ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

ADAPTACIÓN DEL MÉTODO DE AGRUPACIÓN K-MEDIAS PARA DATOS FUNCIONALES CORRELACIONADOS ESPACIALMENTE CON APLICACIÓN A DATOS DEL ÍNDICE DE VEGETACIÓN DE DIFERENCIA NORMALIZADA DE LOS PÁRAMOS DEL ECUADOR.

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

PROYECTO DE INVESTIGACIÓN

JEYSSON FABRICIO CHUQUÍN CHUQUÍN
jeysson.chuquín@epn.edu.ec

SHARON ALEXANDRA MAIGUA CASTILLO
sharon.maigua@epn.edu.ec

DIRECTOR: MIGUEL ALFONSO FLORES SÁNCHEZ
miguel.flores@epn.edu.ec

Quito, marzo, 2022

!

DECLARACIÓN

Nosotros, JEYSSON FABRICIO CHUQUÍN CHUQUÍN y SHARON ALEXANDRA MAIGUA CASTILLO, declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que o ha sido previamente presentado para ningún grado o calificación profesional; y, que hemos consultado las referencias bibliográficas que se incluyen en este documento.

La Escuela Politécnica Nacional puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normativivdad institucional vigente.

JEYSSON FABRICIO CHUQUÍN CHUQUÍN

SHARON ALEXANDRA MAIGUA CASTILLO

!

!

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por JEYSSON FABRICIO CHUQUÍN CHUQUÍN y SHARON ALEZANDRA MAIGUA CASTILLO, bajo mi supervisión.

MIGUEL ALFONSO FLORES SÁNCHEZ, PhD.

DIRECTOR

!

AGRADECIMIENTOS

A Dios, por todo.

A mi hermano Byron por ser mi inspiración.

A mis padres Manuel y Aida que con su esfuerzo y dedicación me ayudaron a culminar mi carrera universitaria.

A mis amigos Alex y Marlon por todo lo vivido en la carrera.

A Miguel por brindarme una nueva perspectiva sobre mi futuro profesional.

A Alexa por su paciencia y apoyo incondicional.

Jeysson.

DEDICATORIA

A toda mi familia por haber sido mi apoyo a lo largo de mi vida, y especialmente a mi abuelita Isabel, por ser la persona que me inspira a seguir mejorando día a día y continuar alcanzado mis objetivos.

Jeysson.

AGRADECIMIENTOS

A mis padres Sandra y Alejandro que con su apoyo incondicional, esfuerzo y dedicación me inspiran para no rendirme.

A mis hermanas Sthefany, Priscila y Cristina por los momentos de alegría, los consejos brindados y sus palabras de aliento.

A Miguel por su apoyo y guía en el transcurso de la carrera.

A Jeysson compañero de trabajo de titulación y amigo por las noches llenas de risas y desvelos.

Alexandra.

DEDICATORIA

Dedico este trabajo con todo cariño y amor a mis padres y hermanas por haberme brindado apoyo incondicional a lo largo de mi vida. A todas las personas que aportaron de alguna manera a mi formación profesional y como ser humano. Y en especial, a mis abuelitos que me cuidan desde donde quiera que se encuentren.

Alexandra.

AGRADECIMIENTOS GENERALES

Los autores agradecen los datos otorgados por el proyecto PIJ-17-05 financiado por la Escuela Politécnica Nacional sobre “Los patrones climáticos globales y su influencia en la respuesta temporal y espacial de índices espectrales de la vegetación del páramo en el Ecuador”.

Al Laboratorio Nacional de Cálculo Científico por facilitar el uso del hardware y software del HPC-MODEMAT.

A todos los profesores y personal del Departamento de Matemática de la Escuela Politécnica Nacional, que con su sabiduría, conocimiento y apoyo, motivaron a desarrollarnos como buenas personas y excelentes profesionales.

A Jorge Mateu profesor del departamento de matemáticas de la Universad Jaume I, por sus valiosas observaciones y correcciones que hicieron posible el avance del presente trabajo de titulación.

A Sandra Torres investigadora del Instituto Nacional de Meteorología e Hidrología (INAMHI), por su criterio experto en el área de aplicación.

Finalmente a Miguel Flores, por su aporte en la realización del presente trabajo de titulación, por la guía académica con su experiencia y profesionalismo y por las oportunidades brindadas.

Índice general

Índice de figuras

Índice de cuadros

Resumen

La dependencia espacial en datos medio ambientales es un criterio influyente en procesos de agrupación, dado que los resultados obtenidos brindan información relevante. Como los métodos clásicos no consideran la dependencia espacial, al considerar esta estructura se producen resultados inesperados, proporcionando agrupaciones de curvas que pueden no ser similares en forma y/o comportamiento. En este trabajo se realiza la agrupación mediante el método k-medias modificado para datos funcionales correlacionados espacialmente aplicado a datos del Índice de Vegetación de Diferencia Normalizada (NDVI) de los páramos del Ecuador. Para esto se implementan índices de calidad que permitan obtener el número adecuado de grupos. Con base en la metodología desarrollada en el método de clasificación jerárquico para datos funcionales con correlación espacial, y dado que los datos funcionales pertenecen al espacio de Hilbert de funciones cuadrado-integrables; se desarrolla el análisis considerando la distancia entre curvas a través de la norma $\mathcal{L}^2$, obteniendo una representación reducida de los datos a través de una base finita de tipo Fourier. Luego, se calcula el variograma empírico y se ajusta a un modelo teórico para así ponderar la matriz de distancia entre las curvas por el trazo-variograma y variograma multivariado calculado con los coeficientes de las funciones base, y esta matriz se utiliza para la agrupación de datos funcionales correlacionados espacialmente. Para la validación del método se realizaron varios escenarios de simulación, obteniéndose buenos resultados de correcta clasificación y se complementa con un caso de aplicación a datos del NDVI, donde se obtuvieron cinco regiones distribuidas latitudinalmente.

Capítulo 1

Introducción

Los métodos de agrupación tienen por objetivo identificar grupos homogéneos de observaciones que representan la realización de alguna variable aleatoria X [Jacques, 2014], y en los cuales los miembros dentro de los grupos tienden a ser similares y distintos de otros miembros de otros grupos [Romano *et al.*, 2015]. Estos métodos también son conocidos como métodos de clasificación no supervisada, pues las observaciones no tienen asociadas ninguna etiqueta de grupo o categoría de forma previa [Jain, 2010]; por lo cual, este tipo de métodos se utilizan con frecuencia para detectar patrones especiales en una base de datos, los cuales pueden tener interpretaciones convenientes por parte del usuario [Jacques, 2014].

Uno de los métodos más populares de agrupación es el algoritmo k-medias, pues desde 1955 se ha venido utilizando y adaptando a miles de problemas en diferentes campos de las ciencias [Jain, 2010]. Muchas de estas adaptaciones asumen que las observaciones de tipo escalar, espacial y multivariante, pueden ser representadas como puntos en un espacio Euclídeo de dimensión finita, pero cuando la dimensión del objeto es sumamente grande las realizaciones de la variable aleatoria X pueden tomar valores en un espacio infinito dimensional [Jacques, 2014]; este último tipo de dato es muy común, ya que hoy en día se pueden realizar mediciones en cada instante de tiempo y es conocido como dato funcional. Es así que, este problema se puede abordar desde el análisis de datos funcionales, pues su objetivo es inferir la estructura de los datos trabajando con su naturaleza de dimensión infinita sobre espacios de Hilbert [Luz López García *et al.*, 2015].

Dado que el campo de aplicación es amplio para el uso de estos métodos, nos centraremos en el área medioambiental, pues en esta área se trabaja con datos temporales, los cuales según [Romano *et al.*, 2017], también son considerados como datos funcionales. Aplicar un método de agrupación, omitiendo su estructura espacial, dará como resultado grupos en los cuales las curvas pueden no ser similares en forma y/o comportamiento [Romano *et al.*, 2017]. Para abordar este problema, hasta el momento se han propuesto métodos de agrupación como el jerárquico [Giraldo *et al.*, 2012] y el dinámico [Romano *et al.*, 2017], los cuales usan el trazo-variograma como medida de dependencia espacial, con la diferencia de que en el método jerárquico se realiza el cálculo de manera global, mientras que en el método dinámico se lo realiza de manera local; así, el método jerárquico usa la distancia euclidiana entre las curvas, mientras que el método dinámico usa la distancia euclidiana al cuadrado entre las curvas dentro de cada grupo [Romano *et al.*, 2015]. Al notar la escasez de opciones de métodos, así como la no implementación del método k-medias, contar con más opciones de agrupación que trabajen con este tipo de datos serán de gran ayuda; es por ello que en este trabajo se implementará el

método k-medias para datos funcionales con correlación espacial, para lo cual se usará la metodología aplicada en [Giraldo *et al.*, 2012], y para evaluar el desempeño del método propuesto, se realizarán varios escenarios de simulación, y posteriormente aplicar este método a los datos del NDVI de los páramos del Ecuador.

1.1. Objetivos

1.1.1. Objetivo General

Adaptar el método K-medias para datos funcionales correlacionados espacialmente con aplicación a datos del Índice de Vegetación de Diferencia Normalizada de los páramos del Ecuador.

1.1.2. Objetivos Específicos

1. Adaptar el método funcional espacial de agrupación jerárquico para la generación de la metodología del algoritmo K-medias.

2. Simular varios escenarios para poner a prueba el método K-medias modificado.

3. Aplicar índices de calidad para poder conocer el número de grupos con el que se debería trabajar.

4. Comparar los resultados de clasificación de las simulaciones del método jerárquico funcional espacial con el método k-medias modificado.

5. Aplicar el método K-medias modificado a los datos de Índice de Vegetación de Diferencia Normalizada de los páramos del Ecuador.

1.2. Justificación

El análisis de agrupación funcional tiene como objetivo construir grupos homogéneos dentro de un conjunto de curvas. Teniendo en cuenta la naturaleza de los datos y que la implementación de los métodos probabilísticos es compleja, esta técnica se vuelve complicada [Pérez, 2018]. Los métodos clásicos de agrupación no consideran la dependencia espacial; por ende, pueden agrupar miembros con errores en cuanto a la ubicación espacial de los mismos. Resulta interesante considerar la estructura de la correlación espacial en un contexto de agrupación funcional ya que puede llegar a producir resultados inesperados [Romano *et al.*, 2017].

Para clasificar datos funcionales con dependencia espacial se utiliza el variograma como una medida de correlación espacial e indica la forma en que un punto tiene influencia sobre otro punto dependiendo de su distancia [Giraldo *et al.*, 2012]. Entonces, como los datos funcionales con los que se trabaja pertenecen al espacio de Hilbert de funciones cuadrado-integrables, se desarrolla el análisis considerando la distancia entre curvas a través de la norma $\mathcal{L}^2$. Después se obtiene una representación reducida de los datos a través de una base finita, que puede ser de tipo: Fourier, *B-splines*, *Wavalets*, entre otras. Luego, se calcula el variograma empírico y se ajusta un modelo teórico para así ponderar la matriz de distancias entre las curvas por el trazo-variograma o variograma

multivariado calculado con los coeficientes de las funciones de la base. El proceso de agrupación de datos funcionales correlacionados espacialmente se lleva acabo mediante la matriz ponderada, utilizando como función de peso el trazo-variograma o variograma multivariado y la distancia entre las curvas [Giraldo *et al.*, 2012].

Los métodos de agrupación para datos funcionales con correlación espacial son de gran utilidad en la práctica y sus aplicaciones son variadas, en [Giraldo *et al.*, 2012], se hizo uso del método de agrupación jerárquico aplicado a datos de temperaturas del aire de Canadá y se logró clasificar e identificar zonas en las cuales las temperaturas y ecosistemas fueron similares. En [Haggarty *et al.*, 2015] se aplicó el método de agrupación jerárquico funcional para conocer los niveles de contaminación por nitrato en el río Tweed en Escocia y se consiguió identificar grupos de estaciones de monitoreo de características espacio-temporales similares.

En [Giraldo *et al.*, 2012] se menciona que tener métodos alternativos para realizar este análisis de agrupación de datos funcionales correlacionados espacialmente, es de gran ayuda ya que se contarían con herramientas adicionales para el investigador; por esta razón se propone una adaptación del método k-medias para tratar este tipo de datos, así como indicadores de calidad para conocer el número de grupos que se debería considerar e índices de correlación espacial y temporal para medir la consistencia de los mismos.

Capítulo 2

Marco Teórico

En el presente capitulo se dan a conocer conceptos fundamentales para el desarrollo del presente trabajo de titulación; así, en la sección 2.1 se definen los principales conceptos de estadística espacial; en la sección 2.2 se tratan conceptos de análisis de datos funcionales; en la sección 2.3 se definen tres métodos de agrupación, los cuales son: agrupación de datos funcionales, agrupación de datos espaciales y agrupación de datos funcionales con correlación espacial; y en la sección 2.4 se introducen índices de calidad para datos multivariantes.

2.1. Estadística espacial

La estadística clásica trabaja bajo el supuesto de independencia de los valores observados; estas observaciones son consideradas como realizaciones independientes de la misma función aleatoria; sin embargo, cuando los datos observados están sujetos a una zona geográfica, la hipótesis de independencia no se cumple[Montero *et al.*, 2015]; adicionalmente, la primera ley de la geografía establece que: "todo está relacionado con todo lo demás, pero cosas cercanas están más relacionadas que las cosas distantes"[Tobler, 1970].

2.1.1. Preliminares

2.1.1.1. Función Aleatoria

Dado un dominio $D \subset \mathbb{R}^n$ y un espacio de probabilidad $(\Omega, \mathcal{A}, P)$, una función aleatoria es una función de dos variables $Z(x, \omega)$ tal que para cada $x \in D$, $Z(x, .)$ es una variable aleatoria en $(\Omega, \mathcal{A}, P)$. Por otro lado, $Z(., \omega)$, definido en D como la sección de la función aleatoria en $\omega \in \Omega$ es una realización de la función aleatoria. La función aleatoria $Z(x, \omega)$ se denota como $Z(x)$ y una realización es representada por $z(x)$ [Jean Paul Chiles, 2012].

2.1.1.2. Proceso Estocástico

Según [Jean Paul Chiles, 2012], una función aleatoria es llamada proceso estocástico cuando x varia en un espacio unidimensional como el tiempo; por otro lado, si x varia en más de una dimensión es conocido como campo aleatorio.

2.1.1.3. Proceso Espacial

Sea Z la variable aleatoria de interés y s su ubicación espacia. El proceso espacial es el proceso estocástico

$$\{Z(s) : s \in D\}$$

donde $D \subset \mathbb{R}^d$ es el conjunto formado por todas las ubicaciones espaciales [Bohorquez, 2020].

2.1.1.4. Proceso Temporal

Sea Z la variable aleatoria de interés y t el tiempo. El proceso temporal es el proceso estocástico

$$\{Z(t) : t \in D\}$$

donde $D \subset \mathbb{R}$ es el conjunto formado por todos los instantes de tiempo [Bohorquez, 2020].

2.1.1.5. Proceso Espacio Temporal

Sea Z la variable aleatoria de interés, s su ubicación espacial y t el tiempo de ocurrencia. Un proceso espacio-temporal es un proceso estocástico

$$\{Z(s,t) : (s,t) \in D \times D'\}$$

donde el conjunto $D \subset \mathbb{R}^d$ es el conjunto de todas las ubicaciones espaciales y el conjunto $D' \subset \mathbb{R}$ es el conjunto de todos los instantes de tiempo. Los conjuntos D y D' pueden ser continuos o discretos, fijos o aleatorios [Bohorquez, 2020].

2.1.2. Clases de Datos Espaciales

Dependiendo de las características del dominio espacial D del proceso estocástico de interés, se tienen los siguientes tipos de datos espaciales principales, los cuales son geoestadística, datos de área o *lattice*, procesos o patrones espaciales puntuales y datos georreferenciados.

2.1.2.1. Geoestadística

Es el conjunto de métodos aplicados a datos espaciales, donde las ubicaciones espaciales s provienen de un conjunto continuo y fijo $D \subset \mathbb{R}^d$; cabe resaltar que el propósito de la geoestadística es la interpolación. Además, si el conjunto D no es continuo se puede llegar a obtener predicciones carentes de sentido, y se considera D fijo pues los puntos en el espacio se seleccionan a conveniencia o bajo un esquema de muestreo probabilístico [Giraldo, 2008].

2.1.2.2. Datos de Área

En este tipo de datos las ubicaciones espaciales pertenecen a un conjunto discreto contable y fijo $D \subset \mathbb{R}^d$. Las ubicaciones pueden estar espaciadas regular o irregularmente, usualmente son irregulares pues sus separaciones no siguen un patrón predecible.[Cressie, 1991].

2.1.2.3. Patrones Puntuales

En este caso, las ubicaciones espaciales pertenecen a un conjunto discreto o continuo y aleatorio $D \subset \mathbb{R}^d$. Con este tipo de dato se realizan análisis con el propósito de determinar si la distribución de las mediciones en la región es aleatoria, agregada o uniforme [Giraldo, 2008].

2.1.2.4. Datos Georreferenciados

En este caso, las variables de interés tienen asociadas las coordenadas de las ubicaciones donde fueron medidas, las que pueden ser geográficas, planas o cartesianas [Giraldo, 2008].

2.1.3. Geoestadística

La palabra geoestadística fue establecida por Hart en el año de 1954. El significado de esta palabra se lo puede deducir a partir de su prefijo geo, que hace referencia a la tierra; por lo tanto, la geoestadística es una rama de la estadística enfocada en el análisis y la modelización de la variabilidad espacial de fenómenos que ocurren dentro de una zona geográfica; además, es considerada como una disciplina híbrida entre Geología, Matemática, Estadística y Minería de datos [Cressie, 1991].

2.1.3.1. Variable Regionalizada

Una variable regionalizada es un proceso estocástico con dominio continuo contenido en un espacio euclidiano d-dimensional $\{Z(s) : s \in D \subset \mathbb{R}^d\}$.

En el caso de que las mediciones sean realizadas en una superficie; es decir, $d = 2$, $Z(s)$ puede asociarse a la variable aleatoria ligada a ese punto del plano, s representa las coordenadas geográficas y Z la variable en cada una de ellas. La medición de esta variable aleatoria puede representar la magnitud de una variable ambiental medida en un conjunto de coordenadas en la región de estudio [Giraldo, 2008].

Un proceso estocástico espacial se caracteriza por su distribución de probabilidad de dimensión finita; es decir, la distribución de probabilidad conjunta de un conjunto de variables $Z(s_1), ..., Z(s_k)$ para todo $k \in \mathbb{N}$ y para todos los puntos $s_1, ...s_k \in D$ [Bohorquez, 2020].

2.1.3.2. Función de Distribución Conjunta

Considérese una función/campo aleatorio $Z = \{Z(s) : s \in D\}$ y k ubicaciones espaciales $s_1, ..., s_k \in D$. El vector aleatorio $\{Z(s_1), ..., Z(s_k)\}$ está caracterizado por su función de distribución conjunta:

$$F_{s_1,...,s_k}(z_1, ..., z_2) = P[Z(s_1) \leq z_1, ..., Z(s_k) \leq z_k]$$

2.1.3.3. Función de Media

La esperanza o el momento de primer orden de un campo aleatorio o proceso estocástico espacial $Z(s)$, es una función no aleatoria de s

$$E[Z(s)] = \mu(s) \quad \text{donde} \quad \mu(s_i) = E(Z(s_i)) \; i = 1, ..., k$$

En cada sitio s dado, $\mu(s)$ representa la media alrededor de la cual se distribuyen los valores tomados por las realizaciones de la funciones aleatoria [Giraldo, 2008]. Si la esperanza del campo aleatorio varía con la ubicación espacial, se le conoce como deriva o *drift* del campo aleatorio.

En geoestadística existen tres elementos de segundo orden, los cuales son: varianza, convarianza y variograma.

2.1.3.4. Función de varianza

La varianza de un campo aleatorio o proceso estocástico espacial $Z(s)$ respecto a $\mu(s)$, está definida por:

$$V(s) = \sigma^2(s) = Var[Z(s)] = E[(Z(s) - \mu(s))^2] \quad \text{donde} \quad V(s_i) = \sigma^2(s_i) = V(Z(s_i)) \; i = 1, ..., k$$

2.1.3.5. Función de Covarianza

Sean $Z(s_i)$ y $Z(s_j)$ dos variables aleatorias de un proceso estocástico espacial, la covarianza es una función de separación espacial de s_i y s_j y está definida por:

$$C(s_i, s_j) = C(s_i - s_j)$$

$$C(s_i, s_j) = C(Z(s_i), Z(s_j))$$

$$C(s_i, s_j) = E[(Z(s_i) - \mu(s_j))(Z(s_j) - \mu(s_j))]$$

$$C(s_i, s_j) = E[Z(s_i)Z(s_j)] - \mu(s_i)\mu(s_j)$$

Nótese que:

$$C(s, s) = C(0) = \sigma^2(s)$$

2.1.3.6. Función de correlación o Correlograma

La correlación lineal de dos variables aleatorias de un proceso estocástico espacial $Z(s_i)$ y $Z(s_j)$, está definida por:

$$\rho(s_i, s_j) = \frac{C(s_i, s_j)}{\sigma(s_i)\sigma(s_j)}$$

2.1.3.7. Función de semi-variograma

El semi-variograma entre dos variables aleatorias de un proceso estocástico espacial $Z(s_i)$ y $Z(s_j)$ está dado por:

$$\gamma(s_i, s_j) = \gamma(s_i - s_j) = \frac{1}{2}V[Z(s_i) - Z(s_j)],\ \forall s_i, s_j \in D$$

2.1.3.8. Estacionariedad de Funciones Aleatorias

Una variable aleatoria regionalizada vista como una realización de un proceso estocástico espacial o función aleatoria $\{Z(s) : s \in D\}$, probabilísticamente adquiere sentido cuando es posible inferir toda o parte de la ley de probabilidad del proceso estocástico o campo aleatorio. Es claro que inferir la ley de probabilidad cuando solo se tiene una realización del proceso estocástico espacial o función aleatorio sería imposible; para poder hacer inferencia consistente, se necesitan de varias realizaciones, pero en la realidad solo existe una, inclusive solo una parte de las realizaciones está disponible en las ubicaciones de la muestra; para dar solución a este problema, se introduce el supuesto de estacionariedad, que significa que la ley de probabilidad espacial del proceso estocástico espacial o función aleatoria, o parte de este, es invariante respecto a traslaciones; es decir, las propiedades probabilísticas de un conjunto de observaciones no dependen de las ubicaciones donde fueron medidas, solo dependen de su separación; además, como se mencionó, las realizaciones no son independientes [Montero *et al.*, 2015]. Así, se distinguen tres tipos de estacionariedad:

- Estacionariedad fuerte o de primer orden.
- Estacionariedad débil o de segundo orden.
- Estacionariedad intrínseca .

Estacionariedad Fuerte o de primer orden

Este supuesto se refiere a que el proceso alcanza un estado de equilibrio. Formalmente, una variable regionalizada es estacionaria si su función de distribución no varía con la traslación del vector h; esto es, si para:

$$Z(s) = [Z(s_1), ..., Z(s_n)]'$$
$$Z(s+h) = [Z(s_1+h), ..., Z(s_n+h)]'$$

se tiene que:

$$F_{Z_1,...,Z_n}(s_1, ..., s_n) = F_{Z_1,...,Z_n}((s_1+h), ..., (s_n+h))$$

Nótese que esta condición es muy restrictiva, por lo que normalmente se relaja a condiciones de segundo orden, las que limitan el supuesto de estacionariedad a los dos primeros momentos del proceso estocástico o campo aleatorio [Montero *et al.*, 2015].

Estacionariedad de Segundo Orden

Sea $\{Z(s) : s \in D\}$ un proceso estocástico o campo aleatorio, se dice débilmente estacionario o de segundo orden si tiene momentos de segundo orden finitos; es decir, que la covarianza existe y se verifica que:

- La esperanza existe, es constante a través del dominio D, y no depende de la ubicación s; es decir:

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

- La covarianza existe para todo par de variables aleatorias, $Z(s)$ y $Z(s+h)$, y solo depende del vector de separación h entre las ubicaciones s y $s+h$; es decir, depende de la dirección y la distancia de separación, y no de sus ubicaciones absolutas y se tiene que:

$$Cov(Z(s), Z(s+h)) = C(h), \forall s \in D \ y \ h$$

  La estacionariedad de la covarianza implica que la varianza $Var(Z(s))$ existe, es finita y no depende de s; es decir:

$$Var(Z(s)) = \sigma^2(s) = C(s,s) = C(s-s) = C(0) = \sigma^2$$

La estacionariedad de segundo orden implica la siguiente relación entre entre la función de semivarianza y la de autocovarianza:

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

Nótese que si un proceso estocástico o campo aleatorio es estacionario de segundo orden, entonces también será un proceso estocástico estacionariamente fuerte; sin embargo, esto no ocurre en el sentido contrario.

Se debe tener en cuenta que la estacionariedad de segundo orden implica la existencia de la varianza del proceso estocástico o campo aleatorio. Existen fenómenos con varianza infinita lo que imposibilita su modelización utilizando procesos de este tipo. Sin embargo, existen casos en los que los incrementos, $Z(s+h) - Z(s)$, tienen varianza finita y que además son un proceso estacionario de segundo orden; este tipo de proceso estocástico o campo aleatorio se conoce como un proceso con estacionariedad intrínseca [Montero *et al.*, 2015].

Estacionariedad Intrínseca

Sea $\{Z(s) : s \in D\}$ un proceso estocástico o campo aleatorio; se dice que es un proceso con estacionariedad intrínseca, si:

- $Z(s)$ tiene esperanza finita y constante para todo punto en el domino, lo cual implica que la esperanza de los incrementos sea cero.

$$E(Z(s+h) - Z(s)) = 0$$

- Para cualquier vector h, la varianza del incremento está definida y es una función única de la distancia.

$$V(Z(s+h) - Z(s)) = E(Z(s+h) - Z(s))^2 = 2\gamma(h)$$

2.1.3.9. No estacionariedad de Funciones Aleatorias

Sea $\{Z(s) : s \in D\}$ un proceso estocástico o campo aleatorio para el cual la media y/o la función de covarianza depende de la ubicación espacial; es decir, no son traslacionalmente invariantes; se dice que $Z(s)$ es una función aleatoria no estacionaria.

Cuando el proceso estocástico o campo aleatorio $\{Z(s) : s \in D\}$ tiene media no constante y varia con la ubicación espacial, y sus incrementos de primer orden $Z(s+h) - Z(s)$ son no estacionarios, se dice que la función aleatoria es no intrínseca [Montero *et al.*, 2015].

2.1.3.10. Isotropía

Identificar la estacionariedad en un campo temporal es fácil, ya que solo existe una dirección de variación. Por otro lado, en el campo espacial existen múltiples direcciones, por lo que se debe asumir que el fenómeno es estacionario en todas ellas. Cuando la esperanza de la variable cambie respecto a las direcciones o cuando la covarianza o correlación dependan del sentido en el que se determinen, entonces se dirá que no se tiene estacionariedad [Giraldo, 2008].

Ahora bien, si $C(h)$ y/o $\gamma(h)$, son funciones únicas de la magnitud $\|h\|$

$$Cov(Z(s), Z(s+h) = C(\|h\|))$$
$$\frac{1}{2}Var(Z(s+h) - Z(s)) = \gamma(\|h\|)$$

el proceso posee función de covarianza y/o semivarianza isotrópica. La estacionariedad posibilita combinar pares de datos con la misma diferencia en magnitud de separación de las coordenadas, y si los vectores de diferencias tienen la posibilidad de ser reemplazados con distancias escalares, entonces el campo se dice isotrópico [Bohorquez, 2020]. Si la correlación entre los datos no es dependiente de la dirección en la que esta se calcule se plantea que el fenómeno es isotrópico [Giraldo, 2008]; en caso opuesto un campo aleatorio que es estacionario pero no isotrópico se desenvuelve de forma distinta según las diferentes direcciones del espacio; por lo cual, no es suficiente conocer cuánto permanecen distanciadas las ubicaciones, sino además es necesario conocer la orientación de esa distancia; a dichos campos se los conoce como campos aleatorios anisotrópicos [Bohorquez, 2020].

2.1.3.11. Correlación Espacial

El análisis estructural es la primera etapa en el desarrollo de un análisis geoestadístico; a partir de este se puede expresar la estructura de dependencia espacial o correlación que existe entre los datos medidos de una variable mediante funciones como el variograma y/o covariograma [Giraldo, 2008].

Covariograma

Como se definió anteriormente, la función de covarianza de un proceso estocástico espacial o campo aleatorio esta dada por:

$$\begin{aligned} C(s_i, s_j) &= C(Z(s_i), Z(s_j)) \\ &= E((Z(s_i) - \mu(s_i))(Z(s_j) - \mu(s_j))), \ \forall s_i, s_j \in D \end{aligned}$$

Bajo la hipótesis de estacionariedad de segundo orden, la función de covarianza tiene las siguientes propiedades:

- Solo depende del vector de separación h entre las ubicaciones espaciales:

$$C(h) = E((Z(s+h) - \mu)(Z(s) - \mu)), \ \forall s, s+h \in D \subset \mathbb{R}^d$$

  donde μ es la media constante del proceso estocástico espacial o campo aleatorio. El covariograma muestra el comportamiento de la correlación entre $Z(s)$ y $Z(s+h)$.

  Cuando la función de covarianza solo depende de la distancia entre las ubicaciones s y $s+h$ se conoce como **proceso isotrópico**. Cuando depende de la distancia y de la dirección del vector de separación h se conoce como un **proceso anisotrópico**.

- Está acotado por la varianza del proceso estocástico espacial o campo aleatorio en el origen; es decir:

$$|C(h)| \leq C(0) = Var(Z(s))$$

- Es una función par; es decir:

$$C(h) = C(-h)$$

- Es una función definida positiva, en términos de diferencia entre coordenadas $s_i - s_j = h$:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j C(s_i - s_j) \geq 0, \forall n \in \mathbb{N}^*, \forall \lambda_1, ..., \lambda_n \in \mathbb{R}, \forall s_1, ..., s_n \in D \subset \mathbb{R}^d$$

  Esta condición es la más importante y necesaria para que una función de covarianza esté bien definida [Montero *et al.*, 2015].

- Si $C_k(h), \forall k \in \mathbb{N}$ son funciones de covarianza en $\mathbb{R}^d$ y:

$$\lim_{k\to\infty} C_k(h) = C(h), \forall h \in \mathbb{R}^d$$

  entonces, $C(h)$ es una función de covarianza en $\mathbb{R}^d$, teniendo en cuenta que este límite existe $\forall h$.

- Toda combinación lineal de funciones de covarianza con coeficientes positivos, también es una función de covarianza.

- El producto de funciones de covarianza es también una función de covarianza.

Semivariograma

El semivariograma es el instrumento más usado para describir la dependencia espacial de variables regionalizadas, pues este cubre un mayor espectro de estas variables en comparación de la función de covarianza, incluyendo la estacionariedad intrínseca de funciones aleatorias, en donde la covarianza no puede ser definida [Montero *et al.*, 2015].

Como se definió anteriormente, el semivariograma está dado por:

$$\gamma(s_i - s_j) = \frac{1}{2} Var(Z(s_i) - Z(s_j)), \forall s_i, s_j \in D$$

Bajo el supuesto de estacionariedad de segundo orden o de hipótesis intrínseca (sin media/drift), se escribe como:

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

la cual muestra como la disimilitud entre $Z(s+h)$ y $Z(s)$ se desenvuelve con la distancia h.

Como se vio anteriormente, bajo el supuesto de estacionariedad de segundo orden, el semivariograma y el covariograma verifican la siguiente relación:

$$C(h) = C(0) - \gamma(h)$$

En la práctica se utiliza el semivariograma en lugar del covariograma puesto que este no necesita conocer la media de la función aleatoria [Montero *et al.*, 2015].

Para que una función de semivariograma sea válido necesita verificar las siguientes propiedades teóricas:

- Por definición $\gamma(0) = 0$, aunque es común que presente una discontinuidad en el origen; esta discontinuidad se la conoce como efecto pepita.

- Es una función par; es decir:

$$\gamma(h) = \gamma(-h)$$

- Siempre toma valores mayores o iguales que cero:

$$\gamma(h) \geq 0$$

- Bajo estacionariedad intrínseca, es una función condicionalmente negativa; es decir:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(s_i - s_j) \leq 0, \forall n \in \mathbb{N}^*, \forall\lambda_1, ..., \lambda_n \in \mathbb{R} : \sum_{i=1}^{n}\lambda_i = 0, \forall s_1, ..., s_n \in D \subset \mathbb{R}^d$$

- El semivariograma de una proceso estocástico o campo aleatorio estacionario de segundo orden es finito; es decir, su comportamiento tiende a una línea horizontal a medida que se incrementa la separación de las coordenadas; sin embargo, el de una función aleatoria intrínsecamente estacionaria sin media o *drift* puede crecer al infinito; es decir, el semivariograma no se estabiliza; en este caso el semivariograma puede ser al menos intrínsecamente estacionario si crece más lento que una ecuación de segundo grado; es decir:

$$\lim_{|h|\to\infty}\frac{\gamma(h)}{|h|^2} = 0$$

El semivariograma de un proceso estocástico espacial estacionario de segundo orden depende de los siguientes parámetros (ver Figura 2.1):

- Silla: Es la asíntota superior del semivariograma. Solamente los procesos estacionarios de segundo orden tienen silla; en estos casos la silla es de $C(0) = \sigma^2$.

- Rango: Es la distancia sobre la cual los puntos ya no ejercen influencia sobre otros; es decir, son independientes. Los puntos que se encuentran separados por una distancia menor o igual al rango se consideran espacialmente correlacionados, mientras que los puntos separados por una distancia mayor se consideran independientes [Bohorquez, 2020].

- Efecto Pepita: Es una discontinuidad en el origen, $\gamma(h) \to c_0$ cuando $h \to 0$.

