temperature in the Basin of Mexico. Journal of Hydrology, 336(3-4), 231-249.

[4] Qiao, P., Lei, M., Yang, S., Yang, J., Guo, G., & Zhou, X. (2018). Comparing ordinary kriging and inverse distance weighting for soil as pollution in Beijing. Environmental Science and Pollution Research, 25(16), 15597-15608.

[5] Chen, T., Ren, L., Yuan, F., Yang, X., Jiang, S., Tang, T., ... & Zhang, L. (2017). Comparison of spatial interpolation schemes for rainfall data and application in hydrological modeling. Water, 9(5), 342.

[6] Adhikary, S. K., Muttil, N., & Yilmaz, A. G. (2017). Cokriging for enhanced spatial interpolation of rainfall in two Australian catchments. Hydrological processes, 31(12), 2143-2161.

[7] Li, J., & Heap, A. D. (2014). Spatial interpolation methods applied in the environmental sciences: A review. Environmental Modelling & Software, 53, 173-189.

[8] Arun, P. V. (2013). A comparative analysis of different DEM interpolation methods. The Egyptian Journal of Remote Sensing and Space Science, 16(2), 133-139.

[9] Berndt, C., & Haberlandt, U. (2018). Spatial interpolation of climate variables in Northern Germany-Influence of temporal resolution and network density. Journal of Hydrology: Regional Studies, 15, 184-202.

[10] Matheron, G. (1965). Les variables rgionalisées et leur estimation: une application de la théorie des fonctions aléatoires aux sciences de la nature. Masson et CIE.

[11] Sarma, D. D. (2010). Geostatistics with applications in earth sciences. Springer Science & Business Media.

[12] Oliver, M. A., & Webster, R. (2015). Basic steps in geostatistics: the variogram and kriging (pp. 15-42). New York, NY: Springer International Publishing.

[13] Xiao, N. (2015). GIS algorithms. Sage.

[14] Edelsbrunner, H., & Morozovy, D. (2014). Persistent homology: theory and practice.

[15] Oliver, M. A., & Webster, R. (2014). A tutorial guide to geostatistics: Computing and modelling variograms and kriging. Catena, 113, 56-69.

[16] Munch, E. (2017). A user's guide to topological data analysis. Journal of Learning Analytics, 4(2), 47-61.

[17] Otter, N., Porter, M. A., Tillmann, U., Grindrod, P., & Harrington, H. A. (2017). A roadmap for the computation of persistent homology. EPJ Data Science, 6(1), 17.

[18] Ghrist, R. (2008). Barcodes: the persistent topology of data. Bulletin of the American Mathematical Society, 45(1), 61-75.

[19] Giusti, C., Ghrist, R., & Bassett, D. S. (2016). Two's company, three (or more) is a simplex. Journal of computational neuroscience, 41(1), 1-14.

[20] Ghrist, R. (2008). Three examples of applied and computational homology. Lab Papers (GRASP), 18.

[21] Carstens, C. J., & Horadam, K. J. (2013). Persistent homology of collaboration networks. Mathematical problems in engineering, 2013.

[22] Maletić, S., Zhao, Y., & Rajković, M. (2016). Persistent topological features of dynamical systems. Chaos: An Interdisciplinary Journal of Nonlinear Science, 26(5), 053105.

[23] Ghrist, R. W. (2014). Elementary applied topology (Vol. 1). Seattle: Createspace.

[24] Boissonnat, J. D., Chazal, F., & Yvinec, M. (2018). Geometric and topological inference (Vol. 57). Cambridge University Press.

[25] Hatcher, A. (2002). Algebraic topology, Cambridge Univ. Press, Cambridge xii.

[26] Zomorodian, A. J. (2005). Topology for computing (Vol. 16). Cambridge university press.

[27] Dey, T. K., Edelsbrunner, H., & Guha, S. (1999). Computational topology. Contemporary mathematics, 223, 109-144.

[28] Munkres, J. (2013). Topology: Pearson New International Edition. Pearson.

[29] Mendelson, B. (1990). Introduction to topology. Courier Corporation.

[30] Saveliev, P. (2016). Topology Illustrated. Peter Saveliev.

[31] Herstein, I. N. (2006). Topics in algebra. John Wiley & Sons.

[32] Pivato, M. (2003). Visual Abstract Algebra.

[33] Rudin, W. (1964). Principles of mathematical analysis (Vol. 3). New York: McGraw-hill.

[34] Kreyszig, E. (1978). Introductory functional analysis with applications (Vol. 1). New York: wiley.

[35] Tausz, A. (2013). The phom package: User's manual.

[36] Pebesma, E. (2019). The meuse data set: a brief tutorial for the gstat R package.

[37] Antonanzas-Torres, F. (2014). Geostatistics examples in R: ordinary kriging, universal kriging and inverse distance weighted.

[38] Fasy, B. T., Kim, J., Lecci, F., & Maria, C. (2014). Introduction to the R package TDA. arXiv preprint arXiv:1411.1830.

[39] Kummerow, C., Barnes, W., Kozu, T., Shiue, J., & Simpson, J. (1998). The tropical rainfall measuring mission (TRMM) sensor package. Journal of atmospheric and oceanic technology, 15(3), 809-817.

[40] Simpson, J., Adler, R. F., & North, G. R. (1988). A proposed tropical rainfall measuring mission (TRMM) satellite. Bulletin of the American meteorological Society, 69(3), 278-295.

[41] Thiele, O. W. (1987). On requirements for a satellite mission to measure tropical rainfall. Nasa Reference Publication

# Índice de figuras

# Índice de tablas