  Existen dos causas principales para que se produzca este efecto. La primera es la existencia de error de medición (EM); sucede una vez que no es viable repetir una medición en la ubicación s sin error, evidenciando su variabilidad. La segunda causa es el efecto de micro-escala (MS); el cual se produce ya que existe un proceso espacial que opera a distancias más pequeñas de las que fueron consideradas en el muestreo. Si ES y MS son diferentes de cero, el semivariogrma presentará una discontinuidad puntual en el origen, conocido como efecto pepita y representado de la siguiente manera [Bohorquez, 2020]:

$$c_0 = \sigma^2_{EM} + \sigma^2_{MS}$$

- Silla Parcial: Si un semivariograma tiene efecto pepita c_0 y silla $C(0)$, la diferencia $c_p = C(0) - c_0$ es la silla parcial del semivariograma, la cual representa la varianza de la variable $Z(s)$ sin el efecto pepita.

La relación entre el covariograma y el semivariograma se modifica para tomar en cuenta el efecto pepita de la siguiente manera:

$$\gamma(h) = c_0 + c_p\gamma'(h)$$
$$C(h) = \begin{cases} c_p(1-\gamma'(h)) & si\ h > 0 \\ c_0 + c_p & si\ h = 0 \end{cases}$$

donde $\gamma'(h)$ es el modelo de variograma, c_p es la silla parcial y c_0 es el efecto pepita [Bohorquez, 2020].

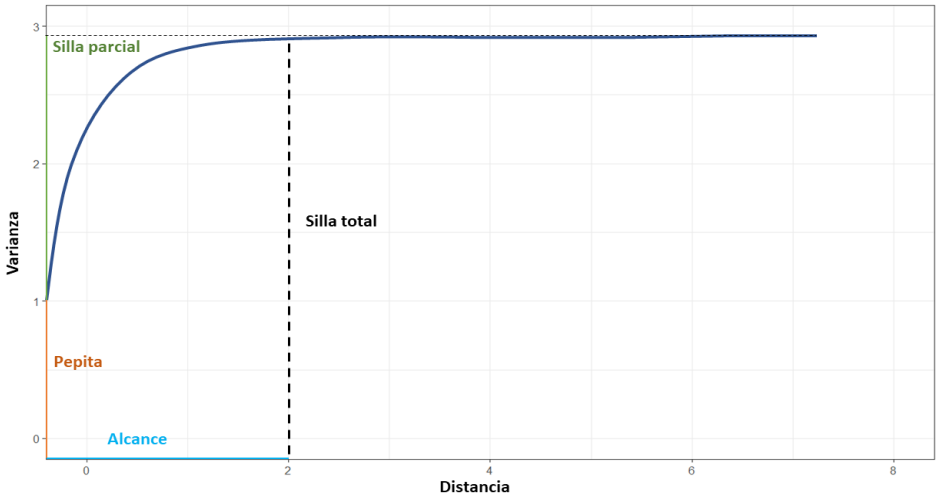


Figura 2.1: Partes de un variograma

A partir de este punto se omitirá el prefijo "semi" y se tratará al semi-variograma solo por variograma.

2.1.4. Variograma

En la práctica, se trabaja con las realizaciones del proceso estocástico o campo aleatorio en estudio; es decir, se cuenta con un conjunto de datos georeferenciados en un dominio dado. Utilizando estos datos, se puede inferir la estructura de dependencia espacial del fenómeno [Montero *et al.*, 2015].

Por lo general, la descripción de la distribución espacial se limita a los primeros momentos. La esperanza o el momento de primer orden involucra un solo sitio a la vez y no proporciona información sobre la dependencia espacial. Por otro lado, la covarianza, correlograma y variograma o momentos de segundo orden están definidos mediante un par ubicaciones; así, estos momentos proporcionan información de la continuidad y dependencia espacial de la variable regionalizada [Emery, 2013].

2.1.4.1. Disimilitud contra Separación

La variabilidad de una variable regionalizada $Z(s)$, se mide en diferentes escalas y se calcula a través de la disimilitud entre los valores z_{s1} y z_{s2} ubicados en los puntos s_1 y s_2 del domino espacial D. La medida de disimilitud $\hat{\gamma}$ de estos valores está dada por:

$$\hat{\gamma}(s_1, s_2) = \frac{(z_{s1} - z_{s2})^2}{2}$$

Los dos puntos s_1 y s_2 en el espacio geográfico pueden unirse por un vector $h = s_1 - s_2$ como se muestra en la Figura 2.2.

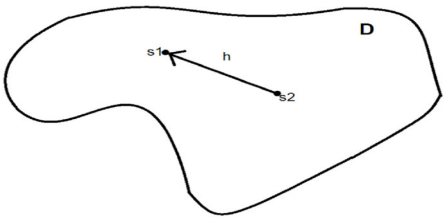


Figura 2.2: Distancia

Si $\hat{\gamma}$ es dependiente de la separación y orientación del par de ubicaciones descritos por el vector h, se tiene que [Wackernagel, 2003]:

$$\hat{\gamma}(h) = \frac{1}{2}(z(s_1 + h) - z(s_1))^2$$

La disimilitud es simétrica con respecto de h, pues es una cantidad al cuadrado, por lo que, su representación será mostrada utilizando todos los pares de datos de la muestra en el conjunto D. El gráfico de la disimilitud $\hat{\gamma}$ contra la separación espacial absoluta h, es llamado nube del variograma [Wackernagel, 2003].

La disimilitud con frecuencia se incrementa con la distancia, pues los puntos cercanos tienden a ser similares [Wackernagel, 2003] (ver Figura 2.3).

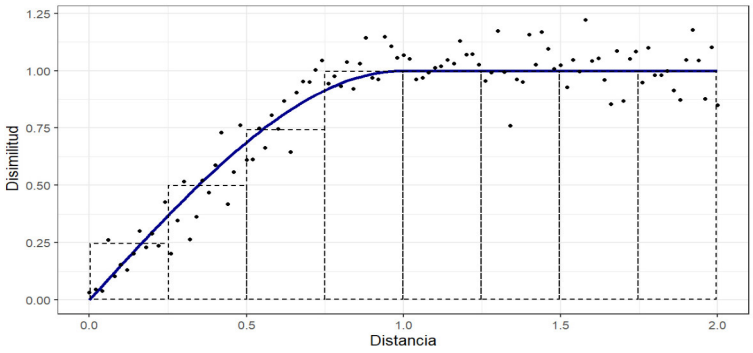


Figura 2.3: Disimilitud respecto de h.

2.1.4.2. Variograma Empírico

Teniendo en cuenta que el variograma $\gamma(h)$ es la varianza de la variable de incrementos $Z(s+h) - Z(s)$, el estimador clásico se basa en la estimación de esta varianza mediante

el método de momentos y está definido para un vector de separación h de la siguiente manera:

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{N(h)} (Z(s+h) - Z(s))^2, \ h \in \mathbb{R}^d$$

donde:

$$N(h) = \{(s_i, s_j) : s_i - s_j = h, i, j = 1, ..., n\}$$

siendo $N(h)$ el conjunto de todos los pares de ubicaciones cuya separación corresponde a un vector h y $|N(h)|$ es la cardinalidad de $N(h)$. Nótese que $N(-h) \neq N(h)$, aunque $\hat{\gamma}(-h) = \hat{\gamma}(h)$, y la media μ no necesita ser estimada.

Al ser un estimador de la varianza muestral es sensible a datos atípicos; es por ello que [Cressie, 1991] propuso estimadores resistentes a datos atípicos, asumiendo normalidad marginal del campo aleatorio [Bohorquez, 2020]; sin embargo, la diferencia entre estos estimadores es pequeña [Cressie, 1991].

Tolerancia en los parámetros de cálculo

En la práctica, los datos están distribuidos irregularmente en el dominio espacial D, por lo que el número de pares $|N(h)|$ que se utilizan en el cálculo del variograma empírico, $\hat{\gamma}(h)$ para un vector h dado, es por lo general escaso; dando como resultado una estimación del variograma empírico errónea, provocando que sea imposible de interpretarlo y modelarlo. Para corregir esto, se suelen añadir algunas tolerancias de cálculo sobre las distancias y direcciones, teniendo así que:

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

donde $N^+ = \{(s_i, s_j) : s_i - s_j \in T(h)\} = \bigcup_{h' \in T(h)} N(h')$

Siendo $T(h)$ una región de tolerancia alrededor de h; es decir, $T(h) = [h - \Delta h, h + \Delta h]$ en el caso unidiminesional. En el caso de dimensión dos o tres, existen tolerancias sobre la longitud de h y sobre su orientación [Emery, 2013] (véase Figura 2.4):

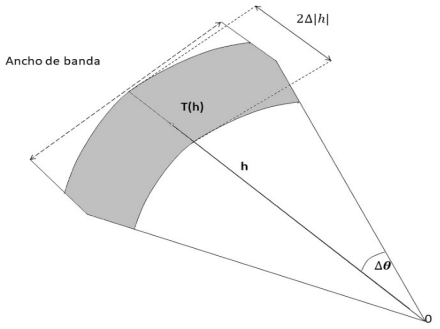


Figura 2.4: Región de tolerancia $T(h)$ alrededor del vector h.

El ancho de banda limita la división del cono de tolerancia a una expansión máxima. En el espacio de tres dimensiones, se introducen dos anchos de banda, horizontal y vertical [Emery, 2013].

Propiedades del variograma empírico

- El variograma empírico $\hat{\gamma}(h)$ es un estimador insesgado del variograma teórico [Montero *et al.*, 2015]:

$$E(\hat{\gamma}(h)) = \gamma(h)$$

- La varianza relativa es un indicador de robustez:

$$\frac{Var(\hat{\gamma}(h))}{(\gamma(h))^2}$$

  A medida que más alta sea dicha varianza, más susceptible es el variograma empírico de fluctuar en torno a su valor esperado, y más difícil se vuelve la inferencia estadística.

  Los principales factores que influyen en la varianza relativa son [Emery, 2013]:

  - Norma del vector h: la varianza relativa crece cuando la distancia aumenta.
  - La irregularidad de la malla de muestreo, tiene la posibilidad de causar grandes variaciones en el cálculo del variograma empírico, incluso a distancias pequeñas.
  - A medida que más reducido es el número de pares de datos, más grandes son las variaciones.
  - La existencia de datos atípicos, puesto que en el cálculo del variograma empírico, se elevan los valores al cuadrado.

- Las direcciones de cálculo del variograma empírico deben tener en consideración la anisotropía de la variable regionalizada; en la situación de isotropía, donde los variogramas direccionales son semejantes, se puede tener en cuenta un variograma omnidireccional, determinado por:

$$\bar{\gamma}^+(h) = \frac{1}{2|N^+(h)|} \sum_{N^+(h)} (Z(s+h) - Z(s))^2$$

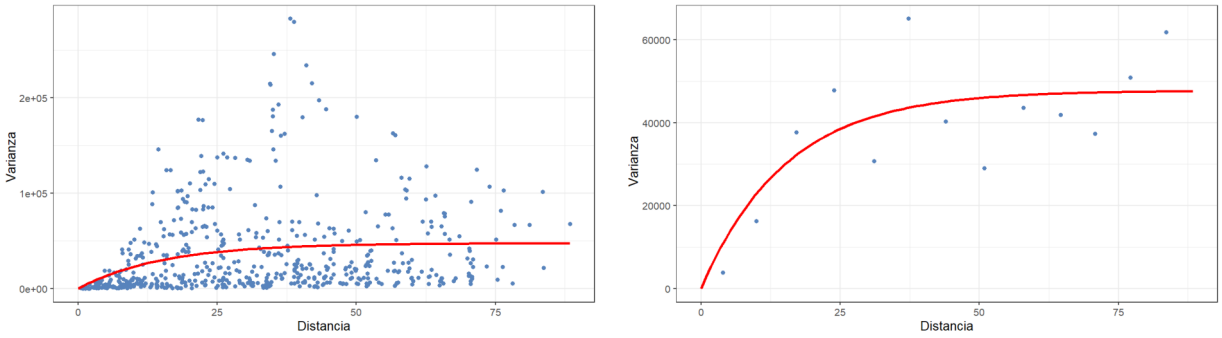
  donde: $|N^+(h)| = \{(s_i, s_j) : s_i - s_j \approx h\}$ [Montero *et al.*, 2015].

- Si el variograma presenta comportamiento anisotrópico es necesario realizar transformaciones de las coordenadas que permitan obtener variables regionalizadas anistrópicas a partir de los modelos isotrópicos [Wackernagel, 2003].

Nube variográfica

Teniendo presente que bajo el supuesto de estacionariedad, los valores $\hat{\gamma}(s_i, s_j)$ son estimadores insesgados de los valores que corresponden $\gamma(s_i, s_j)$. La recopilación de pares de distancias y sus correspondientes valores de variograma, $\{(h, \gamma(h)) : h = s_i - s_j\}$, es conocido como variograma empírico y su gráfico como nube variográfica. Para mejorar la conducta del variograma empírico como un estimador del variograma teórico $\gamma(h)$ es necesario aplicar un tipo de suavización, puesto que se espera que la función $\gamma(h)$ varíe suavemente en función de h, por lo cual se disminuye la varianza sin añadir sesgo llevando a cabo un promedio de los valores de $\hat{\gamma}$ sobre rangos o *bins* adecuados de distancia entre puntos s_i, s_j [P.J. Diggle, 2007].

Si el diseño de muestreo es una grilla regular la suavización puede ser lograda sin introducir sesgo, simplemente promediando todos los valores $\hat{\gamma}(h)$ para cada h diferente. Sin embargo, si el diseño es irregular, para un rango de ancho m se define un variograma γ_k, para un entero positivo k, como el promedio de todos los $\hat{\gamma}(h)$ para el cual la respectiva distancia h satisfaga $(k-1)m < h \leq km$; entonces, γ_k es aproximadamente una estimación insesgada de $\gamma(h_k)$; por conveniencia se adopta $h_k = (k - 0{,}5)m$, que es el punto medio del intervalo respectivo [P.J. Diggle, 2007].



(a) Nube variográfica.

(b) Bin variograma

Figura 2.5: Variogramas

2.1.4.3. Estimación del variograma

Se sabe que no se puede utilizar de manera directa el variograma emprírico, puesto que está definido para ciertas distancias y direcciones. Por otro lado, debido a los parámetros de tolerancia, está sujeto a aproximaciones, puesto que en el proceso de cálculo se utiliza un número limitado de datos. Para dar solución a esto, se ajusta un modelo teórico de variograma a partir del variograma empírico, siendo esta etapa la más significativa de todo análisis geoestadístico, puesto que nos da información de la continuidad espacial de la variable en estudio [Emery, 2013].

Una función de variograma teórico es ajustada a la sucesión de disimilitudes medias; nótese que este ajuste involucra una interpretación del comportamiento en el origen y el comportamiento para largas distancias, más allá del rango del variograma empírico. El ajuste puede llevarse a cabo de manera empírica mediante criterio experto, puesto que en la mayoría de casos prácticos no es fundamental el que tan bien la función de variograma se ajuste a la secuencia de puntos; lo más relevante es el tipo de continuidad que se asume para la variable regionalizada y la hipótesis de estacionariedad asociada

al proceso estocástico espacial o campo aleatorio; estas suposiciones sirven de guía en la selección de una función de variograma teórico [Wackernagel, 2003].

Comportamiento en el origen

Si las distancias son cercanas a cero; es decir, la variable regionalizada es regular en el espacio, más regular será el variograma en el origen. Generalmente, se suelen diferenciar tres tipos de comportamiento para el variograma en el origen (ver Figura 2.6) [Emery, 2013]:

- Parabólico: Corresponde a una variable regionalizada bastante regular en el espacio.
- Lineal: Corresponde a una variable regionalizada continua, pero no tan regular.
- Discontinuo: Corresponde a una variable regionalizada más errática; es decir, con discontinuidades en la distribución espacial de sus valores. La diferencia entre dos datos bastante cercanos es grande, lo que provoca el efecto pepita.

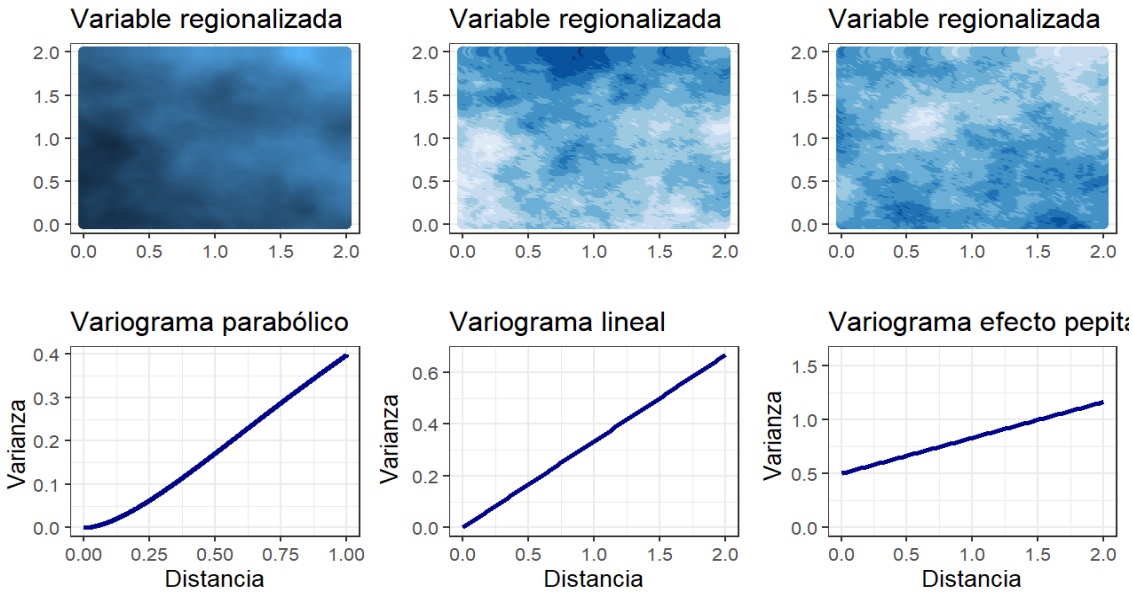


Figura 2.6: Comportamientos del variograma en el origen.

Comportamiento para distancias muy grades

El variograma habitualmente crece a partir del origen y se estabiliza a partir de una distancia ϕ, alrededor de una silla σ^2; en este caso se conoce que esa silla es igual a la varianza a priori $C(0) = \sigma^2$.

Como se vio anteriormente, las variables aleatorias $Z(s)$ y $Z(s+h)$ están correlacionadas, si la longitud del vector de separación h es inferior a la distancia ϕ o alcance; más allá de $|h| = \phi$, el variograma es constante e igual a su silla, y las variables $Z(s)$ y $Z(s+h)$ son independientes. Los variogramas con silla, se denominan *modelos de transición* (ver Figura 2.7) [Emery, 2013].

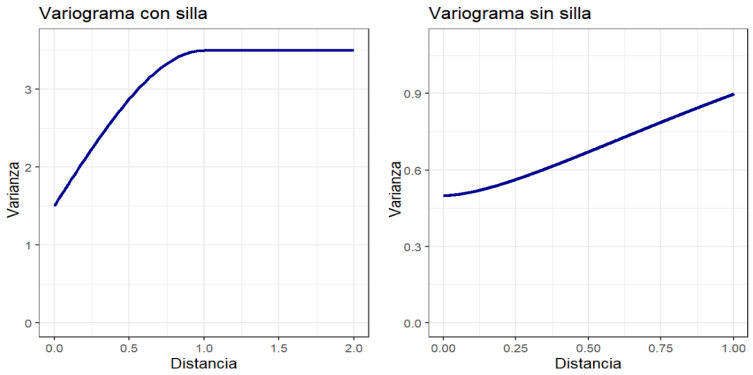


Figura 2.7: Lado izquierdo: Variograma con silla y alcance. Lado derecho: Variograma sin silla y alcance.

Existen casos en los cuales el variograma crece infinitamente una vez la distancia se incrementa y no posee ni silla ni rango; en este caso, la varianza es infinita $C(0) = \infty$ y no existe la función de covarianza ni el correlograma. Cuando se tiene ausencia de silla, es posible que sea una consecuencia del efecto de micro-escala [Emery, 2013].

Criterios para la estimación y ajuste del variograma

Tendencia:

Es fundamental verificar que los datos no presenten tendencia, puesto que cuando la media del proceso $\mu(s)$ no es constante; es decir, el proceso no es estacionario, el variograma empírico calculado a partir de las observaciones es erróneo; este comportamiento de tendencia se puede comprobar mediante un gráfico de dispersión de la variable de respuesta frente a cada coordenada espacial (ver Figura 2.8). En caso de que se evidencie tendencia se sugiere que se incluya un modelo de superficie de tendencia para la media que varía espacialmente, cuando una superficie de tendencia es incluida en el modelo, las dos coordenadas espaciales deben contribuir en este, puesto que la orientación de la región de estudio es arbitraria. En la práctica; al trabajar con datos geográficos de gran escala, se puede esperar que ciertas variables relacionadas a un ambiente físico muestren dependencia en la latitud [P.J. Diggle, 2007].

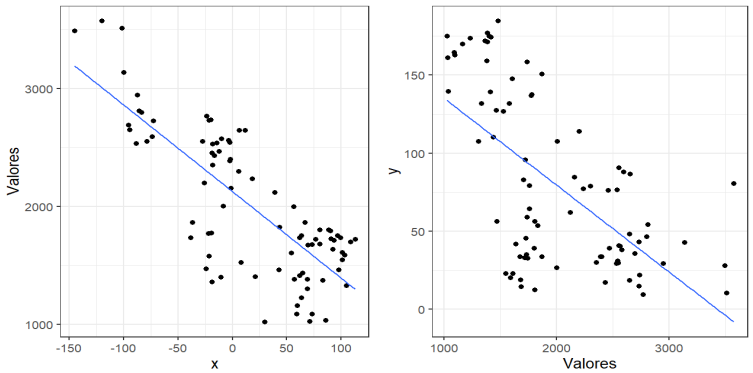


Figura 2.8: Tendencia de los valores con respecto a las coordenadas.

Cuando la función de media no es constante, el variograma empírico atribuye de ma-

nera errónea la variación inducida por esta a la estructura de convarianza de gran escala en el proceso no observado $Z(s)$. Una solución es estimar $\mu(s)$ mediante un modelo de superficie de tendencia o, si la información de la covariable está disponible, mediante un modelo general de regresión, y utilizar los residuos $R_i = Z_i - \hat{\mu}(s_i)$ en lugar de las observaciones para el cálculo del variograma empírico [P.J. Diggle, 2007]. Este comportamiento se puede visualizar en la Figura 2.9.

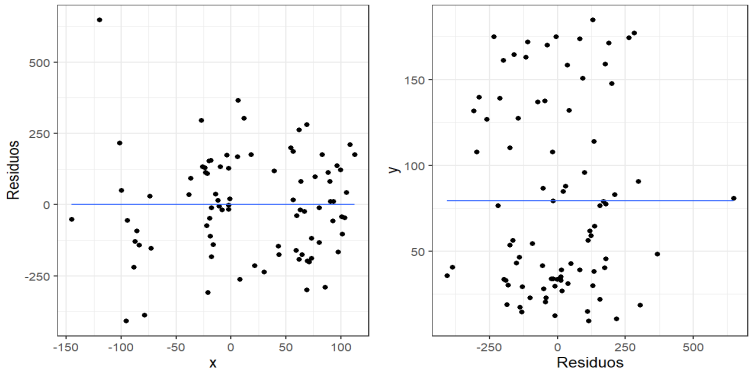


Figura 2.9: Tendencia de los residuos con respecto a las coordenadas.

Establecer la variable correcta:

En la estimación del variograma influye la distribución de los datos, pues estos pueden estar sesgados o tener valores extremos altos o bajos y por tanto su variograma calculado suele presentar un comportamiento errático. Para abordar este problema es recomendable transformar los datos a un espacio Normal o Gaussiano [Gringarten & Deutsch, 1999]. Ver Figura 2.10.

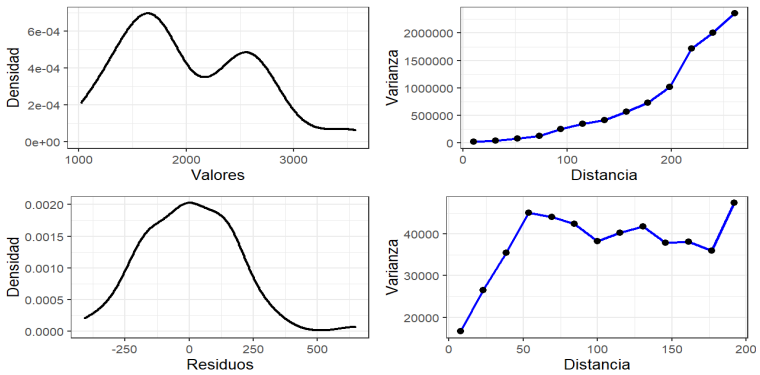


Figura 2.10: Arriba: Variograma empírico calculado a partir de los datos originales. Abajo: Variograma empírico calculado a partir de los residuos.

Verificación de isotropía:

Como bien se conoce, se debe verificar la estacionariedad de los datos, puesto que este influye de manera directa en la estimación del variograma empírico; teniendo en cuenta que un proceso espacial puede estar definido en varias direcciones, este criterio se debe

verificar para todas estas. No obstante, en la práctica la isotropía se estudia por medio del cálculo de funciones de autocovariaza o de semivarianza muestrales en algunas direcciones que habitualmente son: $0^o, 45^o, 90^o$ y 135^o, ver Figura. 2.11.

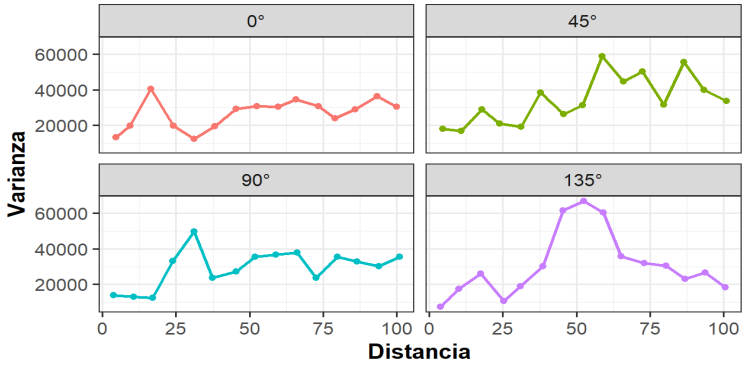


Figura 2.11: Comportamiento del variograma direccional.

Con la verificación de los puntos anteriores, se procede al cálculo del variograma empírico y su posterior ajuste de un modelo teórico por medio de los métodos que se exponen a continuación.

2.1.4.4. Métodos de estimación teórica más utilizados

Métodos con base en la función de verosimilitud

Los parámetros de la función de covarianza pueden ser estimados por medio de los métodos de Máxima Verosimilitud (ML) y Máxima Verosimilitud restringida (REML), los cuales requieren la especificación de la distribución del vector $Z = (Z(s_1), ..., Z(s_n))$; generalmente, se asume normalidad multivariada. Para el caso ML, se tiene que:

$$Z \sim N_n(X\beta, \Sigma(\theta))$$

donde $\Sigma(\theta) = Cov(Z)$ es una matriz de dimensión $n \times n$ y X es una matriz de dimensión $n \times q$ con $q < n$, de variables explicativas, dentro de las cuales comúnmente se encuentran las coordenadas geográficas; de donde el negativo de la función de logverosimil es:

$$L(\beta, \theta) = \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma(\theta|) + \frac{1}{2}(Z - X\beta)'\Sigma^{-1}(\theta)(Z - X\beta)$$

El elemento i, j de la matriz $\Sigma(\theta)$ corresponde a la covarianza espacial entre las variables $Z(s_i)$ y $Z(s_j)$; esto es:

$$Cov(Z(s_i), Z(s_j); \theta)) = C(s_i - s_j; \theta) = C(h; \theta)$$

El estimador $\hat{\theta}$ es sesgado pero asintóticamente eficiente; no obstante, al trabajar con una muestra grande se debe tener en consideración que debido a su conducta iterativa,

se realizarán grandes cantidades de operaciones computacionales debido al cálculo del determinante y de la inversa de la matriz de covarianza [Bohorquez, 2020].

Estimador REML

El estimador REML es una modificación del estimador ML propuesto para disminuir el sesgo, el cual reemplaza la maximización de la verosimilitud del vector Z, por la del vector $A'Z$ tal que: $E(A'Z) = 0$, donde A es una matriz de dimensión n x $(n - p)$, de rango columna completo.

Con esta modificación y dado que

$$Var(A'Z) = A'\Sigma(\theta)A$$

el negativo de la función de logverosímil queda:

$$L(\theta) = \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|A'\Sigma(\theta)A| + \frac{1}{2}Z'(A'\Sigma(\theta)A)^{-1}A'Z$$

Esta función es dependiente únicamente de θ; de esta manera este método no usa la modelización de la superficie de tendencia, sino que se basa directamente en un vector de incrementos de media 0. No obstante, pese a minimizar el sesgo en las estimaciones de θ, aun interviene un costo computacional elevado [Bohorquez, 2020].

Verosimilitud Compuesta (CL)

El método CL que se utiliza para estimar θ, involucra la suma de componentes individuales de funciones logverosimilitud, correspondiente a las distribuciones marginales de las variables de interés. Por consiguiente, la distribución multivariada de Z no es necesaria, puesto que se basa en las distribuciones marginales $f(Z(s_i);\theta)$ y se asume la existencia del gradiente y de la matriz Hessiana de f.

Ahora bien, se suponen conocidas $f(Z(s_i);\theta)$, excepto el parámetro θ; entonces:

$$\log(Z(s_i);\theta) = \ln(f(Z(s_i);\theta))$$

es una función logverosímil y la función de verosimilitud compuesta está definida por:

$$CL(\theta) = \sum_{i=1}^{n}\log(Z(s_i);\theta)$$

Al gradiente de CL, $\nabla CL = CS(\theta)$, se le conoce como función de score compuesta.

Así, los valores estimados de $\hat{\theta}$ se determinan resolviendo el siguiente sistema de ecuaciones:

$$CS(\theta) = \sum_{i=1}^{n}\nabla\log(Z(s_i);\theta) = 0$$

Una de las razones principales para usar las funciones de verosimilitudes marginales es que aunque al inicio no se cumpla el supuesto requerido, es posible aproximarse a este

por medio de alguna transformación [Bohorquez, 2020].

Mínimos cuadrados ponderados

Este método utiliza la matriz de ponderación $W(\theta)$ y suele ser expresado en términos del variograma o de la covarianza, gracias a su equivalencia en procesos estacionarios de segundo orden. Para el caso espacio-temporal, se estima θ para un variograma minimizando la siguiente expresión:

$$(2\hat{\gamma} - 2\gamma(\theta))' W^{-1}(\theta)(2\hat{\gamma} - 2\gamma(\theta))$$

donde la matriz de ponderación $W(\theta)$ está dada por:

$$W(\theta) = Diag(Var(2\hat{\gamma}(h_k))) \approx Diag\left(\frac{2(2\gamma(h_k|\theta))^2}{N(h_k)}\right)$$

con:

$$N(h_k) = \{(i,j) : s_i - s_j = h_k\}$$

Para las ubicaciones $i, j = 1, ..., n$, que producen los primeros k rezagos espaciales $k = 1, ..., K$; generalmente se utilizan los rezagos espaciales hasta la mitad de la distancia máxima entre cualquier par de ubicaciones, debido a que para ubicaciones bastante separadas disminuye notoriamente la cantidad de puntos incluidos en la estimación del variograma. La aproximación de $Var(2\hat{\gamma}(h_k|\theta))$, se obtiene bajo el supuesto de que $Z(s) \sim N(\mu; \sigma^2), \forall s \in \mathbb{R}^d$, y que por consiguiente:

$$(Z(s+h) - Z(s))^2 \approx 2\gamma(h)\chi_1^2$$

El inconveniente que presenta este método es la necesidad de definir clases de rezagos para realizar una estimación empírica de la covarianza o del variograma, pues si no se poseen muchos datos la cantidad de estos en cada una de las clases disminuye, de tal forma que para los primeros rezagos se pueden tener suficientes datos para la estimación de cada $\hat{\gamma}(h_k)$; sin embargo, para los últimos rezagos podrían no existir suficientes datos para la estimación [Bohorquez, 2020].

Modelo lineal de regionalización

En la práctica, puede suceder que el variograma empírico no presente una apariencia o forma simple de ser modelizado por un modelo teórico; no obstante, esto no es un inconveniente, puesto que existe la posibilidad de combinar modelos teóricos de variograma, obteniendo nuevos modelos más complejos.

Teniendo presente que un fenómeno regionalizado podría ser considerado como la suma de diversos subfenómenos independientes, el modelo lineal de regionalización construye un campo aleatorio $Z(s)$ como una combinación lineal de L campos aleatorios independientes, cada uno con media 0 y función de covarianza $C_l(h)$, mutuamente independientes [Bohorquez, 2020].

Así, sea Y_l un subcampo aleatorio de Z para todo $l = 1, ..., L$, se tiene que:

$$Z(s) = \sum_{l=0}^{L} a_l Y_l(s + \mu)$$

con:

$$\begin{aligned} E(Z(s)) &= \mu \\ E(Y_l(s)) &= 0 \\ Cov(Y_l(s), Y_{l'}(s+h)) &= \begin{cases} C_l(h) & si\ l = l' \\ 0 & en\ otro\ caso. \end{cases} \end{aligned}$$

Esto implica que la función de covarianza de $Z(s)$ es una combinación lineal de las L funciones $C_l(h)$ y bajo el supuesto de independencia entre los Y, se tiene que:

$$\begin{aligned} C(h) &= Cov(Z(s), Z(s+h)) \\ &= \sum_{l=0}^{L} \sum_{l'=0}^{L} Cov((Y_l(s), Y_{l'}(s+h))) \\ &= \sum_{l=0}^{L} a_l a_l C_l(h) \\ &= \sum_{l=0}^{L} a_l^2 C_l(h) \end{aligned}$$

por lo tanto, el modelo de convarianza $C(h)$ es:

$$C(h) = \sum_{l=0}^{L} b_l C_l(h),\ con\ b_l = a_l^2 \geq 0$$

De esta forma, como $b_l > 0$ es siempre positivo, y es la silla del modelo básico de covarianza $C_l(h)$. Las condiciones suficientes para que $C(h)$ sea un modelo válido de covarianza son:

- Las funciones $C_l(h)$ son modelos de convarianza válidos.
- $b_l > 0,\ \forall_l = 1, ..., L$

En términos de variogramas, se tiene que:

$$E\left[(Y_l(s) - Y_l(s+h))(Y_{l'}(s) - Y_{l'}(s+h))\right] = \begin{cases} \gamma_l'(h) & si\ l = l' \\ 0 & en\ otro\ caso. \end{cases}$$

Entonces,

$$\gamma(h) = \frac{1}{2} E[Z(s+h) - Z(s)]^2 = \sum_{l=0}^{L} b_l \gamma_l'(h)$$

con $b_l = (a_l)^2 \geq 0$, donde b_l es la contribución en varianza del correspondiente variograma $\gamma_l'(h)$ [Bohorquez, 2020].

2.1.4.5. Modelos teóricos de variograma

En esta sección se presentan las funciones de variograma más comunes, que se definen para el caso isotrópico de funciones aleatorias. Para la representación gráfica de la función de variograma se hace uso de la relación $\gamma(h) = C(0) - C(h)$. Estas funciones se pueden clasificar de la siguiente manera:

- Variogramas con silla o variogramas de transición.
- Variogramas con silla y efecto hueco.
- Variogramas sin silla.

Variogramas con silla

- **Modelo esférico**

  Este modelo es válido solo en $\mathbb{R}^1, \mathbb{R}^2, \mathbb{R}^3$, y está definido por:

$$\gamma(\|h\|;\theta) = \begin{cases} \sigma^2 \left(1{,}5\frac{\|h\|}{\phi} - 0{,}5\left(\frac{\|h\|}{\phi}\right)^3\right) & si\ \|h\| \leq \phi \\ \sigma^2 & si\ \|h\| > \phi \end{cases}$$

  con $\theta = (\sigma^2, \phi)$, donde $\sigma^2 = C(0)$ es el valor del variograma donde alcanza la silla, y ϕ es el rango.

  Este variograma presenta un comportamiento lineal cerca del origen, lo cual sugiere continuidad pero con cierto grado de irregularidad en el proceso estocástico espacial o campo aleatorio. Sin embargo, referente a su comportamiento en distancias largas, alcanza su silla en $\|h\| = \phi$. El uso habitual de este modelo en la práctica es gracias a su rango bien definido, su representación polinomial simple y su validez en $\mathbb{R}^1, \mathbb{R}^2, \mathbb{R}^3$. La principal razón para el uso de este es que se comporta de forma casi lineal hasta que alcanza su silla, y presenta estabilidad en una gran variedad de observaciones.

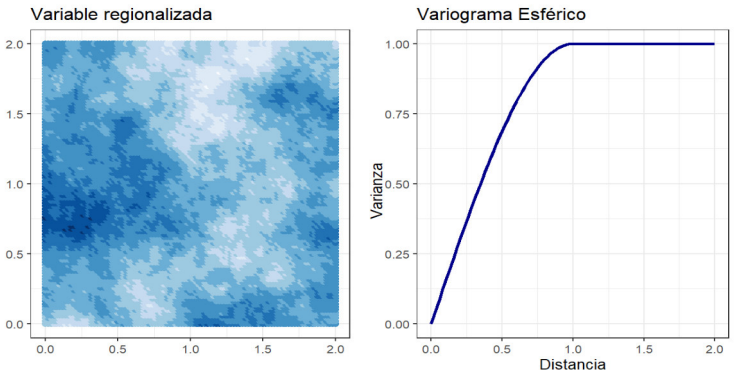


Figura 2.12: Izquierda: Variable regionalizada. Derecha: Variograma esférico.

- **Modelo de efecto pepita puro**

  Este variograma refleja la ausencia de dependencia espacial en el proceso estocástico o campo aleatorio; este modelo puede ser visto como un caso particular del modelo esférico si $\phi \to 0$:

$$\gamma(\|h\|;\theta) = \begin{cases} \sigma^2 & si\ \|h\| = 0 \\ 0 & si\ \|h\| > 0 \end{cases}$$

  con $\theta = \sigma^2$.

  Se debe tener en cuenta que el modelo esférico corresponde a una función aleatoria continua, mientras que el modelo de efecto pepita puro, a una discontinuidad.

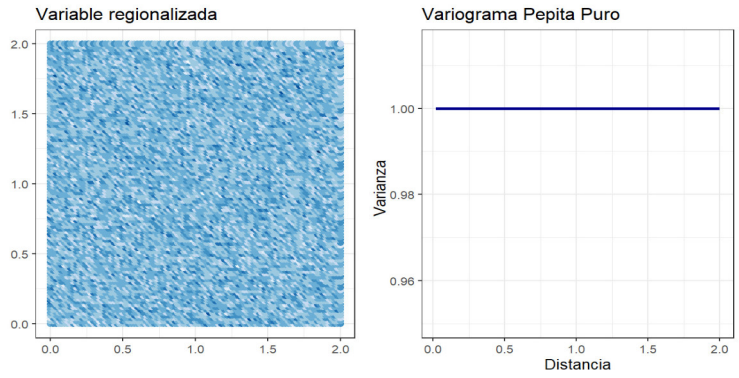


Figura 2.13: Izquierda: Variable regionalizada. Derecha: Variograma pepita puro.

- **Modelo exponencial**

  Este modelo es válido en $\mathbb{R}^d,\ d \geq 1$, y está definido por:

$$\gamma(\|h\|;\theta) = \sigma^2 \left(1 - \exp\left(-\frac{\|h\|}{\phi}\right)\right), \quad \theta = (\sigma^2, \phi)$$

  El modelo exponencial refleja un comportamiento lineal cerca del origen, lo cual indica continuidad pero con un cierto grado de irregularidad en la función aleatoria. Por otro lado, en cuanto a su comportamiento en largas distancias, este alcanza su silla solo asintóticamente cuando $\|h\| \to \infty$. El valor del rango es igual a la distancia para el cual el variograma toma un valor igual al 95 % de la silla.

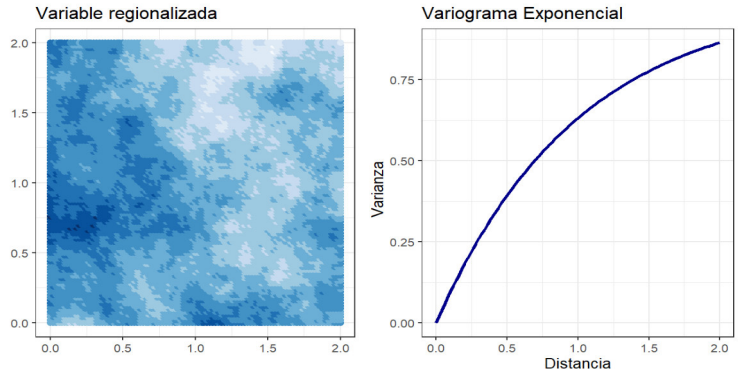


Figura 2.14: Izquierda: Variable regionalizada. Derecha: Variograma exponencial.

- **Modelo Gaussiano**

  Este modelo es válido en $\mathbb{R}^d,\ d \geq 1$, y está definido por:

$$\gamma(\|h\|;\theta) = \sigma^2\left(1 - \exp\left(-\frac{\|h\|^2}{\phi^2}\right)\right), \quad \theta = (\sigma^2, \phi)$$

  La principal característica de este modelo es su forma parabólica cerca al origen; la dependencia espacial se desvanece solo en una distancia que tiende a infinito; por consiguiente, este modelo es considerado poco realista en la práctica.

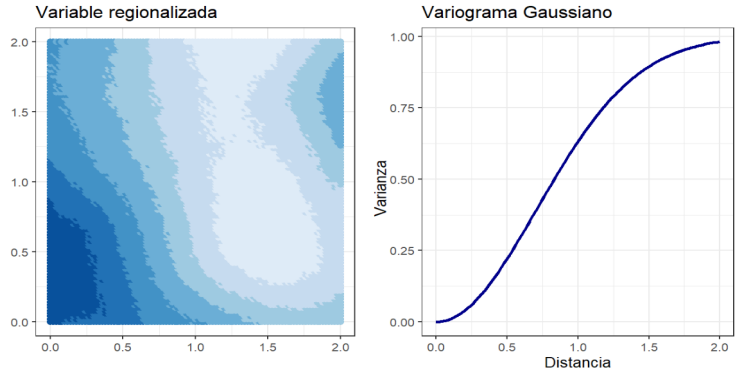


Figura 2.15: Izquierda: Variable regionalizada. Derecha: Variograma Gaussiano.

- **Modelo Cúbico**

  Este modelo es válido en $\mathbb{R}^1, \mathbb{R}^2, \mathbb{R}^3$; generalmente es similar al modelo Gaussiano, puesto que también presenta un comportamiento parabólico cerca del origen. Sin embargo, este modelo alcanza una silla plana a una distancia ϕ, y esta definido por:

$$\gamma(\|h\|;\theta) = \begin{cases} \sigma^2\left(7\frac{\|h\|^2}{\phi^2} - \frac{35\|h\|^3}{4\phi^3} + \frac{7\|h\|^5}{2\phi^5} - \frac{3\|h\|^7}{4\phi^7}\right) & si\ 0 \leq \|h\| \leq \phi \\ \sigma^2 & si\ \|h\| > \phi \end{cases}$$

  con $\theta = (\sigma^2, \phi)$.

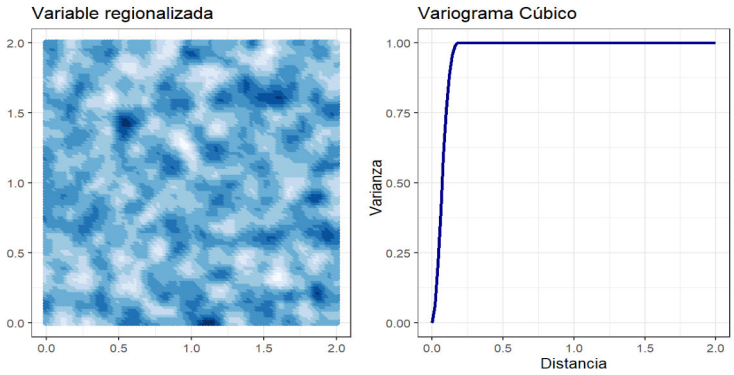


Figura 2.16: Izquierda: Variable regionalizada. Derecha: Variograma cúbico.

- **Modelo Estable**

  Este modelo es válido en $\mathbb{R}^d,\ d \geq 1$ y está definido por:

$$\gamma(\|h\|;\theta) = \sigma^2\left(1 - \exp\left(-\left(\frac{\|h\|}{\phi}\right)^{\alpha}\right)\right), \quad 0 < \alpha \leq 2,\ \theta = (\sigma^2, \phi)$$

  Notar que para $\alpha = 1$, se obtiene el modelo exponencial y con $\alpha = 2$, se convierte en un modelo Gaussiano.

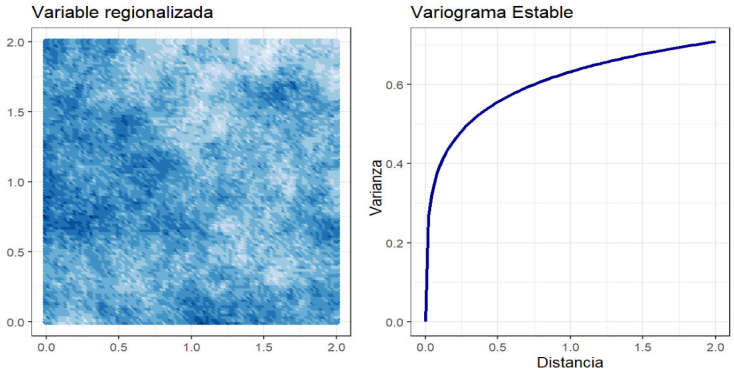


Figura 2.17: Izquierda: Variable regionalizada. Derecha: Variograma estable.

- **Modelo de Cauchy Generalizado**

  Este modelo es válido en $\mathbb{R}^d,\ d \geq 1$, y está definido por:

$$\gamma(\|h\|;\theta) = \sigma^2\left(1 - \frac{1}{\left(1 + \left(\frac{\|h\|}{\phi}\right)^2\right)^{\alpha}}\right), \quad \theta = (\sigma^2, \phi)$$

  Si $\alpha = 1$ es conocido como el modelo de Cauchy.

  Este modelo muestra un comportamiento parabólico cerca del origen y si $\alpha < 2$ alcanza la silla lentamente.

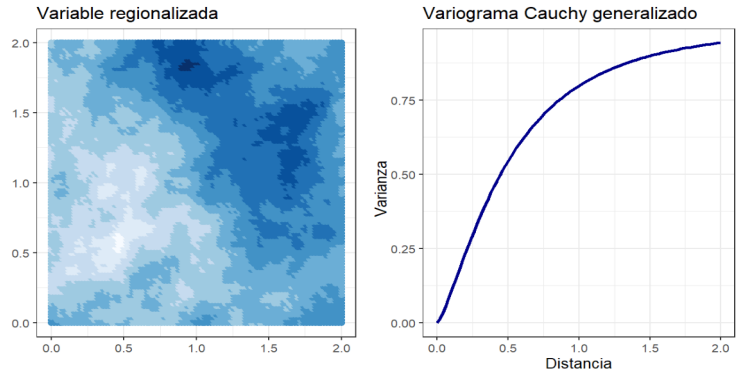


Figura 2.18: Izquierda: Variable regionalizada. Derecha: Variograma Cauchy generalizado.

- **Modelo de K-Bessel o Matérn**

  Este modelo es válido en $\mathbb{R}^d, \quad d \geq 1$, y esta definido por:

$$\gamma(\|h\|;\theta) = \sigma^2 \left(1 - \frac{1}{2^{\alpha-1}\Gamma(\alpha)} \left(\frac{\|h\|}{\phi}\right)^{\alpha} K_\alpha \left(\frac{\|h\|}{\phi}\right)\right), \quad \alpha > 0,\ \theta = (\sigma^2, \phi)$$

  donde K_α es la función de Bessel modificada de segundo tipo de orden α, y que está definida por:

$$K_\alpha(\nu) = \frac{\pi}{2\sin(\alpha\pi)} \left(\sum_{k=0}^{\infty} \frac{1}{k!\Gamma(-\alpha+k+1)} \left(\frac{\nu}{2}\right)^{2k-\alpha} - \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(\alpha+k+1)} \left(\frac{\nu}{2}\right)^{2k+\alpha}\right)$$

  El modelo K-Bessel puede tener cualquier tipo de comportamiento cerca del origen. Para $\alpha = 1/2$, se obtiene el modelo exponencial.

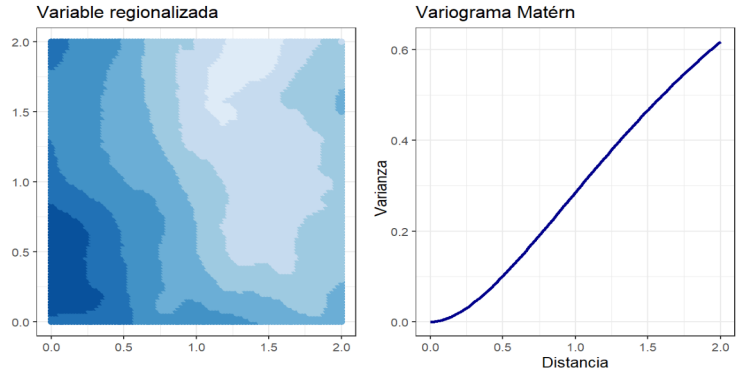


Figura 2.19: Izquierda: Variable regionalizada. Derecha: Variograma Matérn.

Variogramas con efecto hueco

En la práctica la dependencia espacial puede no crecer monótonamente, incluso esta podría ser negativa o presentar alternaciones entre dependencia espacial positiva y negativa. Estos modelos se llaman *modelos de efecto hueco* y se utilizan para definir oscilaciones

que involucran un significado físico. Presentan un comportamiento lineal o parabólico cerca del origen, pueden tener o no tener silla, y pueden ser o no periódicos o pseduperiódicos.

- **Modelo J-Bessel**

  Este modelo está definido por:

$$\gamma(\|h\|;\theta) = \sigma^2\left(1 - \left(\frac{2\phi}{\|h\|}\right)^{\alpha}\Gamma(\alpha+1)J_\alpha\left(\frac{\|h\|}{\phi}\right)\right), \quad \theta = (\sigma^2, \phi)$$

  donde α es el parámetro de forma, ϕ es el parámetro de escala, Γ es la función de Euler, y J_α es la función $J - Bessel$ del primer tipo de orden α dada por:

$$J_\alpha(\nu) = \left(\frac{\nu}{2}\right)^2 \sum_{k=0}^{\infty} \frac{-1^k}{k!\Gamma(\alpha+k+1)}\left(\frac{\nu}{2}\right)^{2k}$$

  y es válido para $\mathbb{R}^d, d \leq 2(\alpha+1)$.

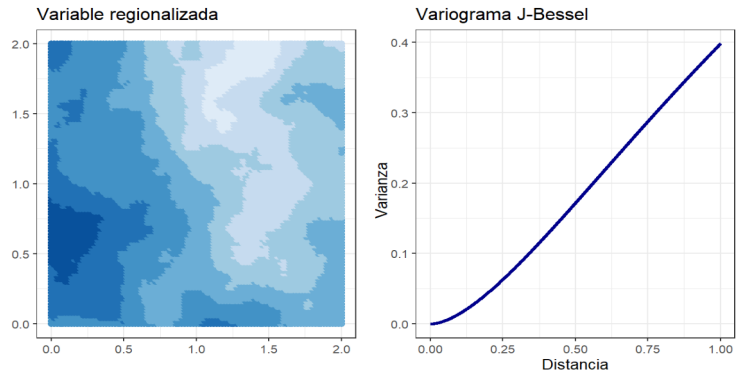


Figura 2.20: Izquierda: Variable regionalizada. Derecha: Variograma J-Bessel.

- **Modelo seno cardinal**

  Este modelo es una particularización del modelo J-Bessel para $\alpha = 1/2$, y es uno de los pocos modelos de efecto hueco válidos en $\mathbb{R}^3$, y está definido por:

$$\gamma(\|h\|;\theta) = \sigma^2\left(1 - \frac{\phi}{\|h\|}\sin\left(\frac{\|h\|}{\phi}\right)\right), \quad \theta = (\sigma^2, \phi)$$

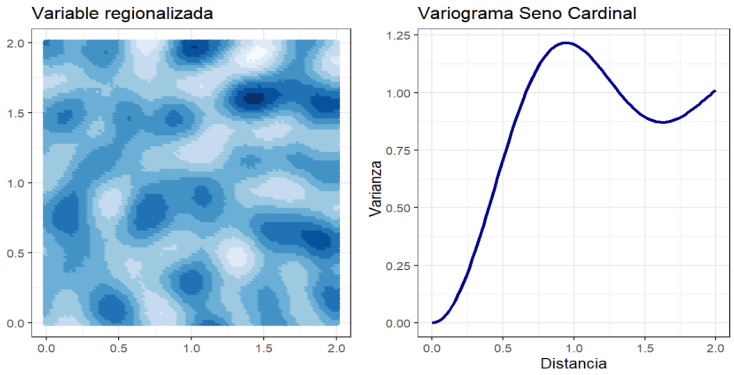


Figura 2.21: Izquierda: Variable regionalizada. Derecha: Variograma seno cardinal.

Variogramas sin silla

Estos modelos corresponden a procesos estocásticos espaciales o campos aleatorios que son intrínsecamente estacionarios, pero no son estacionarios de segundo orden. Estos modelos son no acotados y corresponden a un proceso estocástico espacial o campo aleatorio con capacidad ilimitada para la dispersión espacial y, por consiguiente, ni su covarianza, ni su varianza pueden ser definidas.

- **Modelo Potencia**

  Este modelo es válido en $\mathbb{R}^d, d \geq 1$ y está definido por:

$$\gamma(\|h\|) = (\|h\|)^{\alpha}, \quad 0 < \alpha < 2,$$

  Este variograma no posee ni silla ni rango, sino que crece indefinidamente. Si α es cercano a cero, se dice que es un variograma de efecto pepita, si α es cercano a dos tiene un comportamiento parabólico y si α es igual a uno presenta un comportamiento lineal. Para $\alpha \geq 2$ la condición $\lim_{h\to\infty} \frac{\gamma\|h\|}{h^2} = 0$ no se satisface; es decir, el modelo no es intrínsecamente estacionario.

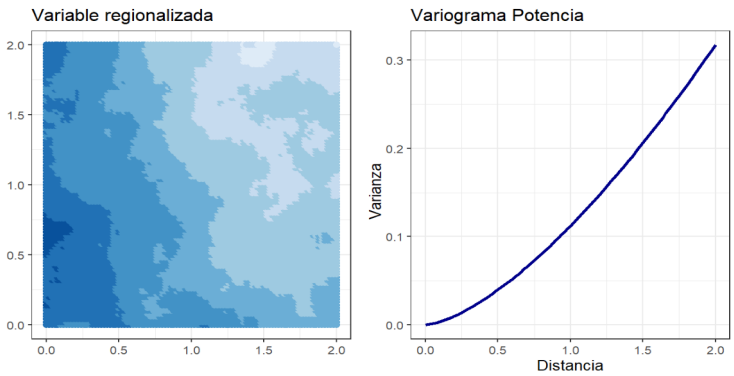


Figura 2.22: Izquierda: Variable regionalizada. Derecha: Variograma potencia.

- **Modelo lineal**

Este modelo es un caso especial del modelo potencia si $\alpha = 1$, y está asociado con la estacionariedad intrínseca, pero no con la estacionariedad de segundo orden y está definido por:

$$\gamma(\|h\|) = \begin{cases} 0 & si\ \|h\| = 0 \\ \|h\| & si\ \|h\| > 0 \end{cases}$$

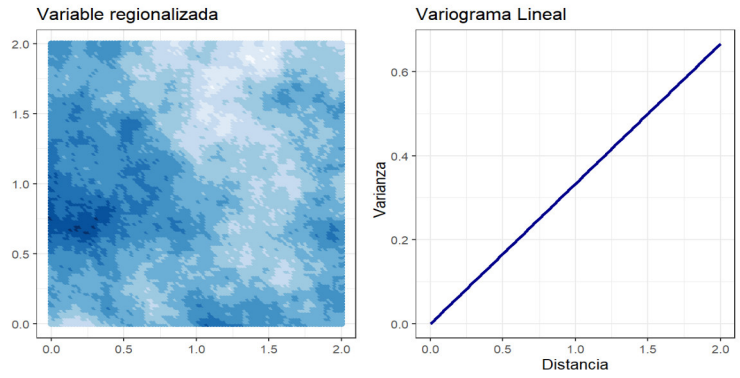


Figura 2.23: Izquierda: Variable regionalizada. Derecha: Variograma lineal.

2.1.4.6. Caso Multivariante

El variograma multivariado fue formalizado en 1991 como una aplicación a procesos estocásticos espaciales o campos aleatorios estacionarios. Para un vector de p variables regionalizadas estacionarias y para toda métrica M, el covariograma multivariado y el variograma multivariado son definidos como:

- Covariograma multivariado:

$$K(h) = E[(Z(s) - \boldsymbol{\mu})M(Z(s+h) - \boldsymbol{\mu})']$$

- Variograma multivariado:

$$\Gamma(h) = E[(Z(s) - Z(s+h))M(Z(s) - Z(s+h))']$$

donde $Z(s)$ es un vector fila de p procesos estocásticos espaciales o campos aleatorios estacionarios de segundo orden, $\boldsymbol{\mu} = E[Z(s)]$ y M es una matriz simétrica definida positiva de tamaño $p \times p$ utilizada como métrica en el cálculo de similaridades [Bourgault *et al.*, 1992].

Asumiendo estacionariedad de segundo orden, la función de autocovarianza multivariada está relacionada con el variograma multivariado por medio de:

$$K(h) = \Gamma(\infty) - \Gamma(h)$$

donde $\Gamma(\infty)$ es la silla del variograma multivariado. Si $\Gamma(0) = 0$ entonces $K(0) = \Gamma(h)$.

El variograma multivariado representa la esperanza matemática de una medida de disimilitud cuadrática, y la función de autocovarianza multivariada representa la esperanza matemática de una medida de similaridad multivariante [Bourgault *et al.*, 1992].

La estimación del variograma multivariado se realiza promediando las disimilitudes multivariadas al cuadrado, de forma similar al variograma tradicional; es decir:

$$\Gamma^*(h) = \frac{1}{2|N(h)|} \sum_{N(h)} d_{ij}^2$$

donde d_{ij} es la disimilitud entre las muestras i y j calculadas con una métrica dada y $N(h)$ como se definió anteriormente [Bourgault *et al.*, 1992].

Ahora bien, para realizar el ajuste de los modelos teóricos de variogramas y variogramas cruzados, se utiliza el Modelo Lineal de Coregionalización.

Modelo lineal de coregionalización (MLC)

Para el cálculo del MLC es necesario introducir las siguientes medidas:

- **Covarianza cruzada:**

$$C_{rr'}(h) = [E(Z_r(s) - E(Z_r(s)))]\,[E(Z_{r'}(s+h) - E(Z_{r'}(s+h)))]$$

- **Correlación cruzada**

$$\rho_{rr'} = \frac{C_{rr'}(h)}{\sigma_r \sigma_{r'}}$$

- **Variograma cruzado**

$$\gamma_{rr'}(h) = \frac{1}{2} E\,[Z_r(s+h) - Z_r(s)]\,[Z_{r'}(s+h) - Z_{r'}(s)]$$

- **El pseudo variograma cruzado**

$$\phi_{rr'} = \frac{1}{2} E\,[Z_r(s+h) - Z_{r'}(s)]^2$$

- **Co-dispersión**

$$v_{rr'}(h) = \frac{\gamma_{rr'}(h)}{\sqrt{\gamma_{rr}(h)}\sqrt{\gamma_{r'r'}(h)}} \in [-1, 1]$$

Al igual que en el caso univariante, el variograma cruzado posee las siguientes propiedades: es invariante bajo traslación, $\gamma_{rr'}(0) = 0$, $\gamma_{rr'}(h) = \gamma_{rr'}(-h)$. Además, esta medida es la más utilizada en las aplicaciones geoestadísticas.

Ahora bien, sea $\mathbb{Z} = (Z_1, ..., Z_p)$ un proceso estocástico o campo aleatorio espacial multiavariante; se tiene que el MLC es una suma proporcional de modelos de covarianza o variogramas. En notación matricial donde $C(h) = [C_{ij}(h)]$ es una matriz de covarianza de dimensión $p \times p$ y de manera similar $\Gamma(h) = [\Gamma_{ij}(h)]$, con

$$C(h) = \sum_{k=1}^{L} B_k C_k(h)$$
$$\Gamma(h) = \sum_{k=1}^{L} B_k \gamma_k(h)$$

donde cada B_k se conoce como *matriz de coregionalización* y:

$$C_{ij}(h) = \sum_{k=1}^{L} b_k(i,j) C_k(h) \quad \forall i, j = 1, ..., p$$
$$\Gamma_{ij}(h) = \sum_{k=1}^{L} b_k(i,j) \gamma_k(h) \quad \forall i, j = 1, ..., p$$

El MLC asume que todos los variogramas simples y cruzados pueden ser expresados como una combinación lineal de modelos básicos (exponencial, Gaussiano, esférico, entre otros) idénticos indexados por k. Una condición suficiente para que el modelo sea válido es que cada una de las matrices B_k sean definidas positivas.

Por construcción del modelo, todos las covarianzas cruzadas son simétricas, y los variogramas cruzados son modelos de variograma válidos.

Ajuste del MLC

El ajuste del MLC puede ser realizado mediante el método de mínimos cuadrados, al igual que en el caso univariante, teniendo en cuenta que la única diferencia es que la silla de cada elemento es reemplazada por una matriz de sillas.

El ajuste es llevado a cabo por el algoritmo propuesto por Goulard (1989) quienes implementaron este algoritmo de manera iterativa para asegurar la positividad de los coeficientes B_k.

Se definen matricialmete el variograma empírico y variograma teórico multivariado como sigue:

$$\hat{\Gamma(h)} = [\hat{\gamma}_{ij}(h)]$$
$$\Gamma(h) = [\gamma_{ij}(h)] = \sum_{k} B_k \gamma_k(h)$$

el criterio de bondad de ajuste es la suma de cuadrados ponderados (SCP) de todos los términos de la matriz de errores $\hat{\Gamma}(h) - \Gamma(h)$ y se suman sobre el conjunto de retardos J utilizados para el ajuste; es decir:

$$SCP = \sum_{k \in J} w(h) \text{traza} \left\{ \left[V(\hat{\Gamma}(h) - \Gamma(h)) \right]^2 \right\}$$

Los ponderadores $w(h)$ son positivos y usualmente iguales al número de pares utilizados para la estimación del variograma en el retardo h. La matriz V es definida positiva y diseñada para equilibrar la influencia de las variables; por lo general es la diagonal de la matriz o inversa de la matriz de varianzas o la matriz identidad. La idea es minimizar el criterio optimizando un B_k a la vez y repitiendo esto hasta que no haya mejora alguna. El residuo para el ajuste actual menos el $k-$ésimo término es:

$$d\Gamma_k(h) = \hat{\Gamma}(h) - \sum_{u \neq k} B_u \gamma_u(h)$$

En ausencia de restricción de positividad, el ajuste óptimo de $d\Gamma_k$ por $B_k\gamma_k(h)$ se obtiene por medio de la cancelación de la derivada de SCP en relación a B_k

$$\frac{\partial SCP}{\partial B_k} = -2V \left[\sum_{h \in J} w(h)\gamma_k(h) \left(d\Gamma_k(h) - \gamma_k(h) B_k \right) \right] V = 0$$

Como V es no singular se tiene que:

$$B_k = \frac{1}{\alpha_k} \sum_{h \in J} w(h)\gamma_k(h) d\Gamma_k(h)$$

donde

$$\alpha_k = \sum_{h \in J} w(h)\gamma_k(h)^2$$

La solución restringida $B_k^+ \geq 0$ es la matriz definida positiva cercana a B_k de acuerdo a la norma definida por V. Dado que es simétrica, la matriz B_k tiene descomposición de la siguiente forma

$$B_k = U_k \nabla_k U_k' \quad \text{con} \quad U_k' V U_k = I_p$$

donde U_k es una matriz de vectores propios de $B_k V$ y ∇_k es la matriz diagonal de sus valores propios.

Por tanto, la solución restringida es

$$B_k' = U_k \nabla_k^+ U_k'$$

donde ∇_k^+ es la matriz ∇_k en el cual los valores propios negativos son reemplazados por ceros. Este algoritmo iterativo siempre converge y la solución no depende del punto de partida [P.J. Diggle, 2007].

2.2. Análisis de datos funcionales (FDA)

Los objetivos fundamentales del FDA son los mismos que la estadística convencional e incluyen: (a) Representar y transformar los datos de manera que faciliten el análisis posterior. (b) Mostrar los datos de manera que se destaquen las variables en estudio. (c) Estudiar fuentes relevantes de patrones y variación de los datos. (d) Explicar la variación de una variable resultante o dependiente utilizando la información de la variable independiente [Ramsay & Silverman, 2001].

En la estadística elemental las observaciones son escalares, pero en la estadística multivariante las observaciones son vectores que pertenecen al espacio Euclídeo $\mathbb{R}^d$, donde d es la dimensión de los vectores observados. Sin embargo, en el FDA las observaciones son funciones o curvas y aunque estas pueden ser medidas en puntos discretos se las pueden pensar como funciones en espacios infinitos dimensionales [Kokoszka, 2017].

Los datos funcionales por tanto son una generalización natural de datos multivariantes, pasando de una dimensión finita a una dimensión infinita [Zhang, 2013]. La alta dimensionalidad de los datos, representa un reto para la teoría y para el cálculo computacional, donde estos retos varían dependiendo de como se tomaron las muestras. Sin embargo, la alta o infinita dimensionalidad de los datos, son una fuente rica en información, lo que conlleva a poder realizar investigaciones y análisis de datos mas robustos [Wang *et al.*, 2016].

La extensa variedad de aplicaciones y herramientas hacen que definir de forma precisa el FDA sea de cierta manera difícil. El FDA surge cuando una variable de interés, en un conjunto de datos, puede ser naturalmente vista como una curva o función suave; entonces el FDA puede pensarse como un análisis de alto nivel estadístico sobre muestras de curvas [Kokoszka, 2017]. En la práctica, los datos funcionales son obtenidos por medio de observaciones sobre el tiempo, espacio u otros dominios continuos; los datos funcionales resultantes pueden ser curvas, superficies u objetos complejos [Zhang, 2013].

Las funciones o curvas pueden ser vistas como realizaciones de un proceso estocástico uno dimensional; usualmente se asume que estas realizaciones pertenecen al espacio de *Hilbert*.

2.2.1. Espacios de Hilbert

El espacio de *Hilbert* generaliza la noción de espacio euclidiano, extendiendo las propiedades de dimensión finita para dimensión infinita.

El Espacio L^2

El espacio $L^2 = L^2[0,1]$ es el conjunto de funciones reales medibles de Lebesgue f definidas en $[0,1]$, tales que satisfacen [Young, 2014]:

$$\int_0^1 f^2(t)dt < \infty$$

Este espacio se le conoce como el *Espacio de Funciones Cuadrado Integrables*.

Si $f, g \in L^2$, la igualdad $f = g$, significa que [Kokoszka, 2017]:

$$\int [f(t) - g(t)]dt = 0$$

Las funciones cuadrado integrables forman un espacio vectorial; esto significa que, si $f, g \in L^2$, entonces se cumple que, para cualquier escalar a, b::

$$af + bg \in L^2$$
$$(af + bg)(t) = af(t) + bg(t),\ t \in [0, 1]$$

Se debe tener en cuenta que los elementos de L^2 y operaciones sobre estos no necesitan estar definidos para todo punto t, sino que están definidos para *casi todo punto* t [Kokoszka, 2017].

Para funciones $f, g \in L^2$, se define el producto interno de dos funciones como:

$$\langle f, g \rangle = \int f(t)g(t)dt$$

El producto interno permite entender la noción de ortogonalidad, lo cual hace que la estructura geométrica del espacio L^2 sea intuitivamente muy similar al espacio euclidiano finito dimensional.

Se dice que dos funciones f, g son ortogonales si:

$$\langle f, g \rangle = 0$$

El producto interno permite definir la noción de distancia entre funciones a través de la norma:

$$\|f\| = \sqrt{\langle f, f \rangle} = \left\{ \int f^2(t)dt \right\}^{\frac{1}{2}}$$

Por otro lado, la distancia entre dos funciones en el espacio L^2 es la norma de su diferencia; es decir:

$$d(f, g) = \|f - g\|$$

Un conjunto de funciones $\{e_1, e_2, e_3, ...\}$ es una base en L^2 si toda función $f \in L^2$ admite una única expansión de la forma:

$$f(t) = \sum_{j=1}^{\infty} a_j e_j(t)$$

y se dice que $\{e_1, e_2, e_3, ...\}$ es una base ortonormal, si:

$$\langle e_j, e'_j \rangle = 0,\ j \neq j'\ y\ \|e_j\| = 1$$

Para una base ortonormal, como por ejemplo la base de Fourier, se tiene que:

$$a_j = \langle f, e_j \rangle$$

y tenemos la igualdad de *Parseval*:

$$\int f^2(t)dt = \|f\|^2 = \sum_{j=1}^{\infty} \langle f, e_j \rangle^2 = \sum_{j=1}^{\infty} \left\{ \int f(t)e_j(t)dt \right\}^2$$

Funciones Aleatorias Cuadrado Integrables

Dado que Z denota una función aleatoria en $\mathbb{R}$. Así, como una variable aleatoria, una función aleatoria es definida en un espacio de probabilidad, digamos Ω; para cada $\omega \in \Omega$, $Z(\omega)$ es una función determinística. Se asume que cada una de las realizaciones $Z(\omega), \omega \in \Omega$, son elementos del espacio L^2 de funciones cuadrado integrables, lo que significa que para cada $\omega \in \Omega$, se verifica que:

$$\|Z(\omega)\|^2 = \int \{Z(\omega)(t)\}^2 < \infty$$

La función $\|Z(\omega)\|$ es por tanto una variable aleatoria; si esta variable aleatoria tiene momento de segundo orden finito; es decir, $E\|Z\|^2 < \infty$, por consiguiente Z es una función aleatoria cuadrado integrable.

Es importante notar la diferencia entre una función cuadrado integrable determinística, donde la integración se define en el intervalo $[0, 1]$, y las funciones aleatorias cuadrado integrables, donde la integración está definida en el espacio de probabilidad Ω [Kokoszka, 2017].

Estimación de la función de media y covarianza

En la práctica, se observa una muestra que consiste de N curvas $Z_1, Z_2, ...Z_N$; se puede ver cada curva como una realización de una función aleatoria Z, o como un elemento aleatorio de L^2 con la misma distribución de Z. Comúnmente se asume que los Z_i son independientes, en particular si las curvas surgen de mediciones sobre sujetos seleccionados aleatoriamente de una población grande.

Ahora bien, suponiendo que $Z_1, Z_2, ..., Z_N$ son *iid* en L^2, y que tienen la misma distribución de Z, la cual se asume de cuadrado integrable. De esta forma se define lo siguiente:

- Función de media:

$$\mu(t) = E[Z(t)]$$

- Función de media estimada:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} Z_i(t)$$

- Función de covarianza:

$$c(t, s) = E[(Z(t) - \mu(t))(Z(s) - \mu(s))]$$

- Función de covarianza estimada:

$$\hat{c}(t,s) = \frac{1}{N}\sum_{i=1}^{N}(Z_i(t) - \hat{\mu}(t))(Z_i(s) - \hat{\mu}(s))$$

- Operador de covarianza:

$$C = E[\langle (Z-\mu), \cdot \rangle (Z-\mu)]$$

- Estimación del operador de covarianza:

$$\hat{C}(z) = \frac{1}{N}\sum_{i=1}^{N}\langle Z_i - \hat{\mu}, z\rangle (Z_i - \hat{\mu}), \quad z \in L^2$$

Nótese que $\hat{C}$ proyecta el espacio L^2 en un subespacio finito dimensional generado por $Z_1, Z_2, ..., Z_N$, mostrando las limitaciones de la inferencia estadística para observaciones funcionales; una muestra finita puede cubrir un objeto infinito dimensional solo con reducida exactitud [Young, 2014].

2.2.2. Reducción de Dimensionalidad

Como se mencionó, un dato funcional es intrínsecamente infinito dimensional, incluso si un gran número de datos son medidos de manera discreta sobre un conjunto finito de cualquier intervalo la dimensionalidad es alta, produciendo dificultades en el análisis de datos [Kokoszka, 2017], como la ralentización de algoritmos estadísticos convencionales o incluso produciendo que algunos de estos sean inviables. En algoritmos de clasificación, el costo computacional y la precisión de clasificación pueden ser mejorados de manera significativa utilizando un subconjunto de mediciones representativas; una razón es que instantes en el tiempo consecutivos usualmente están altamente correlacionados, incluyendo redundancias que pueden incrementar el nivel de ruido y ocultar información útil acerca de la estructura de los datos [Tian, 2010].

Para tratar el problema de alta dimensionalidad, se hace uso de funciones base, las que constituyen una herramienta para almacenar información acerca de las funciones y da la flexibilidad para combinar el poder computacional para ajustar cientos y miles de observaciones. Además, permite expresar los cálculos requeridos mediante matrices [Ramsay J., 2005].

El conjunto de datos más simple, en el contexto de FDA, es una muestra de la forma:

$$z_n(t_{j,n}) \in \mathbb{R}, t_{j,n} \in [T_1, T_2], \;\; n = 1, ..., N, \;\; j = 1, ..., J_n$$

es decir, tenemos N curvas que son observadas en un intervalo común $[T_1, T_2]$. Los valores de las curvas nunca son conocidos en todos los puntos $t \in [T_1, T_2]$, pues solo están disponibles en puntos específicos $t_{j,n}$, los cuales pueden ser diferentes para curvas z_n diferentes. Algunas aplicaciones relevantes del FDA, tratan con situaciones donde el

número de puntos $\{t_{j,n}\}$, por curva son pequeños. La idea fundamental del FDA es que los objetos en estudio sean curvas suaves:

$$\{z_n(t) : t \in [T_1, T_2], n = 1, ..., N\}$$

para los cuáles los valores $z_n(t)$ existen en todo punto t, pero solo son observados en puntos específicos t_j.

El enfoque principal en el FDA es la forma de las funciones observadas o de las funciones que resumen las propiedades de los datos en un sentido específico [Kokoszka, 2017].

Bases de expansión

En la práctica se trabaja con funciones con características que pueden ser impredecibles y complicadas, por lo que se requiere de estrategias para construir funciones que trabajen con parámetros que sean fáciles de estimar y que puedan ser lo más cercano posible a cualquier característica de la curva.

Se utiliza un conjunto de bloques funcionales $\phi_k, k = 1, .., K$ llamados *funciones base*, los cuales son combinados linealmente; es decir, una función $z(t)$ definida en este sentido es expresada de la siguiente forma:

$$z(t) = \sum_{k=1}^{K} c_k \phi_k(t) = c'\phi(t)$$

conocido como *expansión en funciones base*. Los parámetros $c_1, ...c_k$ son coeficientes de la expansión. La expresión matricial usa c para referirse al vector de K coeficientes y ϕ para denotar al vector de tamaño K que contiene las funciones suaves de base que comparten las mismas propiedades. [James Ramsay, 2009].

Usualmente se considera una muestra de funciones de tamaño N:

$$z_i(t) = \sum_{k=1}^{K} c_{ik} \phi_k(t), \;\; i = 1, ..., N$$

en este caso, la notación matricial está dada por:

$$\mathbf{z(t)} = C\phi(t)$$

donde $\mathbf{z(t)}$ es un vector de tamaño N conteniendo las funciones $z_i(t)$, y la matriz de coeficientes C tiene N filas y K columnas [James Ramsay, 2009].

Lo ideal, es que las funciones base tengan características que coincidan con las características de las funciones que se están estimando, lo que permite la obtención de una aproximación adecuada utilizando un número más pequeño que K de funciones; cuánto menor sea K y mejor reflejen las funciones base ciertas características de los datos: se tendrá más grados de libertad para realizar pruebas de hipótesis y calcular intervalos de confianza más precisos, menos cálculo computacional es requerido y es factible que los mismos coeficientes logren describir los datos [Ramsay J., 2005].

El concepto de sistemas de bases no es nuevo; por ejemplo, un polinomio de cualquier orden, no es más que una combinación lineal de las funciones base monomial

$\{1, t, t^2, ..., t^n\}$; no obstante, los polimonios no son muy útiles cuando se trabajan con formas funcionales complejas, por lo que, sistemas de base de tipo splines y Fourier son ampliamente utilizados en la práctica [James Ramsay, 2009].

Bases de Fourier

Probablemente la expansión de base más conocida está dada por las series de Fourier. Esta serie es especialmente útil para funciones extremadamente estables; es decir, funciones en las que no hay características locales fuertes y donde la curva tiende a ser del mismo orden en todas partes. Estas series suelen dar lugar a expansiones uniformes, sin embargo son inadecuadas en cierto grado para los datos que se sospecha que tienen discontinuidades en la función o en las derivadas de bajo orden[Ramsay J., 2005]. En algunas ocasiones, es necesario que las funciones se puedan repetir sobre un cierto período de tiempo T; por ejemplo, pueden ser requeridos para expresar comportamiento estacional en series de tiempo extensas. La serie de Fourier, está dada por [James Ramsay, 2009]:

$$\begin{aligned}\phi_1(t) &= 1\\ \phi_2(t) &= \sin(\omega t)\\ \phi_3(t) &= \cos(\omega t)\\ \phi_4(t) &= \sin(2\omega t)\\ \phi_5(t) &= \cos(2\omega t)\\ &\vdots\end{aligned}$$

donde la constante ω está relacionado con el período T por la relación:

$$\omega = \frac{2\pi}{T}$$

Se observa que, después de la primera función base constante, las funciones base de Fourier son ordenadas de forma sucesiva en pares de seno/coseno, los dos con argumentos en cualquier par que se multiplique por uno de los enteros $1, 2, ...$ hasta un límite superior m. Si las series contienen ambos elementos de cada par, como es usual, el número de funciones base es $K = 1 + 2m$, puesto que, de la forma en que se definió ω, cada función base se repite por si misma después de T unidades de tiempo transcurrido.

Para definir una base de tipo Fourier, solo son necesario dos parámetros: el número de funciones base K y el período T; no obstante, este último valor usualmente puede ser por defecto el rango de los valores t en los cuales están definidos los datos[James Ramsay, 2009].

Estimación de curvas utilizando sistemas de bases por mínimos cuadrados

Se tiene por objetivo ajustar observaciones discretas $y_j, j = 1, ..., n$ utilizando el modelo $y_j = x(t_j) + \epsilon_j$ y se utiliza una base de funciones para la expansión de $x(t)$ de la forma:

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) = \mathbf{c}'\boldsymbol{\phi}$$

donde el vector $\mathbf{c}$ de tamaño K contiene los coeficientes c_k y $\boldsymbol{\phi}$ la matriz de orden $n \times K$ que contiene los valores $\phi_k(t_j)$.

El suavizamiento básico se obtiene si se determinan los coeficientes de la expansión c_k minimizando el criterio de mínimos cuadrados:

$$SMSSE(y|c) = \sum_{j=1}^{n}[y_j - \sum_{k=1}^{K} c_k\phi_k(t_j)]^2$$

en términos matriciales la expresión anterior está dada de la siguiente manera:

$$SMSSE(y|c) = (y - \Phi c)'(y - \Phi c) = \|y - \Phi c\|^2$$

luego, tomando el criterio de derivada de $SMSSE(y|c)$ con respecto a $\mathbf{c}$ tenemos:

$$2\Phi\Phi' c - 2\Phi' y = 0$$

resolviendo para $\mathbf{c}$ tenemos la estimación $\hat{\mathbf{c}}$ que minimiza la solución de mínimos cuadrados:

$$\hat{c} = (\Phi'\Phi)^{-1}\Phi' y$$

así, el vector $\hat{y}$ de valores estimados es:

$$\hat{y} = \Phi\hat{c} = \Phi(\Phi'\Phi)^{-1}\Phi' y$$

Este método básico de aproximación es adecuado cuando se asume que los residuos ϵ_j sobre la curva verdadera son independientes e idénticamente distribuidas con media cero y varianza σ^2 constante.

Ajuste por mínimos cuadrados ponderados

En la práctica se trabaja con errores no estacionarios y/o autocorrelacionados, por lo que se usa el método de mínimos cuadrados ponderados como una extensión del método anterior de la forma:

$$SMSSE(y|c) = (y - \Phi c)'W(y - \Phi c)$$

donde W es una matriz simétrica definida positiva que permite la ponderación de cuadrados y producto de los residuos.

Si la matriz de varianza-covarianza Σ_e para los residuos ϵ_j se conoce, entonces:

$$W = \Sigma_e^{-1}$$

en aplicaciones donde una estimación completa de Σ_e no es factible, los valores de la covarianza de los errores son usualmente asumidos iguales a cero, entonces W es diagonal con los recíprocos de la varianza de los errores asociados con los y_j.

La estimación $\hat{c}$ de los coeficientes del vector c está dada por:

$$\hat{c} = (\Phi' W \Phi)^{-1} \Phi' W' y$$

Suavizamiento de datos funcionales mediante el uso de bases de Fourier

Una función f continua a trozos en el intervalo de $[-\pi, \pi]$, puede ser representada en un sistema de tipo Fourier de la siguiente manera:

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx)$$

donde, para todo k se tiene que:

$$a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos kx dx$$
$$b_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin kx dx$$

esta representación de f se conoce como la expansión en series de Fourier en el intervalo $[-\pi, \pi]$ y donde a_k, b_k son conocidos como los coeficientes de Euler-Fourier de f [Kreider, 1971.] y este tipo de expansión se utiliza para aproximar datos periódicos.

El suavizamiento de datos, utilizando una base de Fourier, siempre implica que se tenga más puntos de los necesarios para modelizar la suavidad de la función en estudio. En la práctica, la frecuencia de la serie no sobrepasa una frecuencia de corte; es decir, que a partir de cierto término, digamos m, todos los coeficientes, a_k, b_k, de Fourier serán prácticamente cero; por lo tanto, la ecuación anterior, puede escribirse de la siguiente manera:

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{m} (a_k \cos kx + b_k \sin kx)$$

esta ecuación se conoce como la expansión de Fourier truncada y es la función suavizada de nuestros datos.

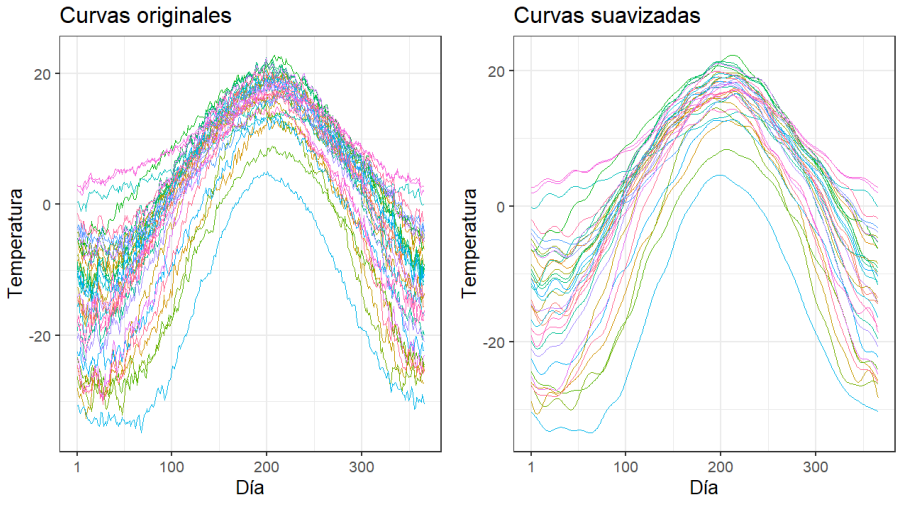


Figura 2.24: Suavización de curvas.

Usualmente, el intervalo donde una función está definida no es $[-\pi, \pi]$ o $[0, \pi]$; también podría ser un intervalo $[a, b]$ arbitrario, por lo cual la expansión en serie de Fourier queda de la siguiente forma:

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \left(a_k \cos \frac{2k\pi x}{b-a} + b_k \sin \frac{2k\pi x}{b-a} \right)$$

donde, para todo k:

$$a_k = \frac{2}{b-a} \int_a^b f(x) \cos \frac{2k\pi x}{b-a} dx$$
$$b_k = \frac{2}{b-a} \int_a^b f(x) \sin \frac{2k\pi x}{b-a} dx$$

2.2.3. Detección de atípicos

El tratamiento de datos atípicos es un aspecto importante en cualquier análisis estadístico, pese a que los datos atípicos influyen de manera significativa en metodologías estadísticas en varios sentidos, su análisis en datos funcionales ha sido poco abordado [Febrero-Bande & de la Fuente, 2012].

De manera intuitiva, se podría pensar en la estadística multivariante para tratar casos de datos atípicos en muestras de datos funcionales. Desafortunadamente, hay varios motivos por las que la estadística multivariante falla en el caso funcional. La primera es que los datos funcionales son realizaciones de un proceso aleatorio suave medido en un conjunto discreto de tiempo, por lo cual la estructura de correlación temporal es ignorada cuando se utilizan métodos de estadística multivariante. La segunda es que por la naturaleza infinto-dimensional, los métodos de estadística multivariante son afectados por la alta dimensionalidad; es decir, los métodos no son capaces de manejar las situaciones en donde el número de mediciones de una o más variables es más grande que el número de individuos en la muestra. La tercera es que pocas asunciones sobre la distribución son impuestas sobre los conjuntos de datos funcionales, mientras que los métodos mutlivariantes están implícitamente sujetos a poblaciones Gaussianas [Febrero-Bande & de la Fuente, 2012].

Los datos atípicos en un conjunto de datos funcionales pueden surgir principalmente por dos razones. La primera es que las curvas pueden tener graves errores de medición, registro o de digitación; estos errores deben ser identificados y corregidos siempre que sea posible. La segunda es que los datos atípicos pueden ser curvas de datos que expresan la realidad, en el sentido de que no tienen errores graves; sin embargo, no se comportan como el resto de curvas. Por tanto, se considera que una curva es un dato atípico si ha sido generado por un proceso estocástico con diferente distribución que el resto de curvas, las cuales se asumen que son idénticamente distribuidas. Una curva puede ser un dato atípico si esta está significativamente alejada del proceso estocástico o si tiene diferente forma y/o comportamiento del resto de curvas [Febrero-Bande & de la Fuente, 2012].

Para identificar datos atípicos en un conjunto de datos funcionales, se hace uso de la *profundidad funcional*. Este concepto fue originalmente introducido en el análisis multivariante para medir la centralidad de un punto en relación con la nube de puntos; es decir, la profundidad da una forma de ordenar los puntos en un espacio Euclídeo desde el centro hacia el exterior; así, los puntos cercanos al centro tendrán una profundidad más grande [Di Blasi *et al.*, 2013]. Por otro lado, para datos funcionales, si una curva

es un dato atípico, esta curva tendrá una profundidad baja, por lo que para detectar la presencia de datos atípicos funcionales basta analizar las curvas con profundidades más bajas [Febrero-Bande & de la Fuente, 2012]. Existen tres medidas principales para calcular la profundidad funcional, las cuales son: *Profundidad de Fraiman y Muniz*, *Profundidad H-modal*, y *Profundidad de proyección aleatoria*. Para mayor detalle revisar [Febrero *et al.*, 2008].

Para la visualización de datos atípicos se han desarrollado varias herramientas gráficas, principalmente propuestas para el caso univariante; sin embargo, para el caso multivariante han sido poco desarrolladas [Dai & Genton, 2017].

2.2.3.1. Gráfico de Magnitudes y Formas

Una de la herramientas gráficas usadas es el **MS-Plot** o **Gráfico de Magnitudes y Formas** propuesto por [Dai & Genton, 2017], el cual se basa en el marco de la periferia funcional direccional, que mide la centralidad de los datos funcionales considerando el nivel y la dirección de sus desviaciones de la región central.

Periferia Funcional Direccional

Teniendo en cuenta que la externalidad es la medida, grado o cualidad de ser atípico; a este concepto se añade el de dirección, ya que la dirección de externalidad es crucial para describir la centralidad de datos funcionales multivariantes.

De manera formal, la externalidad direccional de un dato escalar está definida por:

$$O(Y, F_Y) = \left\{ \frac{1}{d(Y, F_Y) - 1} \right\} v, \quad d(Y, F_Y) > 0$$

donde F_Y es la distribución de una variable aleatoria, d es una medida convencional de profundidad, y v es el vector unitario que va desde la mediana de F_Y a Y. Asumiendo que Z es la única mediana de F_Y para la medida de profundidad d, v puede expresarse como:

$$v = \frac{Y - Z}{\|Y - Z\|}$$

donde $\|.\|$ es la norma L_2.

Ahora bien, si X es una función p-dimensional definida en un dominio $\mathcal{I}$, a partir de la distribución de los datos funcionales F_X, para cada punto fijado $t \in \mathcal{I}$, $F_{X(t)}$ es la función de distribución de $X(t)$ con dimensión p. De esta forma, se definen tres medidas de externalidad direccional para datos funcionales:

1) Media de externalidad direccional:

$$MO(X, F_X) = \int_{\mathcal{I}} O(X(t), F_{X(t)}) w(t) dt$$

2) Variación de externalidad direccional:

$$VO(X, F_X) = \int_{\mathcal{I}} \|O(X(t), F_{X(t)}) - MO(X, F_X)\|^2 w(t) dt$$

3) Externalidad direccional funcional:

$$F0(X, F_X) = \int_{\mathcal{I}} \|O(X(t), F_{X(t)})\|^2 w(t) dt$$

donde $w(t)$ es una función de peso definida en $\mathcal{I}$, la cual puede ser constante o proporcional a la variación local en cada punto de diseño; en particular, se usa $w(t) = \{\lambda(\mathcal{I})\}^{-1}$, donde $\lambda(.)$ es la medida de Lebesgue.

Las tres medidas anteriores de externalidad pueden escribirse en una sola ecuación de la siguiente manera:

$$FO = \|MO\|^2 + VO$$

La ecuación anterior realiza una descomposición total de la externalidad funcional **(FO)** en dos términos: la magnitud de externalidad ($\|\boldsymbol{MO}\|$) y la cantidad de variación de externalidad direccional **(VO)**. Esta descomposición permite una gran flexibilidad para describir la centralidad de un conjunto de datos funcionales y para identificar curvas potencialmente atípicas [Dai & Genton, 2017].

MS-Plot

El ***MS-Plot*** es un gráfico de dispersión de puntos, $(M0', VO)'$, para un grupo de datos funcionales. Cuando la dimensión es más grande que dos, se usa $(\|MO\|, VO)'$, la cual representa la magnitud general de externalidad y la forma de externalidad sin información de la dirección [Dai & Genton, 2017].

De las definiciones de MO Y VO, se esperaría que:

- Las curvas centrales se asignen a la región central baja del MS-Plot; es decir, $\|MO\|$ pequeño y VO pequeño.
- Valores atípicos desplazados se sitúan en la región inferior; es decir, $\|MO\|$ grande y VO pequeño, y las diferentes direcciones de MO indican las diferentes direcciones de sus desplazamientos.
- Los valores atípicos aislados, que están en una pequeña parte del espacio de soporte, se asignan a la región superior central; es decir, $\|MO\|$ pequeño y VO grande.
- Los puntos en la región superior exterior corresponden a las curvas que son sustancialmente atípicas tanto en magnitud como en forma; es decir, $\|MO\|$ grande y VO grande.

2.2.4. ANOVA Para datos funcionales

En la práctica se vuelve un problema decidir si existen o no diferencias en el proceso de interés cuando varían condiciones que pueden afectarlo. Este problema se suele abordar mediante un modelo que supone la existencia de una función base que describe la evolución típica del proceso estudiado, suponiendo que los datos con los que se trabaja se han obtenido agregando variaciones aleatorias a esta función. Entonces, este problema se convierte en un tipo de problema de ANOVA funcional [Cuesta-Albertos & Febrero-Bande, 2010].

Existen varios procedimientos para tratar problemas de tipo ANOVA. En el presente trabajo se tratará el problema de *one-way* ANOVA para datos funcionales univariantes, conocido también como el problema de la *prueba de k-muestras*.

2.2.4.1. *One-Way* ANOVA

Este problema puede ser formulado de la siguiente manera:

- Sea $Z_{i1}(t), Z_{i2}(t), .., Z_{in_i}(t),\ i = 1, ..., k$, donde k hace referencia a los k grupos de funciones aleatorias definidas sobre un intervalo finito $T = [a, b]$ dado.
- Sea $SP(m, \gamma)$ un proceso estocástico con función de media $m(t),\ t \in T$ y función de covarianza $\gamma(s, t), st, t \in T$.
- Asumiendo que $Z_{i1}(t), Z_{i2}(t), .., Z_{in_i}(t)$ son $SP(m_i, \gamma), i = 1, .., k$ i.i.d.

Es interesante comprobar la igualdad de las k funciones de media; es decir:

$$H_0 : m_1(t) = ... = m_k(t),\ t \in T$$
$$H_a : m_1(t) \neq .... \neq m_k(t),\ t \in T$$

A continuación se presentan varias pruebas para este tipo de ANOVA.

CS-Test

El siguiente estadístico fue propuesto en [Cuevas *et al.*, 2004]:

$$V_n = \sum_{1 \leq i < j \leq k} n_i \int_T (\bar{Z}_i(t) - \bar{Z}_j(t))^2 dt$$

Bajo la hipótesis nula y la asunción de que :

$$n_i, n \to \infty \quad \text{tal que} \quad \frac{n_i}{n} \to p_i > 0,\ i = 1, ..., k$$

se demuestra que la distribución aproximada de V_n es la del estadístico:

$$V = \sum_{1 \leq i < j \leq k} n_i \int_T \left(\bar{Y}_i(t) - \sqrt{\frac{p_i}{p_j}} Y_j(t) \right)^2 dt$$

donde $Y_1(t), ..., Y_k(t)$ son procesos Gaussianos independientes de media cero y función de covarianza $\gamma(s,t)$. El valor crítico empírico se calcula mediante remuestreo sobre $Y_i(t), i = 1, .., k$ de un proceso Gaussiano con media cero y función de covarianza $\hat{\gamma}$, definida por:

$$\hat{\gamma}(s,t) = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{ij}(s) - \bar{Z}_i(s))(Z_{ij}(t) - \bar{Z}_i(t))$$

Para el caso heterocedástico y mediante el cálculo del valor crítico empírico, el valor p de V_n puede calcularse como en el caso anterior, mediante procesos Gaussianos independientes con funciones de covarianza dadas por:

$$\hat{\gamma}_i(s,t) = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(Z_{ij}(s) - \bar{Z}_i(s))(Z_{ij}(t) - \bar{Z}_i(t))$$

Prueba basadas en la norma $\mathcal{L}^2$

Este test usa el siguiente estadístico:

$$S_n = \int_T SSR_n(t)dt$$

donde:

$$SSR_n(t) = \sum_{i=1}^{k} n_i(\bar{Z}_i(t) - \bar{Z}(t))^2$$
$$\bar{Z}(t) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Z_{ij}(t)$$
$$\bar{Z}_i(t) = \frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ij}(t)$$

Bajo la hipótesis nula, se tiene que $S_n \sim \beta\chi^2_d$ aproximadamente, donde:

$$\beta = \frac{tr(\gamma^{\otimes 2})}{tr(\gamma)},$$
$$d = (k-1)\kappa,$$
$$\kappa = \frac{tr^2(\gamma)}{tr(\gamma^{\otimes 2})},$$
$$\gamma^{\otimes 2}(s,t) = \int_T \gamma(s,u)\gamma(u,t)du.$$

Esta distribución aproximada es usada para calcular el p-valor de S_n; es decir, $P(\chi^2_d \geq S_n/\beta)$ o su valor crítico $\beta\chi^2_d(1-\alpha)$. Los parámetros β y κ son estimados con base en los datos funcionales por el método de *naive* o el método de *reducción de sesgo*.

Utilizando el método naive, y con el estimador $\hat{\gamma}(s,t)$, se tiene que:

$$\hat{\beta} = \frac{tr(\hat{\gamma}^{\otimes 2})}{tr(\hat{\gamma})},$$
$$\hat{d} = (k-1)\hat{\kappa},$$
$$\hat{\kappa} = \frac{tr^2(\hat{\gamma})}{tr(\hat{\gamma^{\otimes 2}})}.$$

Utilizando el método de reducción de sesgo, se tiene que:

$$\hat{\beta} = \frac{\widehat{tr(\gamma^{\otimes 2})}}{tr(\hat{\gamma})},$$
$$\hat{d} = (k-1)\hat{\kappa},$$
$$\hat{\kappa} = \frac{\widehat{tr^2(\gamma)}}{\widehat{tr(\gamma^{\otimes 2})}},$$
$$\widehat{tr(\gamma^{\otimes 2})} = \frac{(n-k)^2}{(n-k-1)(n-k+2)}\left(tr(\hat{\gamma}^{\otimes 2}) - \frac{tr^2(\hat{\gamma})}{n-k}\right),$$
$$\widehat{tr^2(\gamma)} = \frac{(n-k)(n-k+1)}{(n-k-1)(n-k+2)}\left(tr^2(\hat{\gamma}) - \frac{2tr(\hat{\gamma}^{\otimes 2})}{n-k+1}\right).$$

Prueba F

Este test usa el estadístico:

$$F_n = \frac{\int_T SSR_n(t)dt/(k-1)}{\int_T SSE_n(t)dt/(n-k)}$$
$$SSE_n(t) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{ij}(t) - \bar{Z}_i(t))^2$$

Bajo la hipótesis nula, se tiene que $F_n \sim F_{d_1,d_2}$ aproximadamente, donde $d_1 = (k-1)\kappa$ y $d_2 = (n-k)\kappa$. De manera similar, la distribución aproximada de la prueba F puede ser usada para calcular el p-valor de F_n o su valor crítico; y el parámetro κ puede ser estimado por el método naive o el método de reducción de sesgo.

Cuando las muestras nos son Gaussianas o son pequeñas, estas pruebas no son utilizadas; en este caso, las versiones bootstrap de estas pruebas son usadas para calcular el p-valor de S_n y F_n [Górecki & Smaga, 2015].

Prueba basada en representación de funciones base

Se puede utilizar una función de representación de base para obtener una prueba más robusta.

Asumiendo que se trabaja con k grupos de un proceso estocástico $Z_{ij} \in L_2(T), i = 1..., k, j = 1, ..., n_i$ y sea $\{\phi_l\}$ una base ortonormal de $L_2(T)$.

Esta prueba usa el siguiente estadístico:

$$F = \frac{\frac{1}{k-1}\sum_{i=1}^{k} n_i \|\bar{Z}_i - \bar{Z}\|^2}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i} \|Z_{ij} - \bar{Z}_i\|^2}$$

donde el proceso estocástico, la media funcional muestral y media funcional grupal, escrita en notación matricial a partir de la representación en la base ϕ, están definidos de la siguiente manera:

$$\begin{aligned} Z_{ij}(t) &= c'_{ij}\phi(t), \\ \bar{Z}(t) &= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} c'_{ij}\phi(t), \quad \forall t \in T \\ \bar{Z}_i(t) &= \frac{1}{n_i}\sum_{j=1}^{n_i} c'_{ij}\phi(t) \end{aligned}$$

para todo $t \in T$.

El numerador del estadístico F, mide la variabilidad externa entre las diferentes muestras, mientras que el denominador mide la variabilidad interna dentro de las muestras [Cuevas *et al.*, 2004].

2.3. Agrupación

De manera empírica, agrupación significa encontrar grupos en un conjunto de datos. La primera agrupación en la historia fue de tipo jerárquico, cuando Aristóteles clasificó a los seres vivos. El análisis de agrupación ha sido desarrollado en diferentes campos y con diversas aplicaciones [Hennig *et al.*, 2015].

En el análisis de agrupación el objetivo es examinar datos que no contienen etiquetas, para luego encontrar agrupaciones de los datos; se debe tener en cuenta que se desconoce el número de grupos. Esta metodología es una forma de aprendizaje no supervisado; es decir, no se utilizan datos de entrenamiento. [Andrew B. Lawson, 2002].

Formalmente, si se tiene un conjunto de datos, $D = \{z_1, z_2, ..., z_n\}$, el cual contiene n datos, la tarea de los métodos de agrupación es asignarlos a K subconjuntos disjuntos de D, denotados por $C_1, C_2, ..., C_K$ [Hennig *et al.*, 2015]. Estos métodos con frecuencia se utilizan como un paso preliminar en la exploración de datos para identificar patrones que tengan una interpretación útil para el usuario [Jacques, 2014].

2.3.1. Agrupación de Datos Funcionales

Como se mencionó, el objetivo principal del análisis de agrupación es construir grupos homogéneos de observaciones que representen realizaciones de alguna variable aleatoria Z, y se busca que las observaciones asignadas a los grupos sean similares entre ellas y lo más diferente posible de las observaciones asignadas a otros grupos [Tarpey & Kinateder, 2003]. En el espacio finito dimensional, Z es un vector aleatorio con valores en $\mathbb{R}^p$, $Z = (Z_1, ..., Z_p), p \geq 1$. Un caso particular es cuando las variables aleatorias toman valores en un espacio infinito dimensional, usualmente un espacio de funciones definidas en algún

conjunto continuo $\mathcal{T}$; de esta forma los datos son representados por medio de curvas y la variable aleatoria que corresponde a los datos es un proceso estocástico $Z = \{Z(t), t \in \mathcal{T}\}$. En la actualidad este tipo de datos son más fáciles de observar y se han desarrollado herramientas para poderlos almacenar y procesar [Jacques, 2014].

Se conocen cuatro metodologías para la agrupación de datos funcionales, las cuales son:

- Métodos de datos en bruto: Estos métodos consisten en agrupar directamente las curvas sobre la base de sus puntos de evaluación.
- Métodos de filtrado: Estos métodos realizan la aproximación de las curvas en alguna base de funciones y luego realizan la agrupación utilizando los coeficientes de la expansión en la base.
- Métodos adaptativos: Estos métodos consideran que la representación de un dato funcional depende del grupo, y se realiza simultáneamente la reducción de dimensión y la agrupación.
- Métodos con base en distancias: Estos métodos usan algoritmos de agrupación con base en distancias específicas para datos funcionales.

En el presente trabajo se hará uso de los métodos de filtrado y los métodos basados en distancias.

2.3.1.1. Métodos de filtrado

Este método consiste de dos etapas para la agrupación de datos funcionales. La primera hace referencia al método de filtrado, el cual reduce la dimensión de los datos, pues consiste generalmente en aproximar las curvas utilizando una base finita de funciones; esta base puede ser de tipo B-splines, Fourier o Wavelets; otra técnica de reducción de dimensión es el análisis de componentes principales funcionales (FACP). La segunda etapa consiste en que a partir de los coeficientes de la base de expansión o de los *scores* de los componentes principales, se aplican algoritmos habituales de agrupación para determinar los grupos [Jacques, 2014].

Se debe tener en cuenta que el método de filtrado tiene algunos inconvenientes; uno de ellos es que si las curvas son medidas en diferentes puntos de tiempo, la varianza de la estimación de los coeficientes de base es distinto para cada individuo. Para conjuntos de datos dispersos, varios coeficientes de base podrían tener varianza infinita, haciendo imposible obtener estimaciones razonables [James & Sugar, 2003].

2.3.1.2. Métodos con base en distancias

Estos métodos también son conocidos como métodos no paramétricos y se dividen en dos categorías, la primera agrupa los métodos que hacen uso de distancias o disimilitudes específicas, mientras que la segunda hace uso de heurísticas.

Métodos que usan distancias

Estos métodos aplican técnicas de agrupación no paramétrica como el método k-medias o el método jerárquico, los cuales consideran distancias específicas o disimilitudes

entre curvas. Para este propósito se utiliza una métrica basada en derivadas que mide la proximidad en las curvas z_i y z_i' mediante la siguiente expresión:

$$d_q(z_i, z_i')^2 = \int \left(z_i^{(q)}(t) - z_{i'}^{(q)}(t)\right)^2 dt$$

donde $z^{(q)}$ denota la *q-ésima* derivada de z y $d_0(z, 0)$ es la norma $\mathcal{L}^2$ [Hall, 2007]. Si el cálculo de d_0 se lo realiza utilizando las observaciones discretas de la curva, los métodos no paramétricos son equivalentes a los métodos de agrupación de datos brutos. Por otro lado, si el cálculo de d_0 se lo realiza mediante la representación de las curvas en una base finita, estos métodos son equivalentes a los métodos de filtrado [Jacques, 2014].

Métodos Heurísticos

Estos métodos proponen usar heurísticas para la agrupación de datos funcionales. En [Hébrail *et al.*, 2010] se desarrollaron dos algoritmos de programación dinámica que realizan la agrupación y la estimación de los centros de cada grupo simultáneamente. Por otro lado, en [Yamamoto, 2012] se desarrolló un procedimiento para identificar simultáneamente grupos óptimos de funciones y subespacios óptimos para la agrupación; para esto se define una función objetivo como la suma de las distancias entre las observaciones y sus proyecciones más las distancias entre las proyecciones y las medias de los grupos [Jacques, 2014].

2.3.2. Agrupación de Datos Espaciales

El incremento de los datos espaciales y el uso extendido de bases de este tipo de datos ponen en evidencia la necesidad de generar procesos que permitan la extracción de información espacial; en este sentido se ha desarrollado la minería de datos espaciales, debido a que esta permite descubrir patrones interesantes y previamente desconocidos pero potencialmente útiles [Sumathi *et al.*, 2008]. A medida que se dispone de más y más datos sobre un área espacial es deseable identificar las diferentes funciones y papeles que desempeñan las distintas partes de esta área; en particular, un objetivo bastante deseable es identificar regiones homogéneas y descubrir sus características, creando resúmenes de alto nivel para los datos en estudio y obteniendo información valiosa para planificadores, científicos, entre otros [Cesario *et al.*, 2020].

Los datos espaciales poseen dos atributos distintos: uno espacial y otro no espacial. Los atributos espaciales de este tipo de datos incluyen información relacionada a la ubicación espacial como longitud, latitud, elevación, forma, entre otros. Por otro lado, los atributos no espaciales se utilizan para describir características como nombre, población, tasa de desempleo, entre otros. Se debe tener en cuenta que la información espacial referente a las relaciones entre los datos, tales como la influencia de los datos con los de sus vecinos, usualmente está implícita; por lo cual, para capturar esta información se hace uso de técnicas o instrumentos que la incorporen [Shekhar *et al.*, 2003]. Para este propósito se hace uso de las siguientes técnicas:

- Generalizaciones basadas en descubrimiento de conocimiento.
- Métodos de agrupación.
- Medidas de proximidad agregada.

- Reglas de asociación espacial.

El presente trabajo se enfocará en los métodos de agrupación. Este clase de métodos han sido desarrollados por la estadística a lo largo del tiempo, en primera instancia se trabajó en una dimensión, y se centraron en la búsqueda de medidas de tendencia, variabilidad o dispersión. Luego se extendió para dos dimensiones, que tratan atributos con media o propiedades modales y variabilidad o dispersión, estos atributos son bidimensionales; en el caso de datos espaciales estas cantidades tiene interpretación como la ubicación del centro del grupo y la dispersión del grupo [Andrew B. Lawson, 2002].

Los métodos clásicos de agrupación utilizados para particionar un conjunto de datos espaciales dan como resultado grupos que espacialmente están mezclados o carecen de sentido [Ambroise & Dang, 1997]; para dar solución a este problema, se han desarrollado métodos que toman en cuenta la información espacial de los datos; la aplicación de estos incluyen la identificación de áreas con características o factores similares y el objeto de estudio depende del campo de aplicación.

Una solución que ha sido utilizada a lo largo de los años es ponderar las disimilitudes entre las muestras mediante la función de variograma; es decir:

$$d_{ij}^{*} = d_{ij}\gamma(h)$$

donde $\gamma(h)$ es el variograma ajustado al variograma empírico $\hat{\gamma}(h)$; d_{ij} es la disimilitud entre las muestras i y j [Bourgault *et al.*, 1992].

Por otro lado, en el caso multivariante, en cuanto a similaridad, se tiene que la ponderación está realizada mediante la función de autocovarianza multivariada; es decir:

$$S_{ij}^{*2} = S_{ij}^{2}K(h)$$

Esta ecuación es más robusta en el sentido de que favorece a la formación de grupos que son espacialmete homogéneos.

Por otro lado, en cuanto a disimilitud y utilizando la función de variograma multivariado se tiene que:

$$d_{ij}^{*2} = d_{ij}^{2}\Gamma(h)$$

Se debe tener en cuenta que esta última ecuación es un poco más general ya que permite variogramas multivariados sin silla. Ambas ecuaciones se desempeñan adecuadamente cuando el efecto pepita está presente [Bourgault *et al.*, 1992].

Una vez que los cálculos son realizados para todos los i y j el resultado es una matriz de disimilitud modificada D^{*} con elementos d_{ij}^{*}. Esta matriz se utiliza por una amplia variedad de estrategias de clasificación; por ejemplo, el método jerárquico opera directamente en esta matriz, agrupando primero aquellos individuos para los cuales d_{ij}^{*} es la menor y luego la menos similar según las reglas del algoritmo en uso. Por otro lado, la agrupación dinámica es más apropiada para agrupar individuos de poblaciones que no presentan estructura jerárquica [Oliver & Webster, 1989].

A partir de esta ponderación se pueden aplicar algunos métodos de agrupación que se clasifican de la siguiente forma:

- **Métodos jerárquicos.**

  Este método agrupa los datos en forma de árboles y generalmete se clasifican en dos tipos de enfoque, uno aglomerativo y otro divisivo.

  * El *enfoque aglomerativo*:
    Utiliza una estrategia ascendente para agrupar los objetos; se fusionan los grupos más pequeños en grupos cada vez más grandes hasta que todos los objetos hayan sido agrupados en uno solo. Los métodos más utilizados son AGNES y DIANA.

  * El *enfoque divisivo*:
    Utiliza una estrategia descendente para agrupar los objetos, en este caso los grupos más grandes se dividen en grupos más pequeños hasta que cada objeto forme un grupo por si mismo. Los métodos más utilizados son: CURE, BIRC Y CHAMALEON [Chauhan *et al.*, 2010].

- **Métodos de particionamiento.**

  Los algoritmos de partición organizan los objetos en grupos tales que la dispersión total de cada objeto con respecto a su centro de grupo se minimice. Inicialmente, cada objeto se clasifica como un único grupo; en los siguientes pasos, todos los puntos de datos se reasignan iterativamente a cada grupo hasta que se cumpla un cierto criterio de parada. Bajo este criterio existen los siguientes métodos [Sumathi *et al.*, 2008]: algoritmo del vecino más cercano, algoritmo K-medioides, algoritmo K-medias.

- **Métodos con base en grillas.**

  Los algoritmos de agrupación con base en grillas primero cuantifican el espacio de agrupación en un número limitado de celdas y luego realizan las operaciones necesarias en la estructura de la grilla, las celdas que contienen más de un cierto número de puntos se tratan como densas, la principal ventaja de este enfoque es su rápido tiempo de procesamiento ya que el tiempo es independiente del número de objetos de datos pero depende del número de celdas. Los métodos más utilizados de este tipo son: STRING, Wave Grupo, Cliqué [Sumathi *et al.*, 2008].

- **Métodos con base en densidades.**

  El método considera los grupos como regiones densas de objetos separadas por regiones de baja densidad, que representan el ruido; a diferencia de los métodos de partición se pueden descubrir grupos de manera arbitraria. Los métodos con base en densidades pueden utilizarse para filtrar el ruido y los valores atípicos. Los métodos más utilizados de este tipo son: DBSCAN y OPTICS [Sumathi *et al.*, 2008].

2.3.3. Agrupación de datos funcionales con correlación espacial

Como se ha visto a lo largo de este capítulo, los métodos de agrupación se han ido adaptando a las necesidades para tratar los problemas del mundo real. Puesto que se ha ido generalizando esta técnica para trabajar con datos puntuales, luego datos espaciales y datos funcionales; ahora la nueva problemática es extender aun más estos métodos

de agrupación para datos funcionales con dependencia espacial. Este es el caso cuando las muestras son funciones observadas en diferentes sitios de una región o cuando estas funciones son observadas sobre un conjunto discreto de tiempo. Si los datos funcionales son espacialmente dependientes, los métodos básicos de agrupación de FDA fallan, ya que la estructura espacial es ignorada dando como resultado grupos en donde las curvas no son similares en forma o comportamiento. Es así que, además de considerar las características individuales en cada una de las curvas, se debe considerar la ubicación espacial de la curva y así agrupar aquellas curvas que son homogéneas no solo con respecto a su forma sino también con respecto a su ubicación espacial [Romano *et al.*, 2015].

Para agrupar datos funcionales con correlación espacial se han desarrollado métodos como el dinámico y jerárquico, así como varias formas de capturar la dependencia espacial. En [Giraldo *et al.*, 2012] se propuso el método jerárquico y en [Romano *et al.*, 2013] el método dinámico; ambos métodos utilizan una medida de asociación espacial para resaltar y distinguir la dependencia espacial entre las curvas; esta medida está dada por la función trazo-variograma. Ambos métodos tienen en cuenta la dinámica funcional de las curvas en términos de tiempo y sus relaciones espaciales. El método jerárquico toma en cuenta la dinámica funcional a través de la distancia entre las curvas mediante la norma $\mathcal{L}^2$, mientras que el método dinámico la considera a través de la distancia euclidiana al cuadrado entre las curvas calculando la función de variograma en cada grupo. No obstante, la dependencia espacial influye de manera diferente en el proceso de agrupación; en el caso jerárquico, la dependencia espacial es considerada por la función de trazo-variograma y en el caso dinámico, esta dependencia es considerada también por el trazo-variograma, pero calculado dentro de cada grupo [Romano *et al.*, 2015]. Por otro lado en [Romano *et al.*, 2017] presentan un método para este tipo de datos tomando en cuenta la contribución de cada curva a la variabilidad espacial; de esta forma se define una función de dispersión espacial asociada a cada curva y llevan a cabo la agrupación utilizando el algoritmo k-medias; este algoritmo se basa en la optimización de un criterio de ajuste entre las funciones de dispersión espacial asociadas a cada curva y el representante o centro de los grupos. En el trabajo realizado por [Romano *et al.*, 2011] se extiende la estrategia de agrupación con base en modelos para datos funcionales con correlación espacial; esta estrategia se enfoca en clasificar curvas espacialmente dependientes y obtener un modelo funcional espacial base para cada grupo; el ajuste de estos modelos implica estimar la función trazo-variograma, por lo que proponen un estimador de variograma de kernel.

Ahora bien, dado que el objetivo de este trabajo es adaptar el método k-medias para datos funcionales espacialmente correlacionados, se hará uso de la metodología planteada en [Giraldo *et al.*, 2012].

2.3.3.1. Método jerárquico para datos funcionales con correlación espacial.

Suponiendo que $X_1(t), ..., X_n(t)$ es una muestra de curvas definidas en $t \in T = [a, b] \subseteq \mathbb{R}$ y que además pertenecen al espacio de Hilbert de funciones cuadrado integrables definidas en $[a, b]$; es decir:

$$L_2(T) = \left\{ f : T \to \mathbb{R} : \int_T f(t)^2 < \infty \right\}.$$

De igual manera, se asume que las funciones son expandidas en términos de alguna base de funciones como sigue:

$$X_i(t) = \sum_{l=1}^{K} a_{il} B_l(t) = a_i^T B(t), \quad i = 1, ..., n$$

El análisis de agrupación jerárquico funcional es desarrollado como en el enfoque clásico, pero considerando la distancia entre las curvas $X_i(t)$ y $X_j(t)$ a través de la norma $\mathcal{L}^2$; es decir:

$$d_{ij} = \sqrt{\int_{[a,b]} (X_i(t) - X_j(t))^2 dt}$$

de donde, utilizando la representación de la curva en la base de funciones, se tiene:

$$d_{ij} = \sqrt{\int_{[a,b]} ((a_i - a_j)^T B(t) B(t)^T (a_i - a_j) dt}$$

$$d_{ij} = \sqrt{(a_i - a_j)^T W (a_i - a_j)}$$

donde:

$$W = \int_{[a,b]} B(t) B(t)^T dt$$

donde a_i y a_j son vectores de coeficientes de la base para las curvas i y j. Al utilizar bases ortonormales como la de Fourier, la matriz W es la identidad. Una vez calculada la matriz de disimilitud, el proceso estándar aglomerativo o divisivo es aplicado.

Cuando la estructura espacial es tomada en cuenta, y se considera el entorno geoestadístico, la agrupación permite encontrar grupos de sitios cercanos con características similares. Sea $\{Z(s) = (Z_1(s), ..., Z_m(s)) : s \in D\}$ un proceso espacial m multivariante definido sobre un dominio $D \subseteq \mathcal{R}^d$. Cuando $m = 1$, el proceso de agrupación pondera las disimilitudes d_{ij} entre curvas por:

$$d_{ij}^w = d_{ij} \gamma(h)$$

donde $\gamma(h)$ es el variograma calculado para las distancias entre las ubicaciones i, j. Por otro lado, si $m > 1$ la ponderación es llevada a cabo por:

$$d_{ij}^w = d_{ij} \Gamma(h)$$

donde $\Gamma(h)$ es el variograma multivariado definido por:

$$\Gamma(h) = \frac{1}{2} E[Z(x) - Z(x+h)]^T M [Z(x) - Z(x+h)]$$

con M una matriz simétrica definida positiva usada como métrica. En particular si $M = I$, el variograma multivariado está dado por:

$$\Gamma(h) = \sum_{l=1}^{m} \frac{1}{2} E[Z_l(x) - Z_l(x+h)]^2$$

$$\Gamma(h) = \sum_{l}^{m} \gamma_{ll}(h)$$

donde $\gamma_{ll}(h)$ es el variograma de la *l-ésima* variable. Una alternativa a la matriz M es la inversa de la matriz de varianzas-covarianzas. En este caso el variogrma multivariado es una suma ponderada de los variogramas y variogramas cruzados.

Ahora bien, en el contexto de datos funcionales con correlación espacial, se considera $\{\mathcal{Z}_s(t), s \in D \subseteq \mathbb{R}^d\}$ un proceso aleatorio funcional estacionario isotrópico y $\mathcal{Z}_1(t), ..., \mathcal{Z}_n(t)$ una realización de este proceso aleatorio observado en n sitios con coordenadas $s_1, ..., s_n$ respectivamente. Es así que, para realizar el análisis de agrupación la estructura espacial se la considera a través del trazo-variograma, definido por γ, de la siguiente manera:

$$\gamma(h) = \frac{1}{2} E\left[\int_{[a,b]} (X_i(t) - X_j(t))^2 dt\right], \quad h = \|x_i - x_j\|$$

y la ponderación de las distancias entre las curvas está dada por:

$$d_{ij}^w = d_{ij}\gamma(h)$$

Utilizando el método de momentos la estimación de $\gamma(h)$ es la siguiente:

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{s_i, s_j \in N(h)} \int_{[a,b]} (\mathcal{Z}_{s_i}(t) - \mathcal{Z}_{s_j}(t))^2 dt$$

donde $N(h) = \{(s_i, s_j) : \|s_i - s_j\| = h\}$ es el número de elementos distintos en $N(h)$. Luego de estimar el trazo-variograma para una secuencia de P puntos h_p, se ajusta un modelo paramétrico $\gamma(h;\theta)$ a los puntos $(h_p, \hat{\gamma}(h_p)),\ p = 1, ..., P$; este ajuste se lo hace usualmente mediante mínimos cuadrados ordinarios o ponderados. Este trazo-variograma, es un variograma válido debido a que tiene las mismas propiedades que las de un variograma paramétrico ajustado a partir de un conjunto de datos espaciales univariante.

Por otro lado, una segunda alternativa para la agrupación de este tipo de datos, es estimar los variogramas simples y variogramas cruzados a partir de los coeficientes de las funciones base obtenidos en la suavización.

Asumiendo que la curva para cada ubicación de muestra $i = 1, ..., n$ se la ha expandido utilizando los coeficientes de las funciones base, se tiene la siguiente matriz de coeficientes:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nK} \end{pmatrix}_{(nxK)}$$

que forman una realización de un campo aleatorio de K-variables $\{A(s) = (A_1((s), .., A_K(s)) : s \in D \subseteq \mathbb{R}^d\}$ con $E(A_i(t)) = v_i$, y la matriz de variogramas y variogramas cruzados es de la siguiente forma:

$$\Upsilon(h) = \begin{pmatrix} \gamma_{11}(h) & \gamma_{12}(h) & \cdots & \gamma_{1K}(h) \\ \gamma_{21}(h) & \gamma_{22}(h) & \cdots & \gamma_{2K}(h) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{K1}(h) & \gamma_{K2}(h) & \cdots & \gamma_{KK}(h) \end{pmatrix}_{(KxK)}$$

donde $\gamma_{lq} = \frac{1}{2}E[A_l(s_l) - A_q(s_j)^2],\ l, q = 1, ..., K,\ h = \|s_i - s_j\|$; para estimar la matriz anterior se hace uso del modelo lineal de coregionalización (LMC), este método permite modelizar los variogramas y variogramas cruzados de dos o más variables de modo que la varianza de cualquier combinación lineal posible de estas variables es siempre positiva. A partir del modelo estimado de LMC, se calcula el variograma multivariado utilizando los coeficientes de las funciones base.

$$\gamma_{lq}(h) = \sum_{k=1}^{K} b_{lq}^k g_k(h) \quad \forall l, q$$

donde los b_{lq}^k son las sillas o pendientes de los $g_k(h)$ y:

$$\frac{1}{2}E\left[A_{kl}(s) - A_{kl}(s+h)\right]\left[A_{k'l'}(s) - A_{k'l'}(s+h)\right] = \begin{cases} g_l(h) & \text{Si } k = k' \text{ ,} l = l' \\ 0 & \text{En otro caso} \end{cases}$$

y matricialmente se tiene que:

$$\Gamma(h) = \sum_{l=1}^{K} [b_{ij}^l]\gamma_{ll}(h)$$

y finalmente se tiene que:

$$d_{ij}^w = d_{ij}\Gamma(h)$$

2.4. Índices

2.4.1. Selección de número de grupos.

Se debe tener en cuenta que la agrupación producida por los algoritmos de clasificación o modificaciones de estos no es perfecta; puesto que no trabajan bajo la misma metodología, incluso los parámetros de entrada pueden ocasionar diferentes grupos o generar estructuras de grupos diferentes [Arbelaitz *et al.*, 2013]. Sin embargo, es fundamental conocer el número de grupos K, puesto que varios algoritmos de agrupación lo

requieren como parámetro de entrada. El número de grupos puede ser obtenido aplicando el algoritmo varias veces a partir de un posible valor $K_{mín}$ hasta un valor máximo $K_{máx}$ para luego analizar la secuencia de estructuras resultantes, las cuales serán evaluadas por medio de índices y la solución será aquel K con el mejor índice [Rui Xu, 2008]; considerando que mientras mayor sea el K más separados y más diferenciados serán los grupos[Kogan *et al.*, 2006].

Luego de haberse evaluado un índice en particular para un cierto rango de número de grupos, se tiene que decidir con que número de grupos trabajar. En el caso más simple, se puede considerar como una solución aquel número de grupo donde el índice toma el valor máximo o mínimo; no obstante, este criterio no se aplica en la mayoría de los casos. En la práctica, el número de grupos se determina de manera visual por medio del gráfico de la curva de valores del índice; usualmente se elige el valor donde la curva presenta una forma de codo; es decir, un salto positivo o negativo de la curva del índice o pico local [Dimitriadou *et al.*, 2002].

Las medidas de validación para los grupos formados se clasifican en validación externa y validación interna; la principal diferencia radica en el uso o no de información externa para la validación de los grupos. La validación externa usa información que no está presente en los datos para evaluar hasta que punto la estructura del grupo coincide con alguna estructura externa, ya que en este caso se conoce previamente el número de grupos; así, estas medidas pueden utilizarse para la elección de un algoritmo de agrupación óptimo en un conjunto de datos específico. Por otro lado, las medidas de validación interna evalúan la calidad de una estructura de agrupación sin tener información adicional; así, estas medidas pueden ser utilizadas para elegir el mejor algoritmo de agrupación así como el número óptimo de grupos. En la práctica, la información externa no siempre está disponible, por lo cual las medidas de validación interna son la única opción para la validación de los grupos [Charu C. Aggarwal, 2013].

Para cumplir con los objetivos planteados, en esta sección se trabajará únicamente con medidas de validación interna, las cuales están basadas en los criterios de cohesión y separación.

- **Criterio de Cohesión**

  Mide el grado de la relación entre los miembros de un grupo; es decir, se evalúa la compacidad de los grupo en función de la varianza, puesto que un valor menor de varianza indica que los grupos formados son más compactos [Liu *et al.*, 2013]. La medida básica de este tipo es la *Suma de cuadrados intra grupos (**SSW**)*; esta medida de cohesión se utiliza frecuentemente como función objetivo a minimizar en problemas de clasificación no supervisada, debido a que se relaciona la varianza con el número grupos. Gráficamente se puede elegir el número óptimo de grupos observando un codo en la curva de la función [Pérez, 2018]. Este índice está dado por:

$$SSW = \frac{1}{n}\sum_{j=1}^{K}\sum_{x\in C_j}\|c_j - x\|^2$$

  donde K es el número de grupos, c_j es el centroide del grupo j y n es el tamaño del conjunto de datos.

- **Criterio de Separación**

  Mide lo distinto o bien separado que está un grupo de los demás grupos. La *Suma cuadrática de las distancias entre grupos (**SSB**)* se define como el promedio de las distancias de los centroides de cada grupo al centroide del conjunto de datos; esta medida de separación usualmente se usa como función objetivo a maximizar y está dada por la siguiente expresión:

$$SSB = \frac{1}{n}\sum_{j=1}^{K} n_j \|c_j - \bar{x}\|^2$$

  donde K es el número de grupos, n_j es el número de miembros que contiene el grupo j, c_j es el centroide del grupo j, $\bar{x}$ es el centroide del conjunto de datos y n es el tamaño del conjunto de datos.

2.4.2. Índices con base en la suma de cuadrados

Los índices presentados a continuación están desarrollados con base en la suma de cuadrados; es decir, en las medidas presentadas previamente.

- **Ball y Hall**

$$\frac{SSW}{K}$$

  donde K es el número de grupos.

- **Calinski y Harabasz**

$$\frac{K-1}{n-K}\frac{SSB}{SSW}$$

  donde n es el número de datos y K el número de grupos.

- **Hartigan**

$$\log\left(\frac{SSB}{SSW}\right)$$

- **Xu**

$$d\log\left(\sqrt{\frac{SSW}{dn^2}}\right)$$

  donde n es el número de datos y d la dimensión de los datos.

2.4.2.1. Otros Índices

Existen índices de validación interna que no se basan en la suma de cuadrados y son los siguientes:

- Davies-Bouldin(DB)

$$DB = \frac{1}{K} \sum_{i=1, i \neq j} \text{máx} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

Donde K es el número de grupos, σ_i es la distancia promedio entre cada punto en el grupo i y el centroide del grupo, y $d(c_i, c_j)$ es la distancia entre los centroides de los grupos i y j.

- **Coeficiente de Silueta**

$$s(i) = \frac{b(i) - a(i)}{\text{máx}(a(i), b(i))}$$

Donde $a(i)$ es la media de las distancias del individuo i con los miembros del grupo; $b(i)$ es la distancia media entre el individuo i y todos los miembros del grupo más cercano al que no pertenece. El valor de $s(i)$ puede variar entre -1 y 1, donde -1 significa un mal agrupamiento, 0 es indiferente, y 1 buen agrupamiento.

El coeficiente de silueta para un grupo i:

$$SC_i = \frac{1}{n_i} \sum_{i=1}^{n_i} s(i)$$

donde n_i es el número de miembros en el grupo i.

- **Índice de Dunn**

$$Dunn = \min_{1 \leq j \leq K} \left\{ \min_{1 \leq t \leq K, t \neq j} \left(\frac{d(c_j, c_t)}{\text{máx}_{1 \leq k \leq K} d(c_k)} \right) \right\}$$

Donde K es el número de grupos, $d(c_j, c_t)$ es la distancia entre los centroides de los grupos j y t, $d(c_k)$ es el diámetro del grupo; es decir, la máxima distancia entre dos de sus puntos.

2.4.3. Validación de grupos

Una vez que los grupos han sido obtenidos a través del algoritmo de clasificación, es importante medir la robustez de los mismos; para esto se hace uso de índices que consideran la correlación temporal y espacial.

2.4.3.1. Correlación temporal

Como se mencionó un dato funcional también puede verse como una serie temporal, por lo que se puede utilizar una medida de correlación temporal. Es así que, el cálculo de esta correlación se convierte en otra forma de medir que tan bien que están formados los grupos. Esta correlación está dada por:

$$CORT(X_T, Y_T) = \frac{\sum_{t=1}^{T-1}(X_{t+1} - X_t)(Y_{t+1} - Y_t)}{\sqrt{\sum_{t=1}^{T-1}(X_{t+1} - X_t)^2}\sqrt{\sum_{t=1}^{T-1}(Y_{t+1} - Y_t)^2}}$$

donde $CORT(X_T, Y_T)$ recae en el intervalo $[-1, 1]$; el valor $CORT(X_T, Y_T) = 1$ significa que en cualquier período $[t_i, t_{i+1}]$, las series X_T y Y_t crecen o decrecen simultáneamente; es decir, que tienen el mismo comportamiento. El valor $CORT(X_T, Y_T) = -1$ significa que en cualquier período $[t_i, t_{i+1}]$, la serie X_T crece y Y_t decrece o viceversa; es decir, que tienen comportamiento opuesto. Finalmente, cuando el valor de $CORT(X_T, Y_T) = 0$ significa que no existe monotonicidad entre las series X_T y Y_T [Montero & Vilar, 2014].

2.4.3.2. Índices de correlación espacial

La teoría de la autocorrelación espacial ha sido un elemento clave en el análisis geográfico [Chen, 2013], puesto que como establece la primera ley de la Geografía "todas las cosas están relacionadas entre sí, pero las cosas más próximas en el espacio tienen una relación mayor que las distantes". El término "auto" indica que la correlación se realiza con la misma variable pero medida en distintos lugares del espacio [Ramirez, 2015].

Características socioeconómicas y ambientales propias de la geografía, tienden a mostrar cierto grado de similitud, ya que a menos que existan factores de ruptura o de discontinuidad muy marcados, se espera evidenciar cierta homogeneidad espacial [Celemin, 2009]. Así, este criterio permite entender la variación de un fenómeno en un área geográfica de análisis. Si el fenómeno analizado tiende a agruparse en regiones o forma grupos, entonces se evidencia autocorrelación positiva; por otro lado, si el fenómeno tiende a estar disperso, la autocorrelación espacial es negativa, y si el fenómeno muestra comportamiento aleatorio y no estructurado, se dice que no existe autocorrelación espacial; es decir, la presencia o ausencia de un atributo en un lugar determinado no influye en la medida de dicho atributo en puntos cercanos [Siabato & Guzmán-Manrique, 2019].

Por lo tanto, la autocorrelación espacial se interpreta entonces con un índice estadístico descriptivo que mide las formas y las maneras de como se distribuyen los fenómenos analizados en el área espacial [De Bellefon *et al.*, 2018].

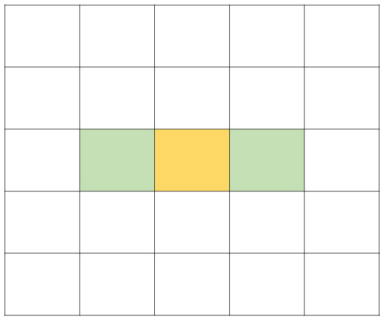
Para capturar esta correlación de tipo espacial, existen varios índices,de los cuales los índices de Moran y de Geary son los más utilizados; el primero es una generalización del coeficiente de correlación de Pearson, y el segundo es similar al estadístico de Durbin-Watson. A diferencia del coeficiente de Geary, el índice de Moran es más significativo para el análisis espacial [Chen, 2013].

2.4.3.2.1 Índice de Moran

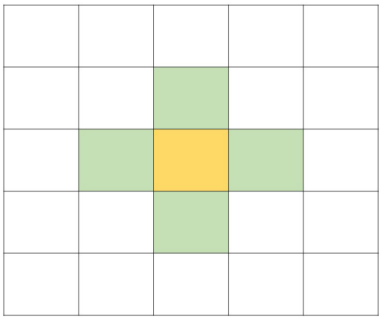
El índice de Moran solo tiene en cuenta los valores de las unidades de análisis determinadas a partir del criterio de vecindad; es decir, la contigüidad física y la distancia:

- **Contigüidad física**

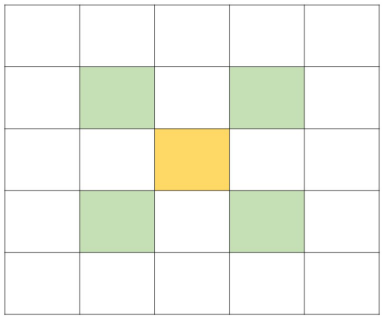
Este criterio toma en cuenta la contigüidad espacial de orden n para un área espacial dividido en unidades espaciales cuadradas y distribuidas uniformemente, ver Figura 2.25.



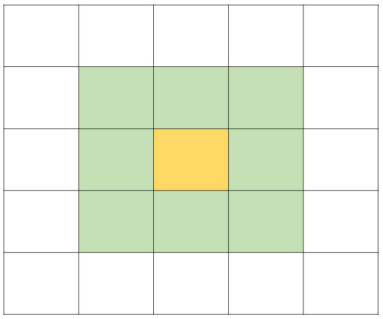
(a) Contigüidad tipo lineal.



(b) Contigüidad tipo lineal bidireccional.



(c) Contigüidad tipo alfil.

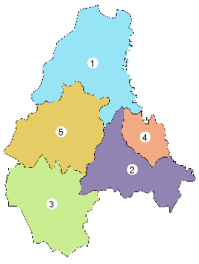


(d) Contigüidad tipo reina.

Figura 2.25: Tipos de contigüidad de primer orden.

- **Contigüidad por distancia**

  Este criterio establece el concepto de vecino a través de una distancia límite d previamente definida por el usuario; así, es habitual considerar d entre los centroides de las unidades espaciales, ver Figura 2.26 [Siabato & Guzmán-Manrique, 2019].



(a) Mapa.

	1	2	3	4	5
1	0	1	0	1	1
2	1	0	1	1	1
3	0	1	0	0	1
4	1	1	0	0	0
5	1	1	1	0	0

(b) Matriz de contigüidad.

Figura 2.26: Contigüidad por distancia.

Por lo consiguiente, en este índice la diagonal de la matriz de contigüidad no se

considera para el cálculo, puesto que se estable la restricción de que una unidad de análisis no es vecina de ella misma. Su naturaleza global se deriva de comparar de manera directa los valores de cada unidad de análisis con la media global del fenómeno, como se muestra en la siguiente ecuación:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n}(z_i - \bar{z})^2}, \quad i \neq j$$

$$I = \frac{\text{Covarianza ponderada}}{\text{Varianza de los valores}}$$

Donde n es el número de unidades de análisis y $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$ corresponde al número total de vecindades.

Este índice tiene como dominio el intervalo $[-1, +1]$, donde si $I < 0$ se evidencia correlación espacial negativa, si $I > 0$ se evidencia correlación espacial positiva y si $I = 0$ el fenómeno se distribuye aleatoriamente [Siabato & Guzmán-Manrique, 2019].

2.4.3.2.2 Índice de Geary

El índice de Geary permite evaluar la asociación espacial. El índice de Geary es similar al coeficiente de Moran; no obstante, el índice de Moran no puede ser sustituido por el coeficiente de Geary y viceversa, debido a que el índice de Moran se calcula a partir de una muestra, mientras que el índice de Geary se calcula a partir de la población [De Bellefon *et al.*, 2018]. A diferencia del índice de Moran, el coeficiente de Geary si incluye el valor de la unidad central en el análisis de autocorrelación. Este índice está definido en el intervalo $[0, 2]$. Cuando $c \in [0, 1]$, existe autocorrelación positiva, si $c \in (1, 2]$, indica autocorrelación negativa, y si $c = 1$, indica ausencia de autocorrelación y este índice está dado por la siguiente expresión [Siabato & Guzmán-Manrique, 2019].

$$C = \frac{n-1}{\sum_{i=1}^{n}(z_i - \bar{z})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i - z_j)^2}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}, \quad i \neq j$$

Nótese que el primer término es un factor de normalización y el numerador del segundo término establece las diferencias entre la unidad de análisis y sus vecinos. Por otro lado, este término no se compara con el valor medio global, sino con el valor medido de cada vecino; por consiguiente, cuanto más grande sea la diferencia entre la unidad de análisis central y sus vecinos, más grande será el numerador [Siabato & Guzmán-Manrique, 2019].

Capítulo 3

Metodología

En este capítulo se estudia el algoritmo K-medias y se desarrolla la metodología con la cual se va a trabajar.

3.1. Algoritmo K-medias

El algoritmo K-medias es un método de *particionamiento*, puesto que genera una partición sencilla de los datos en un intento de recuperar los grupos naturales presentes en los datos [Anil K. Jain, 1988]; este método utiliza la matriz de disimilitud entre los objetos como punto de partida [Bourgault *et al.*, 1992]. Los métodos de particionamiento se utilizan frecuentemente en aplicaciones de ingeniería donde las particiones simples son de interés, puesto que ofrecen una representación eficiente y compacta de grandes bases de datos [Anil K. Jain, 1988].

El problema de particionamiento de grupos formalmente establece que: dado n objetos en un espacio métrico $d-$dimensional, determinar una partición de patrones en K grupos, tal que los objetos en un grupo son más similares entre si que los objetos de los demás grupos [Anil K. Jain, 1988].

La solución a este problema es fácil de ver, puesto que se basa en elegir un criterio y evaluar cada una de las probables particiones que contienen k grupos y después seleccionar la partición que optimice dicho criterio; este razonamiento presenta dos problemas. El primero es encontrar un criterio que abarque la noción de grupo por medio de una expresión matemática que depende principalmente de los parámetros del problema, los cuales por motivos computacionales deben ser simples, pero suficientemente complejos para reflejar la estructura de los datos. La segunda dificultad es el gran número de particiones, incluso para un número de objetos moderado, por lo cual evaluar un criterio sencillo sobre todas las particiones no es factible [Anil K. Jain, 1988]. Además, este problema se ha catalogado como NP-duro, por lo que el uso de heurísticas son alternativas para encontrar una solución [Pérez-Ortega *et al.*, 2020].

Para evitar el problema combinatorio, una función de criterio es evaluada solo en conjuntos pequeños de particiones razonables. Para esto, el procedimiento más común es optimizar la función de criterio utilizando una técnica iterativa. Se empieza con una partición inicial, los objetos se mueven de un grupo a otro en un intento de mejorar el valor de la función de criterio; de esta forma, cada partición es una modificación del anterior; además, solo un número pequeño de particiones son examinados. Este procedimiento hace que los algoritmos sean computacionalmente más eficientes y usualmente convergen a un mínimo local de la función de criterio [Anil K. Jain, 1988].

Con base en el procedimiento anterior, el algoritmo K-medias empieza escogiendo K puntos representativos como los centroides iniciales. Un simple método de inicialización es el propuesto por MacQueen, en el que se toma K centroides aleatoriamente. Luego cada punto se asigna al centroide más cercano con base en una medida de proximidad escogida; usualmente estas medidas son la distancia de Manhattan (norma L_1), distancia Euclidiana (norma L_2), y la del coseno; en general, el algoritmo K-medias trabaja con la distancia Euclidiana. Una vez que los grupos están formados, los centroides de cada grupo se actualizan, luego el algoritmo repite estos dos procesos iterativamente hasta que los centroides no cambian o ningún otro criterio de convergencia se encuentra [Charu C. Aggarwal, 2013].

La función de criterio que utiliza este algoritmo es la *suma de errores cuadráticos* (SEC) o la *suma de los residuos cuadráticos* (SRC). De esta forma el problema de optimización se define de la siguiente manera: dado un conjunto de datos $D = \{z_1, ..., z_n\}$ de n puntos, sea $C = \{C_1, ..., C_K\}$ los grupos obtenidos luego de aplicar el algoritmo K-medias, entonces el SEC es:

$$SEC(C) = \sum_{k=1}^{K} \sum_{z_i \in C_k} \|c_k - z_i\|^2$$

$$c_k = \frac{\sum_{z_i \in C_k} z_i}{|C_k|}$$

donde, c_k es el centroide del grupo C_k. El objetivo es encontrar la agrupación que minimice el valor de SEC; para esto, la asignación iterativa y los pasos de actualización del algoritmo son los que logran minimizar el valor del SEC para un conjunto de centroides dado [Charu C. Aggarwal, 2013].

Por otro lado, la razón de elegir la media de los datos en un grupo como un prototipo de representación del grupo es la siguiente: sea C_k el $k-$ésimo grupo, z_i un punto en C_k, y c_k es la media del grupo k. De esta forma, resolviendo para el representante de C_j el cual minimiza el valor del SEC de la siguiente manera, se obtiene que:

$$SEC(C) = \sum_{k=1}^{K} \sum_{z_i \in C_k} (c_k - z_i)^2$$

$$\frac{\partial}{\partial c_j} SEC = \frac{\partial}{\partial c_j} \sum_{k=1}^{K} \sum_{z_i \in C_k} (c_k - z_i)^2$$

$$= \sum_{k=1}^{K} \sum_{z_i \in C_j} \frac{\partial}{\partial c_j} (c_j - z_i)^2$$

$$= \sum_{z_i \in C_j} 2(c_j - z_i) = 0$$

Por tanto:

$$\sum_{z_i \in C_j} 2(c_j - z_i) = 0$$

$$|C_j| c_j = \sum_{z_i \in C_j} z_i$$

$$c_j = \frac{\sum_{z_i \in C_j} z_i}{|C_j|}$$

Así, el mejor representante para minimizar el valor del SEC de un grupo es la media de los puntos en el grupo. En el algoritmo K-medias, el valor del SEC disminuye con cada iteración, por lo que este comportamiento monótonamente decreciente eventualmente converge a un mínimo local [Charu C. Aggarwal, 2013].

El algoritmo de K-medias referencial es el siguiente [Oyelade *et al.*, 2010]:

Algorithm 1 Algoritmo K-medias referencial.

Require: N:=$\{z_i, ..., z_n\}$. K: Número de grupos.

SEC=$1e^{23}$
Seleccionar K centroides iniciales de los grupos $\{c_j\}$
do
    OldSEC = SEC
    SEC1 = 0
    for j = 1 to K **do**
        $n_j = 0$
    end for
    for i = 1 to n **do**
        for j = 1 to K **do**
            Cálculo de la distancia $d^2(z_i, c_j)$
        end for
        Encontrar el centroide más cercano c_j a z_i
        $c_j = c_j + z_i$
        $n_j = n_j + 1$
        SEC1=SEC1 + $d^2(z_i, c_j)$
    end for
    for j = 1 to K **do**
        n_j = máx$(n_j, 1)$
        $c_j = c_j / n_j$
    end for
    SEC = SEC1
while (SEC < OldSEC)

3.2. Algortimo K-medias funcional espacial

Ahora bien, con base en la metodología desarrollada en el método de clasificación jerárquico presentado en la subsección 2.3.3 para datos funcionales con correlación es-

pacial, y dado que los datos funcionales pertenecen al espacio de Hilbert de funciones cuadrado-integrables, se desarrolla el algoritmo considerando la distancia entre curvas a través de la norma $\mathcal{L}^2$. Esta distancia se obtiene utilizando una representación reducida de los datos mediante una base finita de tipo Fourier. Luego, se calcula el variograma empírico y se ajusta a un modelo teórico para así ponderar la matriz de distancia entre las curvas por el trazo-variograma y variograma multivariado calculado con los coeficientes de las funciones base; a partir de esta matriz se lleva acabo la agrupación de datos funcionales correlacionados espacialmente (Ver Algoritmo 2). La implementación de este algoritmo se encuentra en el apéndice B.

Algorithm 2 Algoritmo K-medias para datos funcionales con correlación espacial.

Require: N:=$\{z_i(t), ..., z_n(t)\}$. D:=$\{s_i, ..., s_n\}$ conjunto de ubicaciones. K: Número de grupos.

SEC=$1e^{23}$
Cálculo de trazo-variograma(γ) o variograma multivariado(Γ).
Seleccionar K centroides iniciales de los grupos $\{c_j(t)\}$
do
    OldSEC = SEC
    SEC1 = 0
    for j = 1 to K **do**
        $n_j = 0$
    end for
    for i = 1 to n **do**
        for j = 1 to K **do**
            Cálculo de la distancia $d^{*2}(z_i(t), c_j(t))$
        end for
        Encontrar el centroide más cercano $c_j(t)$ a $z_i(t)$
        $c_j(t) = c_j(t) + z_i(t)$
        $n_j = n_j + 1$
        SEC1=SEC1 + $d^{*2}(z_i(t), c_j(t))$
    end for
    for j = 1 to K **do**
        $n_j = \text{máx}(n_j, 1)$
        $c_j(t) = c_j(t)/n_j$
    end for
    SEC = SEC1
while (SEC < OldSEC)

$d^*(z_i(t), c_j(t)) = d(z_i(t), c_j(t))\gamma_{ij}(h)$
d_{ir}: distancia entre las curvas $z_i(t)$ y $z_r(t)$.

3.3. Índices de calidad: caso funcional

A continuación se presentan los distintos índices y criterios con un enfoque funcional con base en la metodología presentada en la sección 2.4.

3.3.1. Selección de número de grupos.

Tomando como punto de partida lo expuesto en la sección de índices con base en la suma de cuadrados 2.4.2, se extienden estos conceptos bajo un enfoque funcional, considerando $(\mathcal{Z}_{s_1}(t), ..., \mathcal{Z}_{s_n}(t))$ para $t \in T \subseteq \mathbb{R}$ y $s_i \in D$ para $i = 1, ..., n$ un conjunto de observaciones funcionales georeferenciadas y se asume que cada función pertenece al espacio de Hilbert de funciones cuadrado integrables.

- **Criterio de Cohesión**

  La medida básica de este tipo es la *Suma de cuadrados dentro de los grupos (**SSW**)*, este índice está dado por:

$$SSW = \frac{1}{n} \sum_{j=1}^{K} \sum_{z_s(t) \in C_j} \int_T (c_j(t) - z_s(t))^2 dt$$

  donde K es el número de grupos, c_j es el centroide del grupo j y n es el tamaño del conjunto de datos.

- **Criterio de Separación**

  La *Suma cuadrática de las distancias entre grupos (**SSB**)* está dada por la siguiente expresión:

$$SSB = \frac{1}{n} \sum_{j=1}^{K} n_j \int_T (c_j(t) - \bar{z}(t))^2 dt$$

  donde K es el número de grupos, n_j es el número de miembros que contiene el grupo j, c_j es el centroide del grupo j, $\bar{z}(t)$ es la media funcional del conjunto de datos y n es el tamaño del conjunto de datos.

3.3.1.1. Otros Índices

Existen índices de validación interna que no se basan en la suma de cuadrados y son los siguientes:

- **Davies-Bouldin(DB)**

$$DB = \frac{1}{K} \sum_{i=1, i \neq j} \text{máx} \left(\frac{\sigma_i + \sigma_j}{d(c_i(t), c_j(t))} \right), \quad d(c_i(t), c_j(t)) = \left(\int_T (c_i(t) - c_j(t))^2 dt \right)^{1/2}$$

  Donde K es el número de grupos, σ_i es la distancia promedio entre cada punto en el grupo i y el centroide del grupo, y $c_i(t)$ es el centroide del grupo i.

- **Coeficiente de Silueta**

$$s(i) = \frac{b(i) - a(i)}{\text{máx}(a(i), b(i))}$$

Este coeficiente tiene la misma expresión dada en la parte multivariante, salvo que las distancias utilizadas para su cálculo se realizan a través de la norma $\mathcal{L}^2$. Así, $a(i)$ es la media de las distancias del individuo i con los miembros del grupo; $b(i)$ es la distancia media entre el individuo i y todos los miembros del grupo más cercano al que no pertenece.

- **Índice de Dunn**

  El índice de Dunn en su versión funcional está dado por la siguiente expresión:

$$Dunn = \min_{1\leq j\leq K}\left\{\min_{1\leq l\leq K, l\neq j}\left(\frac{d(c_j(t), c_l(t))}{\max_{1\leq k\leq K} d(c_k(t))}\right)\right\}$$

$$d(c_j(t), c_l(t)) = \left(\int_T (c_j(t) - c_l(t))^2 dt\right)^{1/2}$$

  donde K es el número de grupos, $c_j(t)$ es el centroide del grupo j y $d(c_k(t))$ es el diámetro del grupo; es decir, la máxima distancia entre dos de sus datos funcionales.

3.3.2. Validación de grupos

Para validar los grupos formados se hace uso de índices de correlación espacial de Moran y de Geary con su extensión al campo funcional. Considerando $(\mathcal{Z}_{s_1}(t), ..., \mathcal{Z}_{s_n}(t))$ para $t \in T \subseteq \mathbb{R}$, $s_i \in D$ para $i = 1, ..., n$ un conjunto de observaciones funcionales georeferenciadas y se asume que cada función pertenece al espacio de Hilbert de funciones cuadrado integrables.

3.3.2.1. Índice de Moran

El índice de Moran en el caso funcional está dado por la siguiente expresión:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \int_{t\in T} (z_{s_i}(t) - \bar{z}(t))(z_{s_j}(t) - \bar{z}(t))dt}{\sum_{i=1}^{n} \int_{t\in T} (z_{s_i}(t) - \bar{z}(t))^2 dt}, \quad i \neq j$$

donde w_{ij} define la relación entre las ubicaciones geográficas en una región. Este índice servirá para analizar la correlación espacial entre los grupos y dentro de ellos. Para el primer caso, se considera $z_{s_i}(t)$ como el centroide del grupo i y $\bar{z}(t)$ como la media funcional de todo el conjunto de datos y n corresponde al número de grupos; mientras que en el segundo caso, para algún grupo k, se considera $z_{s_i}(t)$ como los datos funcionales que pertenecen a este grupo, $\bar{z}(t)$ como el centroide de este grupo y n corresponde a los datos funcionales que contiene el grupo k. Se debe considerar que dependiendo del cálculo que se quiera realizar con el índice, w_{ij} estará definida de manera diferente.

3.3.2.2. Índice de Geary

El índice de Geary en el caso funcional está dado por la siguiente expresión:

$$C = \frac{n-1}{\sum_{i=1}^{n}\int_T (z_{s_i}(t) - \bar{z}(t))^2 dt} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\int_T (z_{s_i}(t) - z_{s_j}(t))^2 dt}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}, \quad i \neq j$$

donde w_{ij} define la relación entre las ubicaciones geográficas en una región. Este índice, al igual que en el caso anterior, servirá para analizar la correlación entre los grupos y dentro de ellos. Para el primer caso, se considera $z_{s_i}(t)$ como el centroide del grupo i, $\bar{z}(t)$ como la media funcional de todo el conjunto de datos y n corresponde al número de grupos; mientras que en el segundo caso, para algún grupo k, se considera $z_{s_i}(t)$ como los datos funcionales que pertenecen a este grupo, $\bar{z}(t)$ como el centroide de este grupo y n corresponde al número de datos funcionales que pertenecen a este grupo. Se debe considerar que dependiendo del cálculo que se quiera realizar con el índice, w_{ij} estará definida de manera diferente.

Capítulo 4

Validación y Aplicación

4.1. Estudio de Simulación

Con base en la metodología presentada en la subsección 2.3.3 y una vez realizada la implementación en el software R del algoritmo desarrollado en el sección 3.2, se procede a realizar un estudio de simulación con el objetivo de validar el mismo.

Para poner a prueba el comportamiento del algoritmo bajo escenarios prácticos, es necesario evaluar los procedimientos en datos de los cuales se tiene la respuesta. De esta manera se consideraron varias medias funcionales definidas sobre ubicaciones espaciales particulares y modelos de covarianza espacial. Por consiguiente, se generaron aleatoriamente 30 puntos fijos en cada cuadrante como se ve en la figura 4.1, para un total de 120 puntos y se simularon conjuntos de datos funcionales con correlación espacial bajo el siguiente modelo:

$$Z_{ij} = \mu_i(t) + \epsilon_j(t), \quad j = 1, \ldots, 120, \quad i = 1, 2. \tag{4.1}$$

Donde $\mu_i(t)$ denota la media funcional constante, y $\epsilon(t) \sim \mathcal{N}_{120}(0, \Sigma)$ para cada t con $t = 1, 2, \ldots, 365$. Cada Z_{ij} representa una curva medida en la ubicación j y con media constante μ_i. Se asume que las curvas son mediciones tomadas por día a lo largo de un año. Este estudio se realizó en este sentido para su aplicación a los datos del NDVI de los páramos del Ecuador en la siguiente sección. La estructura espacial de los datos fue dada en términos de la matriz Σ. Para incrementar el número de estructuras y escenarios posibles, se consideraron cuatro modelos de covarianza espacial: *Gauassiano*, *Esférico*, *Matérn* ($\kappa = 0{,}2$) y *Matérn* ($\kappa = 2$). Adicionalmente, se consideraron dos niveles de variabilidad, $\sigma^2 = 2$ y $\sigma^2 = 4$, y dos niveles de correlación espacial, $\phi = 1{,}8$ y $\phi = 2{,}5$. El comportamiento de estos modelos se muestra en la figura 4.2. Tres escenarios fueron considerados para $\mu_1(t)$ y $\mu_2(t)$; en todos los casos, las ubicaciones en los cuadrantes I y III tienen media funcional $\mu_1(t)$, mientras que las ubicaciones en los cuadrantes II y IV tienen media funcional $\mu_2(t)$. En este sentido, las curvas ubicadas en los cuadrantes I y III deberían tener un comportamiento similar al igual que las curvas que están ubicadas en los cuadrantes II y IV; esto debe ser detectado por el algoritmo. Así, se define una variedad de escenarios a través de las medias funcionales. El primer escenario considera las medias $\mu_1(t) = 5$ y $\mu_2(t) = 6$. El segundo escenario las medias $\mu_1(t) = 5$ y $\mu_2(t) = 10$ y el escenario final considera las medias $\mu_1(t) = 5$ y $\mu_2(t) = 15$. Cada escenario descrito previamente considera todas las combinaciones de los parámetros expuestos. Dos simulaciones se muestran en la figura 4.3, donde se puede apreciar las similaridades

y disimilitudes de las curvas en función de su media funcional. Para cada conjunto de datos funcionales simulados, las observaciones fueron suavizadas utilizando una base de tipo Fourier de 65 funciones base. Para cada escenario de simulación, se aplicaron los algoritmos: k-medias funcional con ponderación mediante el trazo variograma y k-medias funcional con ponderación mediante el variograma multivariado.

Si se toma en cuenta la correlación espacial, se aplica la metodología expuesta y el algoritmo k-medias funcional mediante la ponderación de la matriz de distancias por la función de correlación espacial, el algoritmo debería clasificar las ubicaciones en cuatro grupos; es decir, que cada grupo debe estar formado por sitios dentro del mismo cuadrante con medias funcionales iguales y espacialmente cercanos, a pesar de que las medias funcionales indiquen que solo existen dos grupos.

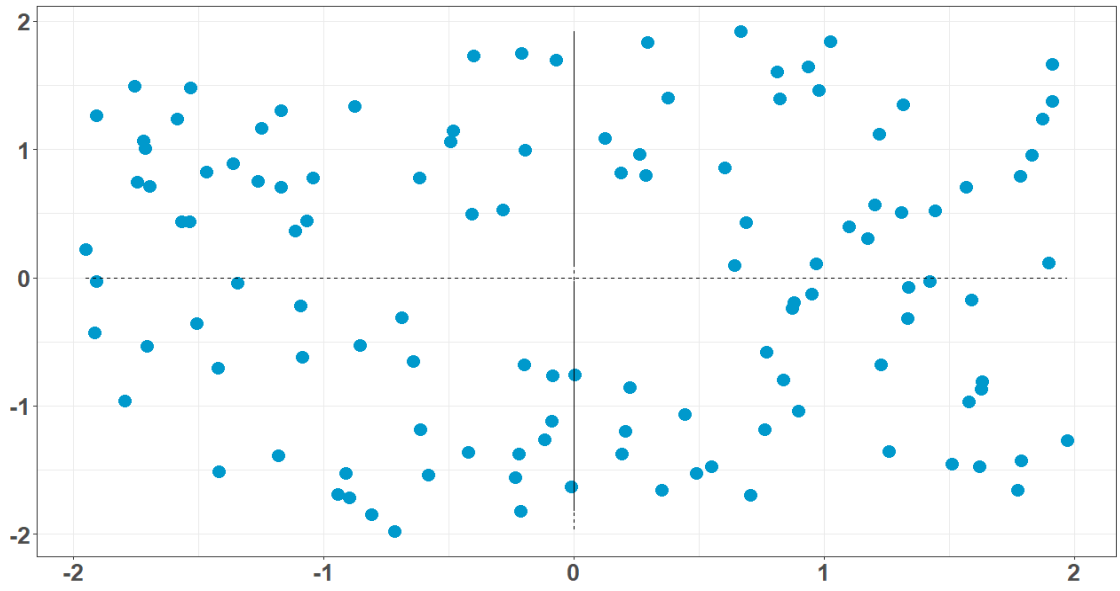


Figura 4.1: Distribución de puntos en el plano.

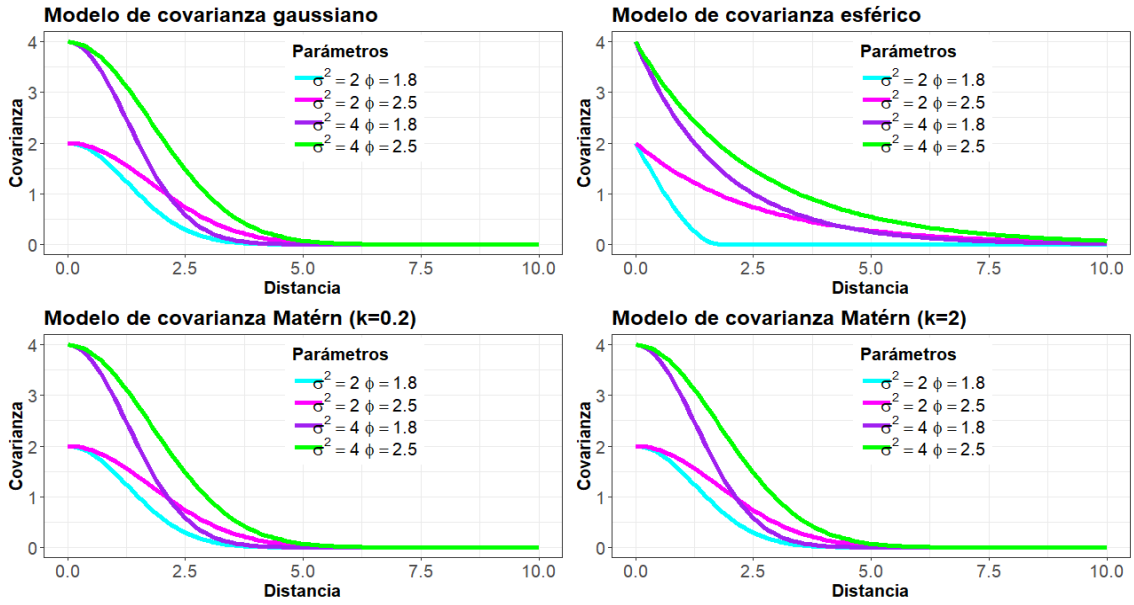
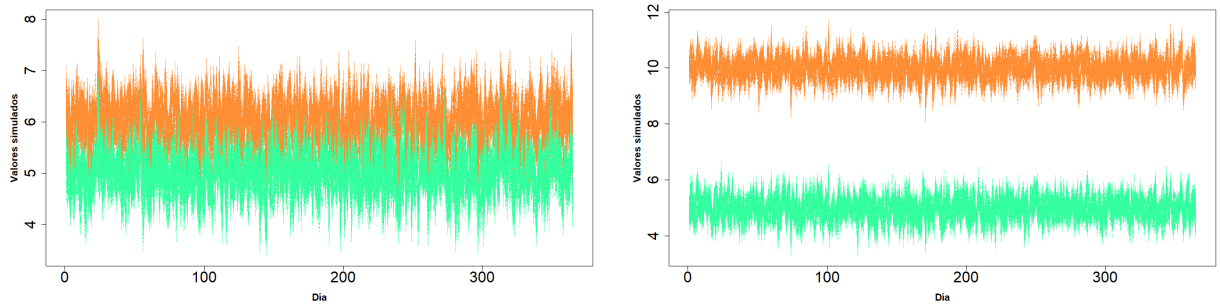


Figura 4.2: Modelos de covarianza.



(a) Medias funcional $u_1(t) = 5,\ u_2(t) = 6$. (b) Medias funcional $u_1(t) = 5,\ u_2(t) = 10$.

Figura 4.3: Curvas simuladas.

En la figura 4.3, se observa la simulación de 120 datos funcionales; en ambos casos el modelo de correlación espacial utilizado fue el modelo gaussiano con parámetros $\sigma^2 = 0{,}2$ y $\phi = 1{,}5$.

4.1.1. Resultados método k-medias funcional espacial

Para cada escenario planteado previamente 1.000 simulaciones fueron realizadas y se aplicaron las dos variantes del algoritmo k-medias en cada caso, **WTV:** ponderación de la matriz de distancias mediante el trazo-variograma y **WMV:** ponderación de la matriz de distancias mediante el variograma multivariado. La clasificación de estos métodos se puede visualizar en las figuras 4.4 y 4.5 respectivamente.

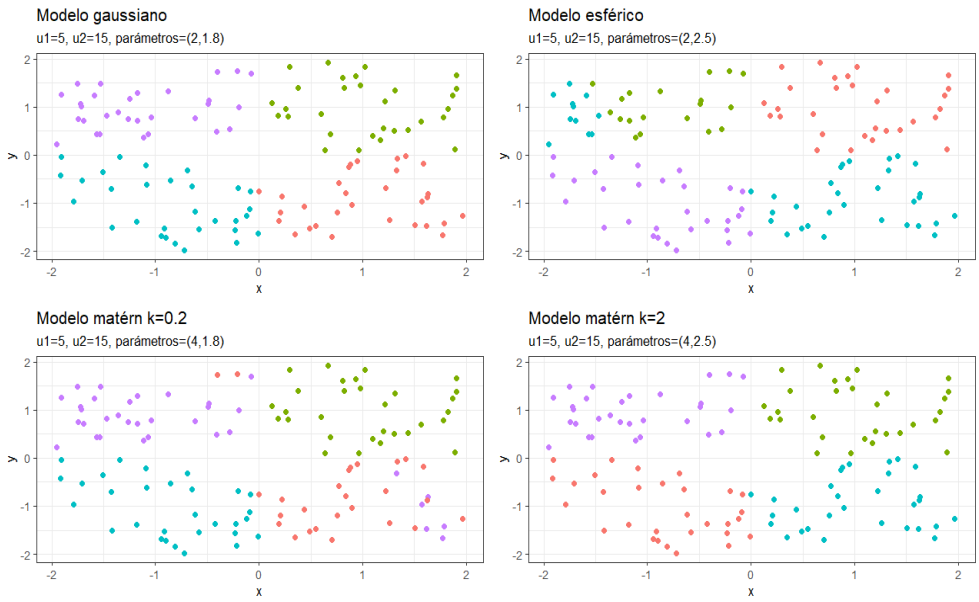


Figura 4.4: Clasificación WTV: media funcional $\mu_1(t) = 5, \mu_2(t) = 15$.

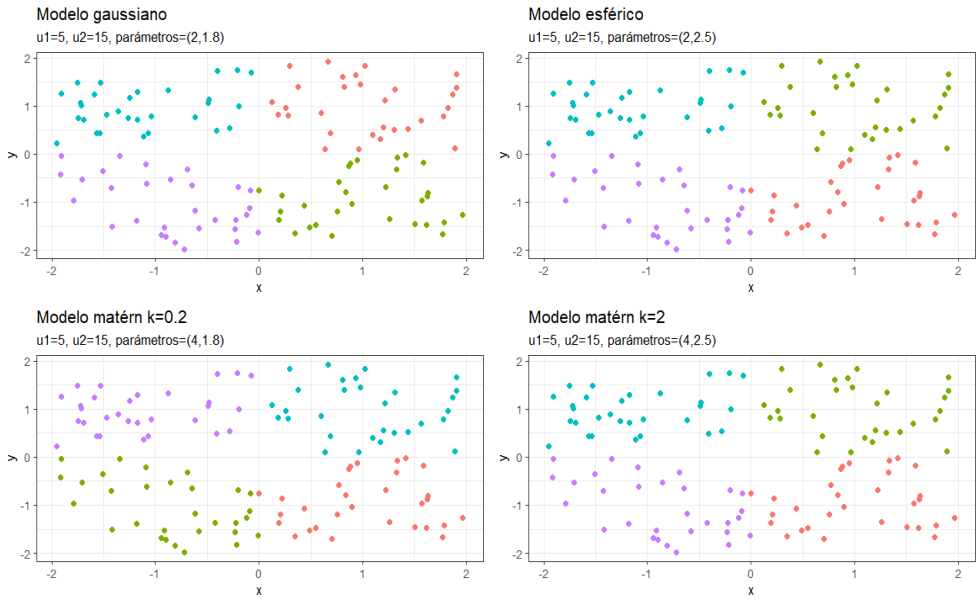


Figura 4.5: Clasificación WMV: media funcional $\mu_1(t) = 5, \mu_2(t) = 15$.

En las siguientes tablas se presentan los valores de porcentaje de correcta clasificación (PCC). Se puede observar que cada método clasifica las ubicaciones en cuatro grupos. Así, el PCC muestra la estructura de dependencia espacial que existe entre las ubicaciones.

- **Escenario 1:** $\sigma^2 = 2$, $\phi = 1{,}8$

Modelo	Método	$\mu_1(t) = 5$ $\mu_2(t) = 6$	$\mu_1(t) = 5$ $\mu_2(t) = 10$	$\mu_1(t) = 5$ $\mu_2(t) = 15$
Gaussiano	WTV	93,76	96,65	96,85
	WMV	93,02	96,29	96,45
Esférico	WTV	90,44	93,14	93,01
	WMV	88,12	93,13	92,22
Matérn	WTV	90,47	92,37	89,49
$\kappa = 0,2$	WMV	92,50	91,46	89,28
Matérn	WTV	98,63	98,81	98,10
$\kappa = 2$	WMV	98,63	98,43	97,45

Cuadro 4.1: PCC: Escenario 1.

- **Escenario 2:** $\sigma^2 = 2$, $\phi = 2{,}5$

Modelo	Método	$\mu_1(t) = 5$ $\mu_2(t) = 6$	$\mu_1(t) = 5$ $\mu_2(t) = 10$	$\mu_1(t) = 5$ $\mu_2(t) = 15$
Gaussiano	WTV	97,37	99,29	98,37
	WMV	97,05	98,69	98,64
Esférico	WTV	92,01	95,08	94,49
	WMV	91,99	95,28	94,78
Matérn	WTV	90,66	92,99	89,92
$\kappa = 0,2$	WMV	93,41	92,83	90,86
Matérn	WTV	99,29	99,24	97,75
$\kappa = 2$	WMV	99,06	99,03	97,36

Cuadro 4.2: PCC: Escenario 2.

- Escenario 3: $\sigma^2 = 4$, $\phi = 1{,}8$

Modelo	Método	$\mu_1(t) = 5$ $\mu_2(t) = 6$	$\mu_1(t) = 5$ $\mu_2(t) = 10$	$\mu_1(t) = 5$ $\mu_2(t) = 15$
Gaussiano	WTV	92,60	96,45	97,13
	WMV	92,07	96,27	96,43
Esférico	WTV	86,93	94,32	93,05
	WMV	84,97	93,75	92,66
Matérn	WTV	87,93	93,04	90,86
$\kappa = 0,2$	WMV	91,20	91,41	91,23
Matérn	WTV	98,04	98,95	98,52
$\kappa = 2$	WMV	98,74	98,51	98,01

Cuadro 4.3: PCC: Escenario 3.

- **Escenario 4:** $\sigma^2 = 4$**,** $\phi = 2{,}5$

Modelo	Método	$\mu_1(t) = 5$ $\mu_2(t) = 6$	$\mu_1(t) = 5$ $\mu_2(t) = 10$	$\mu_1(t) = 5$ $\mu_2(t) = 15$
Gaussiano	WTV	95,94	98,25	99,03
	WMV	96,14	98,54	98,62
Esférico	WTV	90,83	96,35	96,28
	WMV	90,12	95,72	94,88
Matérn	WTV	88,83	93,56	91,56
$\kappa = 0,2$	WMV	91,70	93,87	91,31
Matérn	WTV	99,39	99,40	98,73
$\kappa = 2$	WMV	98,99	98,43	98,28

Cuadro 4.4: PCC: Escenario 4.

De los resultados mostrados en las tablas 4.1-4.4, se observa que el PCC aumenta ligeramente cuando la distancia entre las medias funcionales se incrementa, tanto para el método **WTV** como para el método **WMV**; mostrando resultados en general superiores al 84 % y obteniéndose los valores de PCC más altos en el método Gaussiano. Por otro lado, a pesar de que el método **WMV** realice el proceso de agrupación utilizando más

información de los datos, los valores de PCC no presentan una diferencia notable con respecto al método **WTV**, únicamente logrando diferenciarse por décimas, salvo en el modelo *Matérn* $\kappa = 0{,}2$ del escenario 2 (tabla 4.2) donde se aprecia una diferencia del $2,75\,\%$ para las medias $\mu_1(t) = 5$ y $\mu_2(t) = 6$. Con respecto al modelo *Matérn* se observa que en todos los escenarios, al utilizar el método **WTV**, los valores de PCC crecen al considerar las medias funcionales $\mu_1(t) = 5$ y $\mu_2(t) = 6$ a $\mu_1(t) = 5$ y $\mu_2(t) = 10$ y decrecen en las medias funcionales $\mu_1(t) = 5$ y $\mu_2(t) = 10$ a $\mu_1(t) = 5$ y $\mu_2(t) = 15$. Por otro lado, en el método **WMV** se tiene un comportamiento decreciente para todo par de medias funcionales; no obstante, la diferencia entre los valores de PCC no es significativa.

En el apéndice A se encuentra de manera complementaria un caso adicional en donde se simularon datos bajo el modelo 4.1 y se realizó la clasificación en ocho grupos; y otro escenario en el cual se incrementó el valor de variabilidad.

4.1.1.1. Ruido blanco

Para resaltar la importancia de considerar la dependencia espacial en el proceso de clasificación, se presentan los siguientes escenarios.

4.1.1.2. Clasificación considerando coordenadas

En este caso, se simularon datos sin dependencia espacial; es decir, en el modelo 4.1 se tiene que $\epsilon(t) \sim \mathcal{N}_{120}(0, \sigma^2 I)$; sin embargo, para la clasificación se tomó en cuenta la ubicación espacial de los puntos, obteniendo los siguientes resultados:

Modelo	Método	$\mu_1(t) = 5$ $\mu_2(t) = 5$	$\mu_1(t) = 5$ $\mu_2(t) = 6$	$\mu_1(t) = 5$ $\mu_2(t) = 10$	$\mu_1(t) = 5$ $\mu_2(t) = 15$
Ruido	WTV	40,48	78,06	77,72	76,85
blanco	WMV	69,00	85,29	77,90	76,94

Cuadro 4.5: PCC: Datos simulados sin correlación espacial.

Se puede observar que pese a que no existe correlación espacial entre las ubicaciones los resultados obtenidos son relativamente buenos. Es así que a medida que las medias se alejan el método tiende a clasificar erróneamente; esto se debe a que no existe una estructura espacial definida en los datos que ayude a distinguir los grupos cuando estos se ubican en cuadrantes con la misma media funcional. Para el caso de medias funcionales iguales, el PCC es cercano al $41\,\%$ y $70\,\%$ con los métodos WTV y WMV, respectivamente, puesto que las formas de las curvas son similares y las coordenadas de las ubicaciones no tienen asignadas una forma funcional específica.

4.1.1.3. Clasificación sin considerar coordenadas

Al igual que en el caso anterior, en el modelo 4.1 se tiene que $\epsilon(t) \sim \mathcal{N}_{120}(0, \sigma^2 I)$. En este caso no se consideraron las ubicaciones espaciales para su clasificación; es decir, únicamente se aplicó el algoritmo k-medias funcional obteniendo los siguientes resultados:

Modelo	$\mu_1(t) = 5$ $\mu_2(t) = 5$	$\mu_1(t) = 5$ $\mu_2(t) = 6$	$\mu_1(t) = 5$ $\mu_2(t) = 10$	$\mu_1(t) = 5$ $\mu_2(t) = 15$
Ruido blanco	32,83	54,63	54,32	54,52

Cuadro 4.6: PCC: Clasificación K-medias funcional base.

En este caso, se observa que los valores de PCC no sobrepasan el 55 %. El algoritmo base de k-medias funcional al basarse más en la forma de la curva en el proceso de agrupación, no es capaz de diferenciar que tipo curvas están originadas en las ubicaciones espaciales.

En resumen, los resultados de las simulaciones indican que si existe una estructura de dependencia espacial entre las curvas es importante considerar esta información adicional en los procesos de agrupación. El algoritmo k-medias modificado muestra un óptimo comportamiento en los distintos escenarios, en donde se obtuvieron buenos valores de PCC, en particular cuando la distancia entre las medias funcionales se incrementa. Por otro lado, al comparar los valores de PCC obtenidos entre los algoritmos k-medias modificado y k-medias funcional base se observó que el primero generó valores superiores al trabajar con este tipo de datos.

4.1.2. Resultados método jerárquico funcional espacial

De manera similar al caso del método k-medias funcional espacial, y para cumplir otro de los objetivos planteados, 1.000 simulaciones fueron realizadas para cada escenario mencionado previamente y las dos variantes del algoritmo jerárquico funcional espacial mencionados en [Giraldo *et al.*, 2012]. Para cada caso se aplicó: **WTV:** ponderación de la matriz de distancias por medio del trazo-variograma y **WMV:** ponderación de la matriz de distancias por medio del variograma multivariado. La clasificación de estos métodos se puede visualizar en las figuras 4.6 y 4.7 respectivamente.

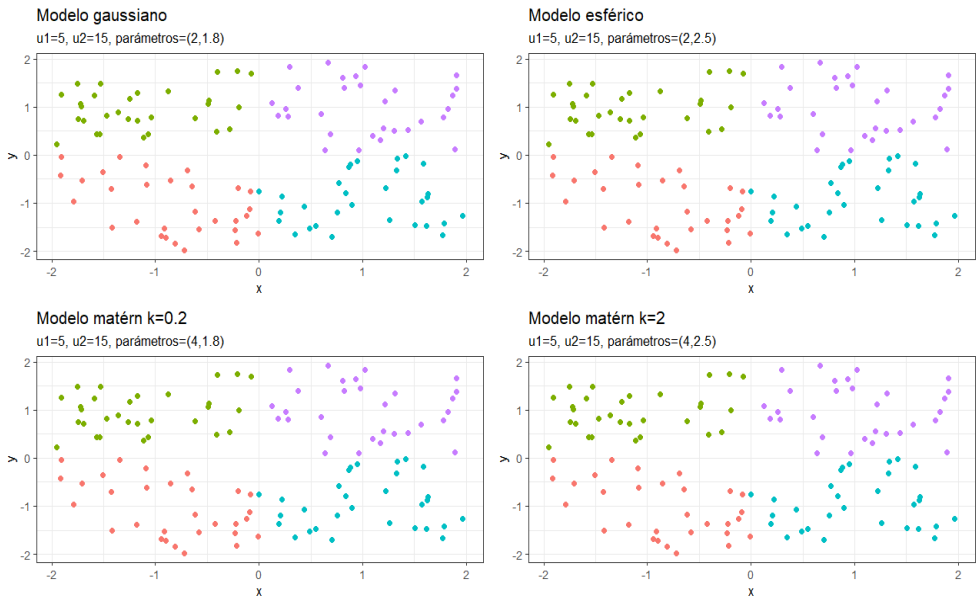


Figura 4.6: Clasificación WTV: media funcional $\mu_1(t) = 5, \mu_2(t) = 15$.

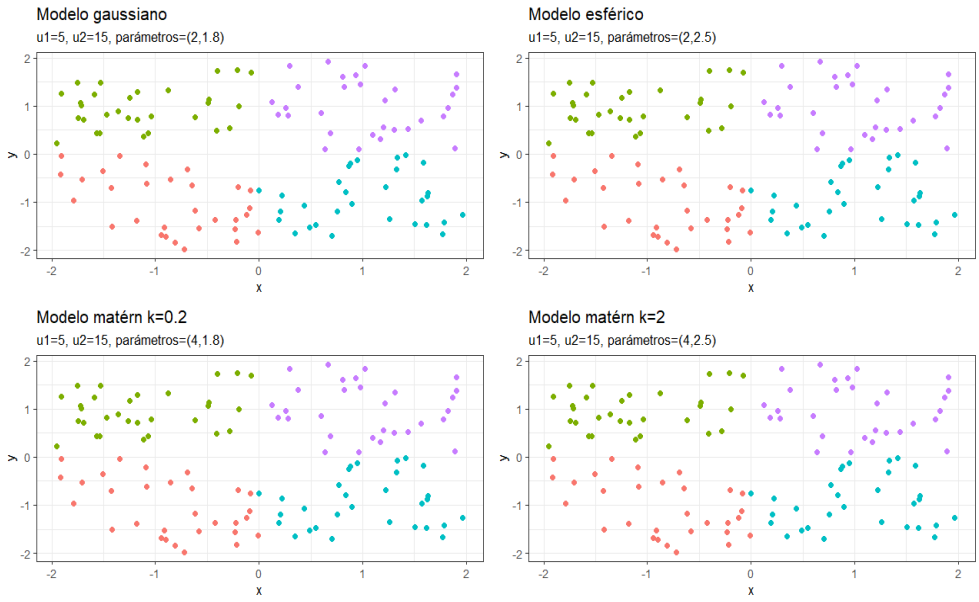


Figura 4.7: Clasificación WMV: media funcional $\mu_1(t) = 5, \mu_2(t) = 15$.

En las siguientes tablas se presentan los valores de PCC, en las cuales cada método clasifica las ubicaciones en cuatro grupos. Así, el PCC muestra la estructura de dependencia espacial existente entre las ubicaciones.

- **Escenario 1:** $\sigma^2 = 2$, $\phi = 1{,}8$

Modelo	Método	$\mu_1(t) = 5$ $\mu_2(t) = 6$	$\mu_1(t) = 5$ $\mu_2(t) = 10$	$\mu_1(t) = 5$ $\mu_2(t) = 15$
Gaussiano	WTV	93,94	98,34	98,24
	WMV	95,23	98,59	98,72
Esférico	WTV	84,29	98,05	98,39
	WMV	89,29	96,95	97,02
Matérn	WTV	97,54	100	100
$\kappa = 0,2$	WMV	97,97	100	100
Matérn	WTV	98,10	100	100
$\kappa = 2$	WMV	97,49	100	100

Cuadro 4.7: PCC: Escenario 1.

- **Escenario 2:** $\sigma^2 = 2$, $\phi = 2{,}5$

Modelo	Método	$\mu_1(t) = 5$ $\mu_2(t) = 6$	$\mu_1(t) = 5$ $\mu_2(t) = 10$	$\mu_1(t) = 5$ $\mu_2(t) = 15$
Gaussiano	WTV	96,39	99,93	99,95
	WMV	98,25	99,94	99,93
Esférico	WTV	90,94	97,55	97,57
	WMV	92,55	96,56	96,31
Matérn	WTV	98,72	100	100
$\kappa = 0,2$	WMV	98,60	100	100
Matérn	WTV	98,65	100	100
$\kappa = 2$	WMV	98,79	100	100

Cuadro 4.8: PCC: Escenario 2.

- Escenario 3: $\sigma^2 = 4$, $\phi = 1{,}8$

Modelo	Método	$\mu_1(t) = 5$ $\mu_2(t) = 6$	$\mu_1(t) = 5$ $\mu_2(t) = 10$	$\mu_1(t) = 5$ $\mu_2(t) = 15$
Gaussiano	WTV	92,19	98,37	98,36
	WMV	94,13	98,55	98,16
Esférico	WTV	77,31	97,23	97,98
	WMV	84,62	96,69	97,11
Matérn	WTV	97,54	100	100
$\kappa = 0,2$	WMV	98,45	100	100
Matérn	WTV	98,13	100	100
$\kappa = 2$	WMV	98,28	100	100

Cuadro 4.9: PCC: Escenario 3.

- **Escenario 4:** $\sigma^2 = 4$**,** $\phi = 2{,}5$

Modelo	Método	$\mu_1(t) = 5$ $\mu_2(t) = 6$	$\mu_1(t) = 5$ $\mu_2(t) = 10$	$\mu_1(t) = 5$ $\mu_2(t) = 15$
Gaussiano	WTV	96,19	99,97	99,96
	WMV	96,64	99,99	99,96
Esférico	WTV	87,79	97,14	97,38
	WMV	92,04	96,47	96,81
Matérn	WTV	98,72	100	100
$\kappa = 0,2$	WMV	92,04	100	100
Matérn	WTV	98,79	100	100
$\kappa = 2$	WMV	99,26	100	100

Cuadro 4.10: PCC: Escenario 4.

De los resultados mostrados en las tablas 4.7-4.10, se observa que el PCC incrementa ligeramente cuando la distancia entre las medias funcionales se incrementa, tanto para el método **WTV** como para el método **WMV**, mostrando resultados en general superiores al 80 %, a excepción del escenario 3 (tabla 4.9) cuando el modelo es esférico, puesto que se obtuvo un valor de PCC del 77,31 % siendo el más bajo. Por otro lado, se observa

que ambas ponderaciones dan resultados similares mostrando que no existe una notable diferencia entre estos.

4.1.2.1. Ruido blanco

Para resaltar la importancia de considerar la dependencia espacial en el proceso de clasificación, se presentan los siguientes escenarios.

4.1.2.2. Clasificación considerando coordenadas

En este caso, se simularon datos sin dependencia espacial; es decir, en el modelo 4.1 se tiene que $\epsilon(t) \sim \mathcal{N}_{120}(0, \sigma^2 I)$, sin embargo, para la clasificación se tomó en cuenta la ubicación espacial de los puntos obteniendo los siguientes resultados:

Modelo	Método	$\mu_1(t) = 5$ $\mu_2(t) = 5$	$\mu_1(t) = 5$ $\mu_2(t) = 6$	$\mu_1(t) = 5$ $\mu_2(t) = 10$	$\mu_1(t) = 5$ $\mu_2(t) = 15$
Ruido	WTV	44,54	84,73	87,44	87,34
blanco	WMV	79,65	99,21	88,60	88,62

Cuadro 4.11: PCC: Datos simulados sin correlación espacial.

Se puede observar que pese a que no existe correlación espacial entre las ubicaciones los resultados obtenidos son buenos, puesto que se tiene que a mayor distancia de separación de las medias los valores de PCC aumentan en el caso de la ponderación realizada mediante el trazo-variograma (WTV). Este comportamiento se debe a que el algoritmo al momento de tomar las coordenadas como un parámetro de entrada, en el proceso de agrupación, toma las ubicaciones que están más cercanas y que al mismo tiempo tengan una forma similar de las curvas. Por otro lado para el caso de ponderación mediante el variograma multivariado, los valores de PCC son superiores a los de la primera ponderación; sin embargo, se puede notar que a medida que las medias se alejan el método tiende a clasificar erróneamente; esto se debe a que no existe una estructura espacial definida en los datos que ayude a distinguir los grupos cuando estos se ubican en cuadrantes con la misma media funcional. Para el caso de medias funcionales iguales, el PCC es cercano al 50 % y 80 % en el método WTV y WMV respectivamente, puesto que las formas de las curvas son similares y las coordenadas de las ubicaciones no tienen asignadas una forma funcional específica.

4.1.2.3. Clasificación sin considerar coordenadas

Al igual que en el caso anterior, se tiene que $\epsilon(t) \sim \mathcal{N}_{120}(0, \sigma^2 I)$; sin embargo, en este caso no se consideraron las ubicaciones espaciales para su clasificación; es decir, solo se aplicó el algoritmo jerárquico funcional.

Modelo	$\mu_1(t) = 5$ $\mu_2(t) = 5$	$\mu_1(t) = 5$ $\mu_2(t) = 6$	$\mu_1(t) = 5$ $\mu_2(t) = 10$	$\mu_1(t) = 5$ $\mu_2(t) = 15$
Ruido blanco	33,02	59,06	58,48	58,54

Cuadro 4.12: PCC: Clasificación jerárquico funcional base.

En este caso, salvo para el caso de medias funcionales iguales donde el valor de PCC es el peor, se observa que estos valores no sobrepasan el 60 %. Al igual que el algoritmo k-medias funcional no es capaz de diferenciar qué tipo curvas están originadas en las ubicaciones espaciales.

4.2. Caso de aplicación

En esta sección se aplicará el algoritmo K-medias modificado sobre una base de datos del Índice de Vegetación de Diferencia Normalizada (NDVI). Como se ha mencionado, la dependencia espacial de los datos ambientales es un criterio influyente en los procesos de agrupación, debido a que los resultados obtenidos proporcionan información relevante, especialmente para la caracterización de las zonas geográficas resultantes [Zapata-Rios *et al.*, 2021]. Por otro lado, los estudios de NDVI son importantes puesto que se utilizan principalmente para evaluar la salud de los cultivos, ayudar a predecir las zonas de peligro de incendio, medir la biomasa, evaluar los efectos de la transpiración de las plantas en los balances de agua y energía de la superficie [Jiang *et al.*, 2006] y otros. Este índice se calcula a partir de la luz visible e infrarroja cercana reflejada por la vegetación.

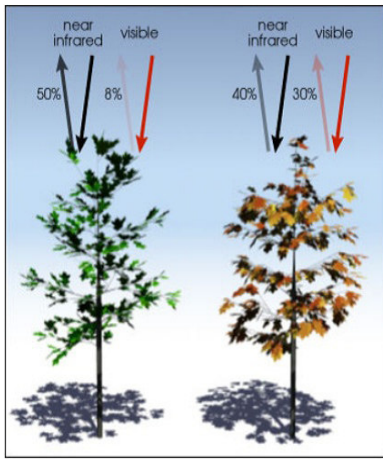


Figura 4.8: Fuente: climatedatalibrary.cl

En la figura 4.8, la vegetación sana (izquierda) absorbe la mayor parte de la luz visible que incide sobre ella y refleja una gran parte de la luz infrarroja cercana. Y la vegetación poco saludable o escasa (derecha) refleja más luz visible y menos luz infrarroja cercana [KheirkhahZarkesh *et al.*, 2014].

4.2.1. Datos

Los datos fueron proporcionados por el proyecto PIJ-17-05 desarrollado por la Escuela Politécnica Nacional sobre "Los patrones climáticos globales y su influencia en la respuesta temporal y espacial de índices espectrales de la vegetación del páramo en el Ecuador". La base de datos cuenta con 10.000 pixeles distribuidos a través de la cordillera de los Andes; en cada ubicación se tiene como observaciones 365 promedios diarios del NDVI desde 2001 hasta 2018. En la figura 4.9 se observa el comportamiento de este conjunto de datos.

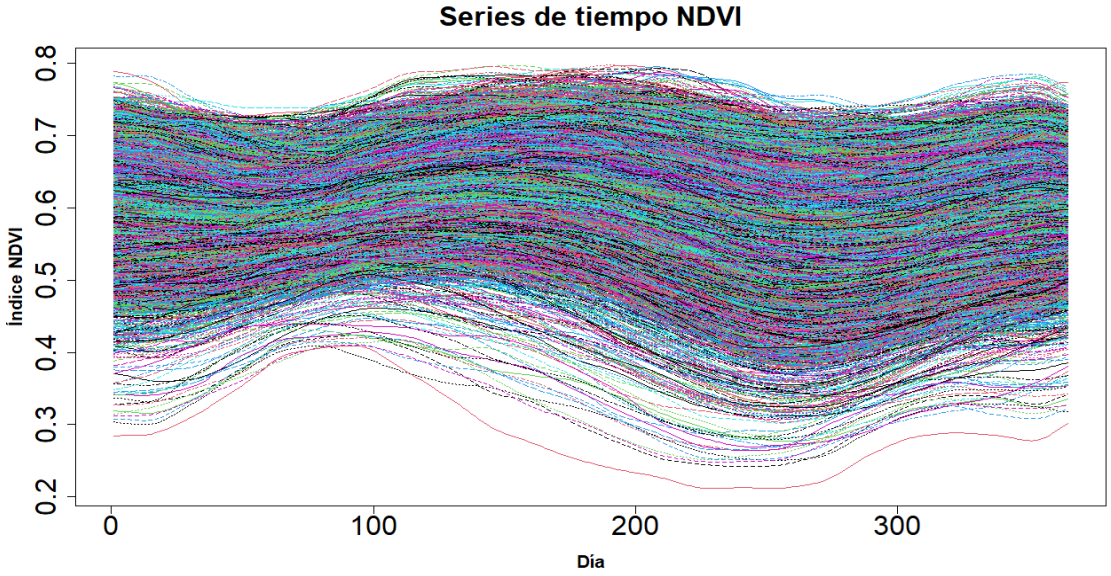


Figura 4.9: Mediciones del NDVI.

En la figura 4.10 se muestra un mapa de calor de un día de observaciones, donde las zonas azules corresponden a un valor bajo de este índice, indicando zonas áridas y las zonas rojas corresponden a un valor alto, indicando zonas de gran vegetación. Sin embargo, existen sectores en donde a distancias pequeñas se presenta un cambio brusco de estos valores; este fenómeno aporta variabilidad a los datos. Esta variabilidad existente puede tener dos razones de ser: la primera propia de errores de medición y la segunda propia de la zona geográfica, puesto que el Ecuador es un país megadiverso y existen ecosistemas que son diferentes en una misma área [Zapata-Rios *et al.*, 2021].

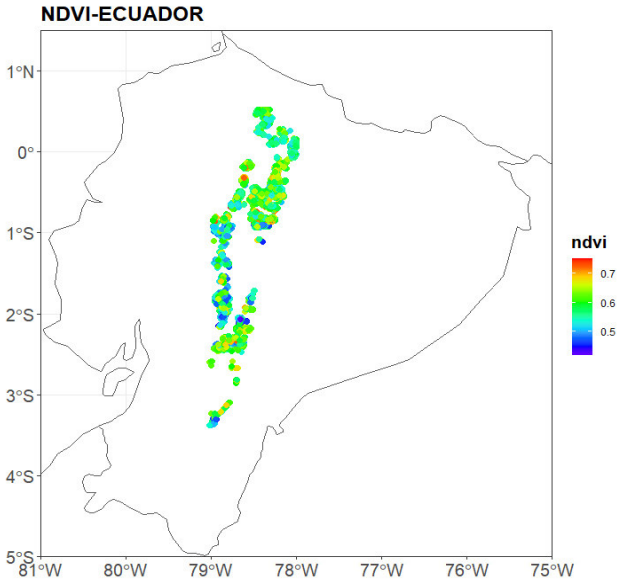
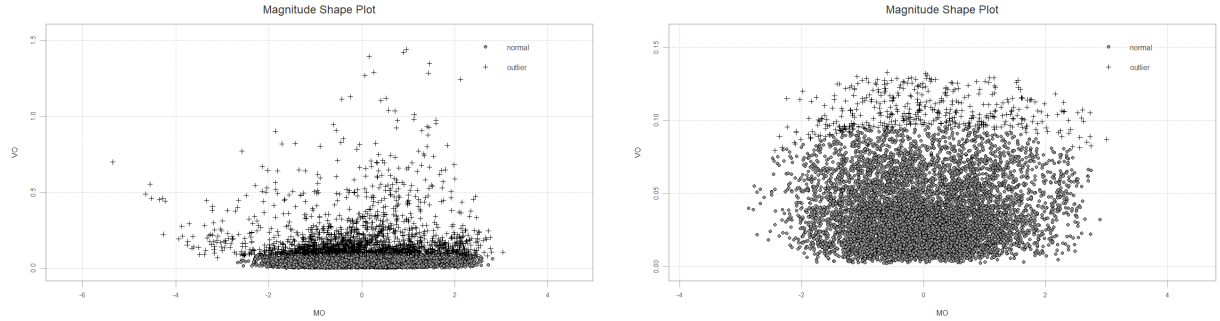


Figura 4.10: Mapa de calor.

4.2.2. Metodología

Siguiendo la metodología presentada en los capítulos previos y utilizando el método MS-Plot con la distancia de proyección de profundidad se procede a depurar la base de datos; para esto se identificaron los datos atípicos los cuales representan el 16,23 % de la información; por lo que se procedió a retirarlos pues no se considera como una pérdida significativa de datos. Es así que se trabajó con información de 8.377 pixeles.



(a) Antes (b) Después

Figura 4.11: Detección y eliminación de curvas atípicos.

Ahora, para tratar el problema de la alta dimensionalidad de esta base de datos se hizo uso de 65 bases de expansión de tipo Fourier. Este número de bases fue obtenido aplicando la función *optim.basis* de la librería *fda.usc*.

Luego, se verifica que los datos no presenten tendencia alguna; en este caso, se corrigió la tendencia presente mediante una regresión de segundo orden dada por la siguiente ecuación:

$$Z_i = \hat{\alpha} + \hat{\beta}_{1\,\text{Longitud}_i} + \hat{\beta}_{2\,\text{Latitud}_i} + \hat{\beta}_{3\,\text{Longitud}_i^2} + \hat{\beta}_{4\,\text{Latitud}_i^2} + \hat{\beta}_{5\,\text{Longitud}_i * \text{Latitud}_i}, \quad i = 1, ..., N$$

donde N es el tamaño del conjunto de datos, $\hat{\alpha}$, $\hat{\beta}_k$, $k = 1, .., 5$ son los coeficientes de la regresión, *Longitud* representa la coordenada **x** y *Latitud* representa la coordenada **y**. Los coeficientes de la regresión son los siguientes:

Coeficiente	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
Valor	-201309,49	-5124,89	-2199,00	-32,58	1,52	-27,95

Cuadro 4.13: Coeficientes del modelo de regresión.

En la figura 4.12 se puede apreciar que la base de datos no presenta tendencia ni en las coordenadas **x** ni en las coordenadas **y**.

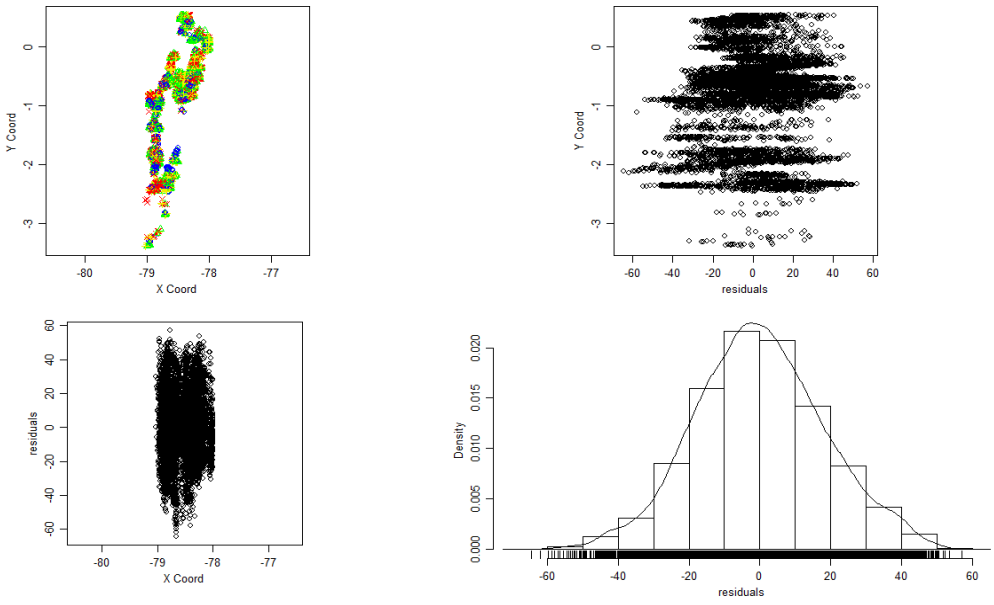


Figura 4.12: Tendencia de los datos.

Después se verifica un aspecto fundamental que es el criterio de isotropía. En la figura 4.13 se observa que para los diferentes ángulos considerados el variograma se encuentra acotado; por lo tanto, este requerimiento se cumple.

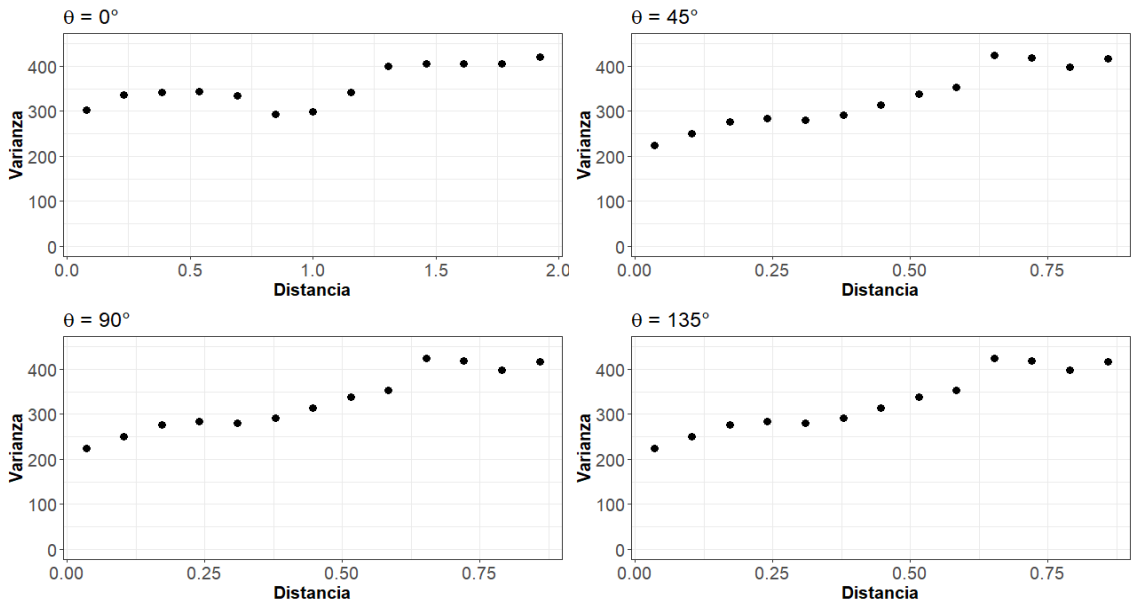


Figura 4.13: Ángulos para el criterio de isotropía: $\theta = 0^o, 45^o, 90^o, 135^o$.

Ahora bien, dado que el algoritmo requiere la estimación del variograma, se ajustó un variograma teórico utilizando las funciones *eyefit* y *variofit* del software estadístico **R** y se obtuvieron los siguientes parámetros: un modelo **"powered.exponential"** como modelo de covarianza, un efecto pepita de 300,09, una silla de 200,06, un rango de 1,93, la distancia máxima de 2 y finalmente el parámetro kappa igual a 1,89. Este variograma se observa en la figura 4.14

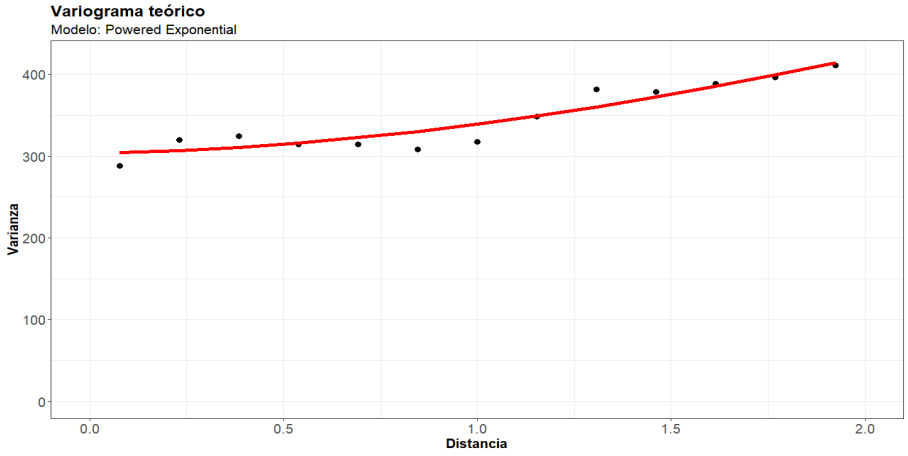
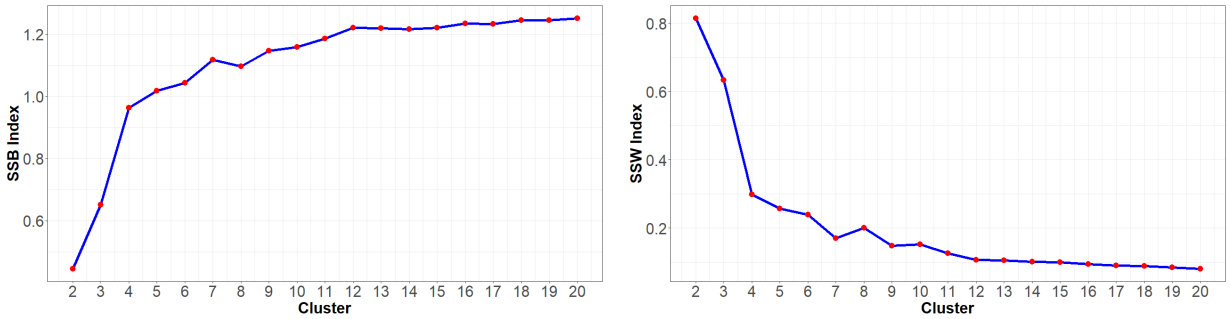


Figura 4.14: Variograma teórico.

Cómo se ha mencionado, el algoritmo K-medias requiere como argumento de entrada el número de grupos; por tanto, se hizo uso de los índices expuestos en la subsección 3.3.1



(a) Índice SSB (b) Índice SSW

Figura 4.15: Selección del número de grupos.

De los resultados obtenidos mostrados en la figura 4.15 y conjunto con el criterio experto, proporcionado por la MSc. Sandra Torres investigadora medioambiental del Instituto Nacional de Meteorología e Hidrología (INAMHI), se tomó la decisión de trabajar con 5 grupos.

Una vez verificados los criterios y obtenidos los parámetros iniciales se procede a aplicar el algoritmo K-medias modificado validado en la sección 4.1.

4.2.3. Resultados

Ahora se presentan los resultados obtenidos una vez aplicado el algoritmo K-medias funcional y el algoritmo k-medias funcional espacial. Estos resultados muestran los grupos distribuidos sobre el mapa de demarcación hidrográfica del Ecuador, puesto que es natural pensar que la disponibilidad de agua en las zonas tiene cierta influencia con el NDVI. Por otro lado, se conoce que una cuenca hidrográfica se define como una unidad territorial en la cual el agua que cae por precipitación se reúne y escurre a un punto común o que fluye al mismo río, lago, o mar [CNRH, 2002]. Además, se muestran las correlaciones temporales y espaciales resultantes.

Estos resultados fueron obtenidos mediante el uso del software R y los recursos computacionales proporcionados por el Laboratorio Nacional de Cálculo Científico HPC MODEMAT.

4.2.3.1. K-medias funcional

Se muestran los resultados obtenidos al aplicar el algoritmo K-medias funcional sobre el mapa de demarcación hidrográfica del Ecuador, imagen izquierda en la figura 4.16. Se puede observar claramente que no existe un patrón claro en la distribución espacial de los grupos. También se presentan las correlaciones temporales y espaciales. En cuanto a las correlaciones temporales se obtuvieron valores positivos y altos al comparar cada centroide con sus respectivos miembros y entre los miembros de cada grupo; esto debido a la naturaleza del algoritmo. Por otro lado, se calculó el índice global de Moran que resultó ser de 0,127, indicando que existe correlación espacial positiva y baja; es decir, los grupos podrían estar solapándose, lo que efectivamente sucede; esto se aprecia en el gráfico 4.16. Además, se midió la correlación espacial dentro de cada grupo mediante el índice de Geary, obteniéndose valores positivos y cercanos a cero, lo que indica un indicio de correlación espacial positiva.

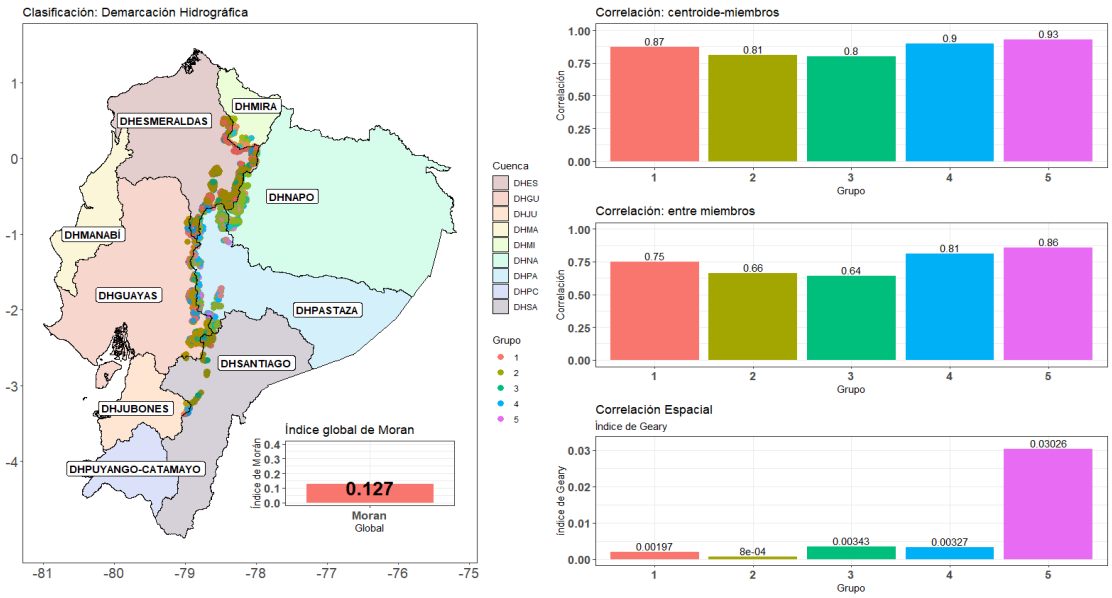


Figura 4.16: Resultados y correlaciones.

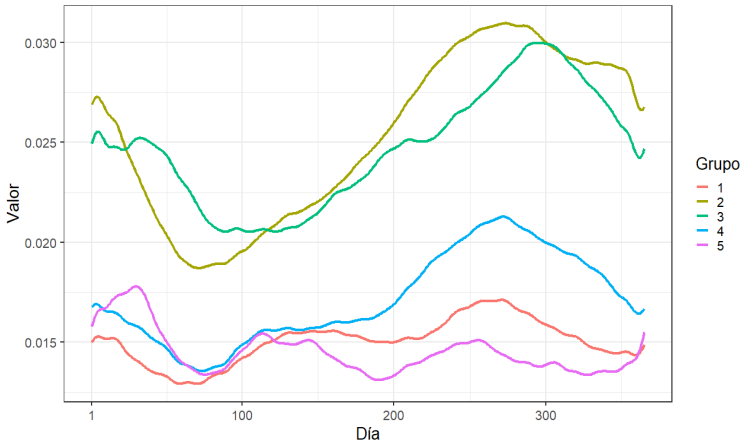


Figura 4.17: Curvas de desviaciones estándar.

Para tener una mejor visualización de los resultados obtenidos se presentan los grupos en el espacio y su respectivo conjunto de curvas. En la figura 4.18 se observa que el grupo uno, con 1.789 miembros, se encuentra disperso en el mapa. El grupo dos, con 4.543 miembros, presentó un patrón espacial, debido a que está conformado por la mayor parte del conjunto de datos. El conjunto de curvas del grupo uno tienen mayor similitud que las del grupo dos, lo que se verifica en la figura 4.17 puesto que los valores de la curva de desviación estándar del grupo uno son menores que los del grupo dos. Este comportamiento se mantiene en las correlaciones temporales, pues la correlación entre el centroide y miembros fue de 0,87 y 0,81; la correlación entre miembros fue de 0,75 y 0,66 para el grupo uno y dos respectivamente (ver figura 4.18). Además, se observa que espacialmente el grupo dos está mejor formado que el grupo uno, pues su correlación es la más cercana a cero.

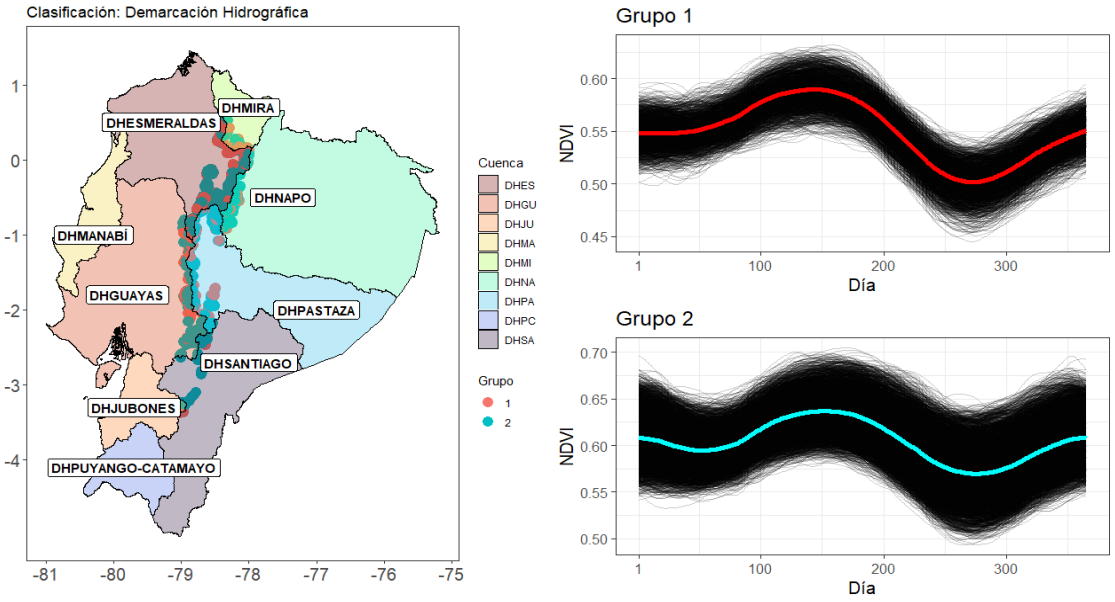


Figura 4.18: Clasificación: Grupo 1 y Grupo 2.

Los grupos tres y cuatro, con 1.005 y 951 miembros respectivamente, se observan en la figura 4.19. En cuanto a la ubicación espacial de los grupos, como se puede observar en el mapa, se encuentran dispersos y no hay gran diferencia en los valores de correlación espacial de los mismos. Por otro lado, el conjunto de curvas del grupo tres es menos compacto que el conjunto de curvas del grupo cuatro, lo que se evidencia en la figura 4.17. Por otro lado, el grupo cuatro es más homogéneo que el grupo tres pues la correlación temporal entre el centroide y los miembros fue de 0,9 y 0,8; la correlación entre miembros fue de 0,81 y 0,64 respectivamente.

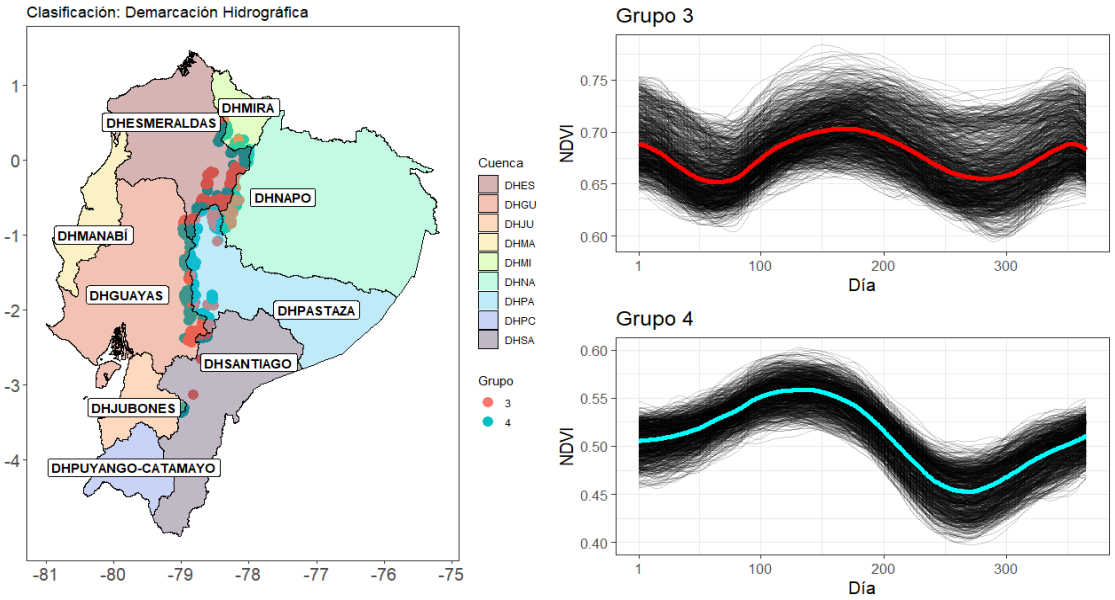


Figura 4.19: Clasificación: Grupo 3 y Grupo 4.

Ahora, en la figura 4.20 se observan los resultados del grupo cinco con 89 miembros. En este caso los valores de la curva de desviación estándar de manera general son los más

bajos (ver figura 4.17). Sin embargo, los valores de las correlaciones temporales son los más altos. De manera similar que los demás grupos no existe un patrón espacial claro y su correlación espacial es la más alejada de cero.

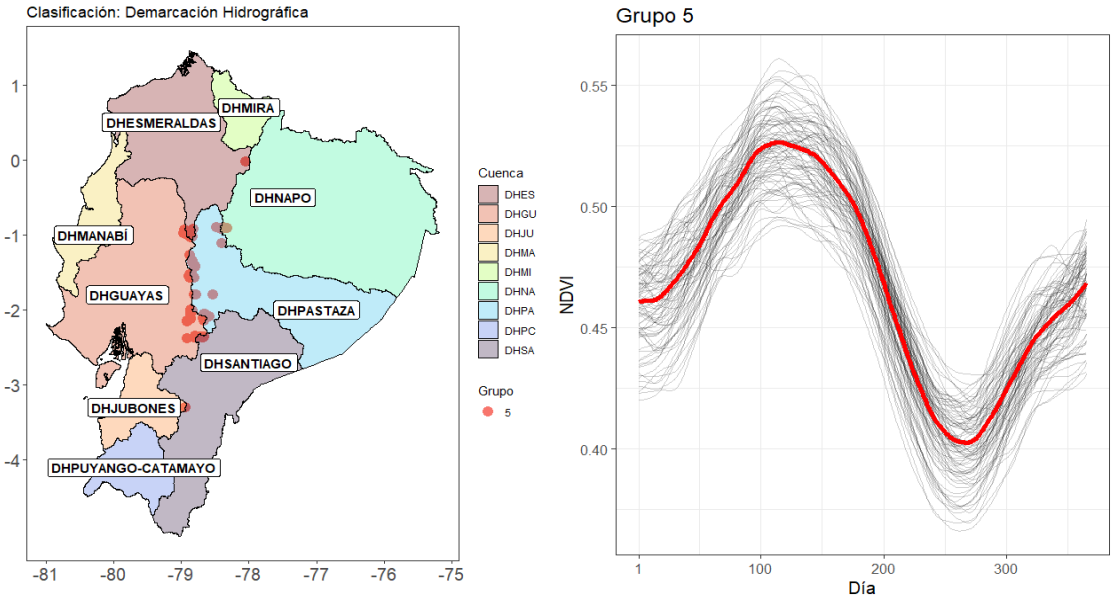


Figura 4.20: Clasificación: Grupo 5.

Finalmente, para comprobar que los centroides de los grupos resultantes no sean iguales se calculó un ANOVA funcional. Como se puede observar en la figura 4.21 el gráfico de densidad indica un p valor de 0; por lo tanto, se rechaza la hipótesis de que los grupos tengan centroides iguales. A partir de este resultado y en conjunto con las correlaciones temporales obtenidas se valida la formación de los cinco grupos.

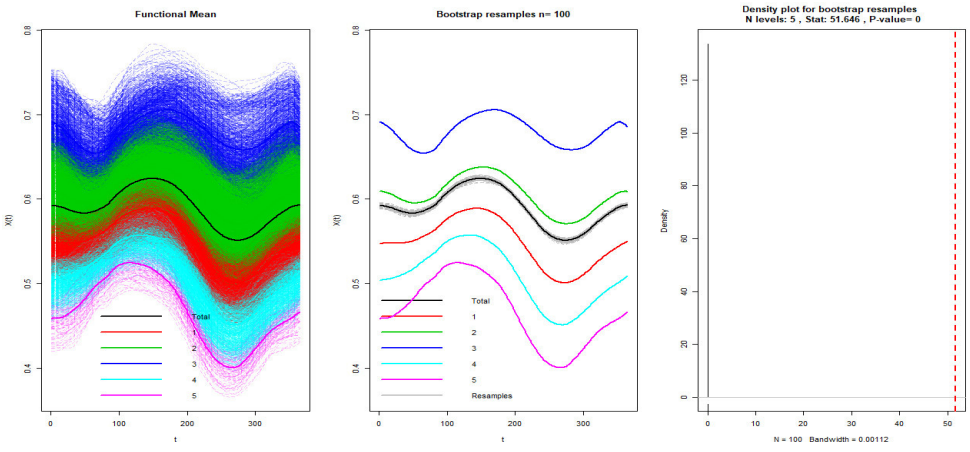


Figura 4.21: ANOVA funcional.

4.2.3.2. K-medias funcional espacial

En este caso, como la base de datos solo dispone de la variable NDVI se utilizó el método de ponderación mediante el trazo-variograma. Así, se obtuvieron los siguientes resultados.

En la figura 4.22 se presenta la distribución de los grupos sobre el mapa hidrográfico del Ecuador y las respectivas correlaciones temporales y espaciales.

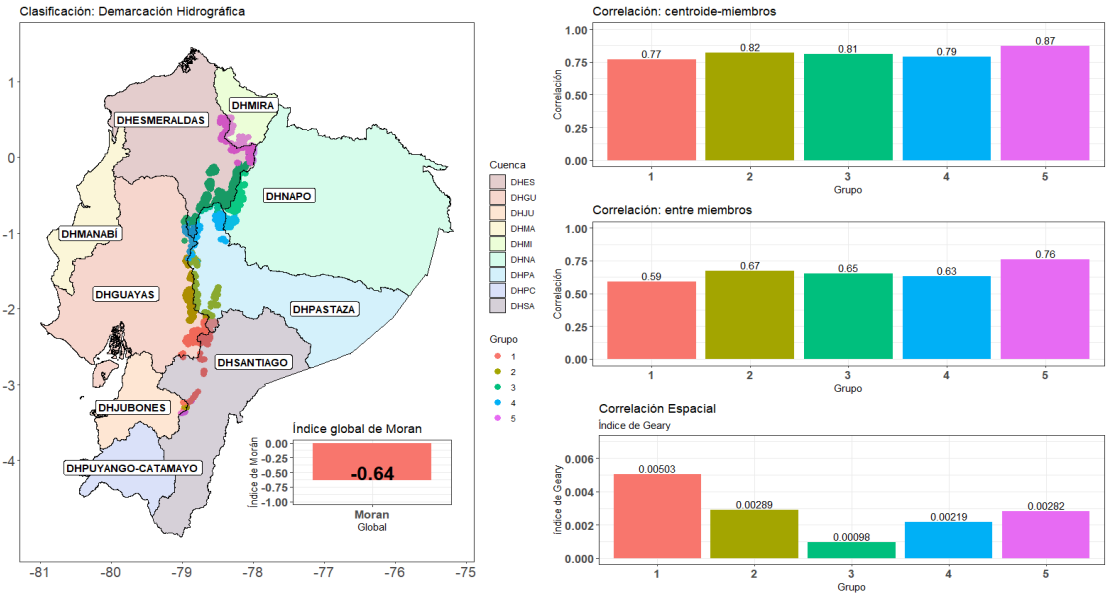


Figura 4.22: Resultados y correlaciones.

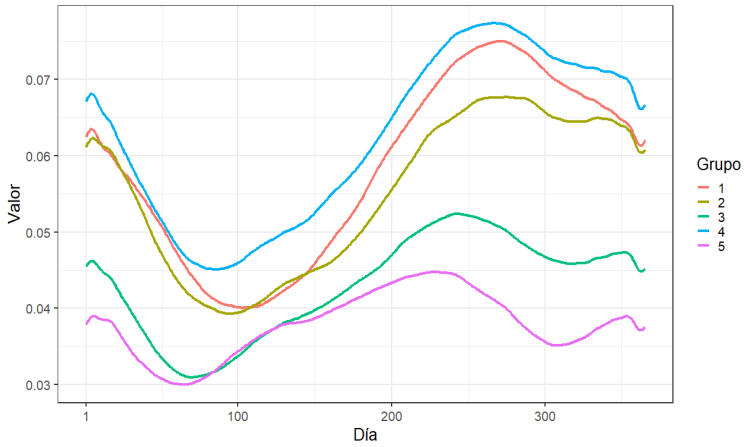


Figura 4.23: Curvas de desviaciones estándar.

A diferencia de los resultados obtenidos bajo el algoritmo K-medias funcional, al considerar la estructura espacial es notorio el cambio en la obtención de grupos, pues en el mapa hidrográfico se puede identificar claramente que los grupos formados se encuentras distribuidos latitudinalmente sobre las principales cuencas hidrográficas. Se observa que las correlaciones temporales del centroide y los miembros como las correlaciones entre miembros de cada grupo son altas y positivas, indicando que los grupos son homogéneos. Con respecto a la correlación espacial, el valor del índice global de Moran indica una correlación espacial negativa y alta; es decir, que los centroides de los grupos están dispersos e indica que de cierta manera los grupos no se estarían solapando. Por otro lado, se calculó la correlación espacial dentro de cada grupo mediante el índice de Geary; estos valores nos indican que existe correlación espacial positiva dentro de cada grupo.

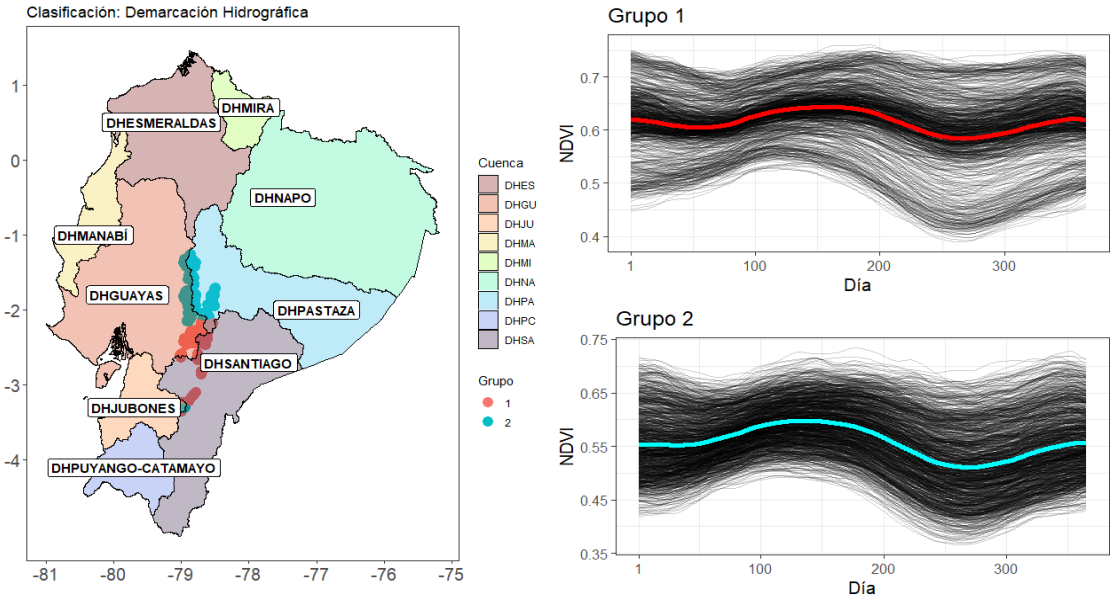


Figura 4.24: Clasificación: Grupo 1 y Grupo 2.

A continuación, se presentan varios gráficos en los que se puede observar de mejor manera la distribución espacial de los grupos obtenidos y su conjunto de curvas correspondiente. En la figura 4.24 se observa que la mayor parte del grupo uno, con 829 miembros, se encuentra sobre las cuencas hidrográfica de los ríos Santiago y Guayas. Por otro lado, el grupo dos, con 1.156 miembros, se encuentra sobre las cuencas hidrográficas de los ríos Pastaza y Guayas. Se observa en la figura 4.23 que la curva de desviación estándar de este último presentó valores más bajos que la curva del grupo uno, por lo cual el grupo dos es menos disperso que el grupo uno; este comportamiento también se evidencia en los valores de correlación temporal, pues la correlación entre el centroide y los miembros fue de 0,77 y 0,82; la correlación entre miembros fue de 0,59 y 0,67 para los grupos uno y dos, respectivamente. La correlación espacial del grupo dos es más cercana a cero debido a que este grupo espacialmente es más compacto que el grupo uno.

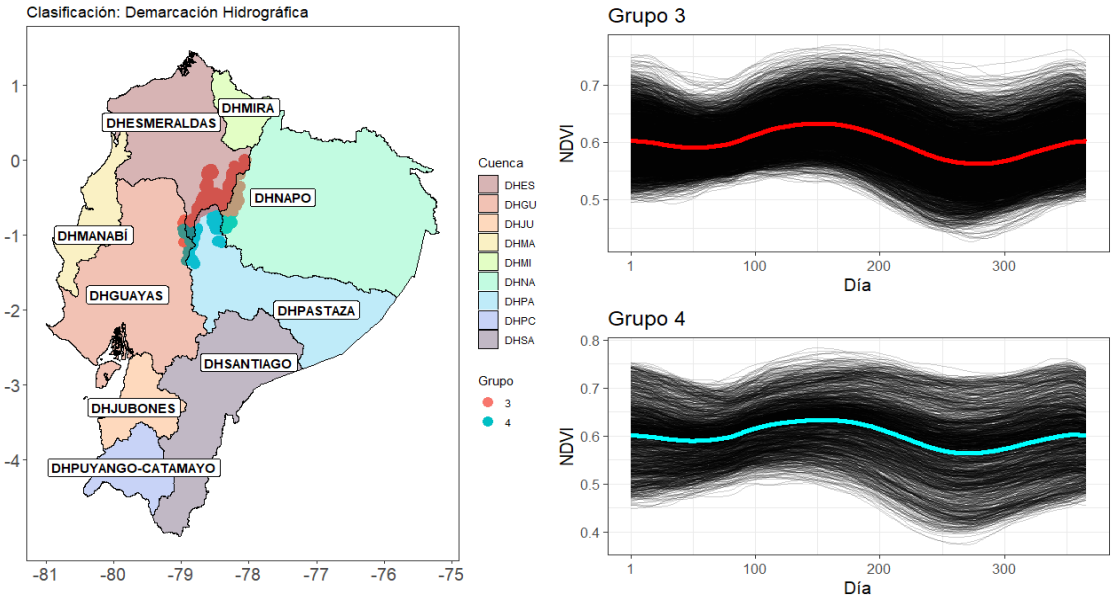


Figura 4.25: Clasificación: Grupo 3 y Grupo 4.

Por otro lado, en la figura 4.25 se presenta la distribución geográfica del grupo tres y cuatro sobre el mapa de demarcación hidrográfica. Se observa que el grupo tres, con 3.803 miembros, se encuentra principalmente sobre las cuencas hidrográficas de los ríos Esmeraldas y Napo; el grupo cuatro, con 1.435 miembros, geográficamente se encuentra sobre las cuencas hidrográficas de los ríos Napo, Pastaza y Guayas. A diferencia de los grupos anteriores los valores de la curva de desviación estándar de este último grupo son más altas, por lo que los miembros de este grupo presentan más dispersión que el grupo tres (ver figura 4.23). La correlación temporal entre el centroide y los miembros fue de 0,79 y 0,81; la correlación temporal entre los miembros fue de 0,65 y 0,63 para los grupos tres y cuatro, respectivamente. En cuanto a la correlación espacial, el grupo tres tiene el valor más pequeño de todos los cinco grupos, por lo que espacialmente es el más compacto (ver figura 4.22).

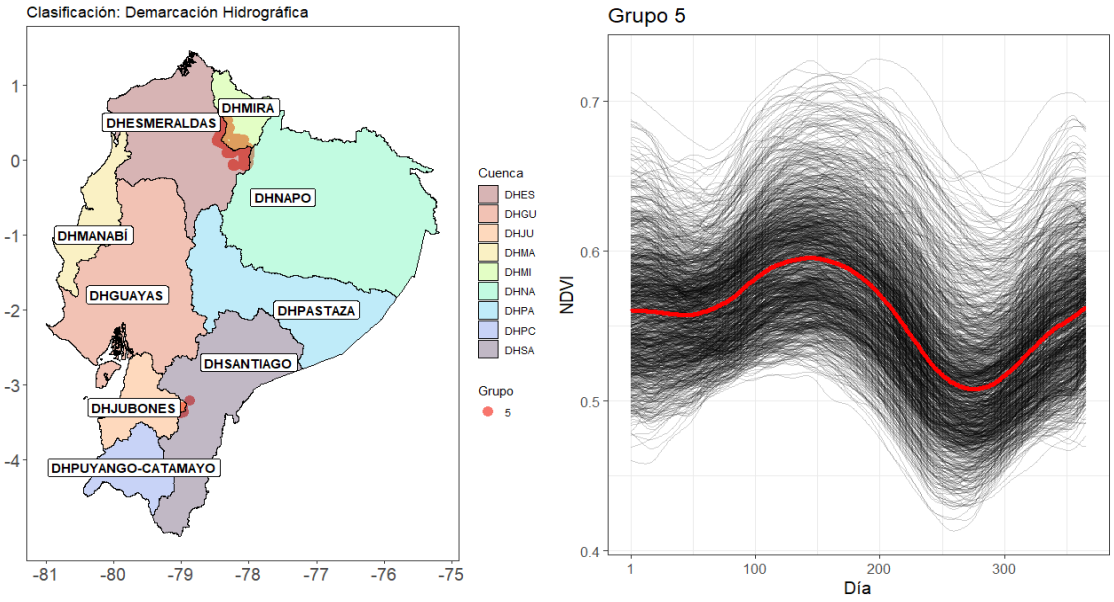


Figura 4.26: Clasificación: Grupo 5.

En la figura 4.26 se tiene al grupo cinco con 1.154 miembros. Este grupo espacialmente se encuentra sobre las cuencas hidrográficas de los río Mira, Esmeraldas y una pequeña porción sobre la cuenca hidrográfica del río Santiago. Los valores de la curva de desviación de este grupo son los más bajos de todos los cinco grupos obtenidos, indicando que este es el grupo más compacto en cuanto a forma; además, en cuanto a correlación temporal entre el centroide y los miembros, así como la correlación entre miembros son las más altas con valores de 0,87 y 0,76, respectivamente. Adicionalmente, tiene correlación espacial con valor de 0,00282, siendo el tercer grupo espacialmente más compacto de los cinco grupos obtenidos.

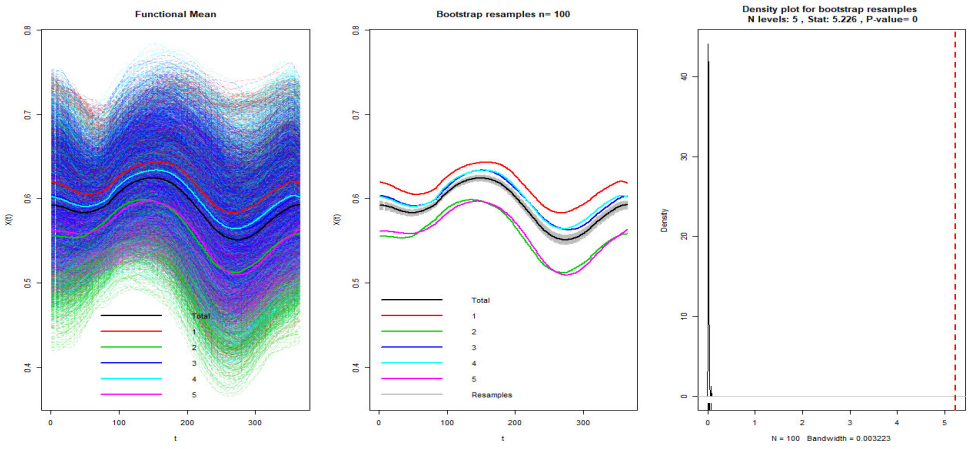


Figura 4.27: ANOVA funcional.

Finalmente, se calculó un ANOVA funcional para comprobar que los centroides de cada grupo no sean iguales. Como se ve en el gráfico de densidad de la figura 4.27 se obtuvo un p valor de 0, con lo cual se rechaza la hipótesis de trabajar con centroides iguales y con los resultados de correlación temporal y espacial que se obtuvieron se valida la formación de los cinco grupos.

Capítulo 5

Conclusiones y recomendaciones

5.1. Conclusiones

Las conclusiones se enfocan en dos aspectos principales; la primera en cuanto a la metodología utilizada y el desempeño del algoritmo K-medias modificado, y la segunda en cuanto a los resultados obtenidos en el caso de aplicación a los datos del NDVI.

- La metodología utilizada para la modificación del algoritmo k-medias es adecuada para la agrupación de datos funcionales con correlación espacial, puesto que como se evidenció en la sección de simulaciones 4.1, los resultados obtenidos en la mayoría de casos superan el 84 % de PCC. Por otro lado, el método de **WTV** ponderación mediante el trazo-variograma arrojó los mejores resultados a pesar de que no considerar toda la información de los datos y por esto su costo computacional es menor. Sin embargo, se evidenció que ambas ponderaciones son sensibles a la ubicación de las medias funcionales, pues al incrementarse la distancia entre las medias funcionales los valores de PCC aumentan.

- El algoritmo k-medias modificado presenta un clara ventaja sobre el algoritmo k-medias funcional tradicional, puesto que los resultados obtenidos con esté último algoritmo no sobrepasan el 55 % de PCC, obteniéndose grupos con distribución espacial casi aleatoria. Por otro lado, pese a que no se presente una estructura espacial clara en los datos simulados, el algoritmo k-medias modificado fue capaz de identificar grupos espaciales homogéneos con medias funcionales iguales, pues los valores de PCC sobrepasan el 75 % a excepción del caso en donde las medias funcionales son iguales. Por tanto, cuando el algoritmo considera información adicional dada por la estructura espacial presente en los datos produce una mejor clasificación, lo que posibilita una caracterización más completa de las zonas espaciales resultantes.

- Al comparar el algoritmo jerárquico funcional espacial con el algoritmo k-medias modificado en los escenarios de simulación, se observó que el primer algoritmo mantiene una ligera ventaja sobre el algoritmo modificado, pues en algunos escenarios los valores de PCC alcanzaron el 100 % y no bajaron del 77 %; esto debido a la naturaleza del algoritmo jerárquico. Por otro lado, el algoritmo modificado tuvo un valor máximo de PCC de 99,40 % y estos valores no bajaron del 84,7 %. Por lo tanto, la agrupación de este último algoritmo resulta ser igual de óptima que el algoritmo jerárquico funcional para datos con dependencia espacial.

- Los datos del NDVI con los cuales se trabajó presentaron una alta variabilidad, lo que dificultó la clasificación espacial para el método. Se observó en la figura 4.10 que la distribución espacial con respecto a los valores del NDVI no presentan un patrón en particular y de la misma manera esto se vio reflejado en la estimación teórica del variograma 4.14. Se debe tener en cuenta que la estimación adecuada del variograma teórico es de suma importancia en cualquier proceso de estadística espacial, en particular en la metodología presentada pues este influye directamente en los resultados obtenidos. Así, se evidenció la presencia del efecto pepita con un valor de 300,09, llegando a su silla de valor 200,06, un alcance de 1,93 y la máxima distancia con la que se trabajó fue de 2, por lo que el comportamiento tanto de los datos como del variograma es muy irregular.

- Los índices SSB y SSW, propuestos en la sección de índices funcionales, tienden a comportarse bajo lo esperado. El primero indica que a mayor número de grupos su valor aumenta; sin embargo, teniendo en cuenta que los valores de este índice son pequeños se evidenció que los grupos formados son compactos con respecto a su forma funcional. El segundo índice indica que a medida que el número de grupos aumenta su valor disminuye, evidenciando que los grupos están separados entre ellos pero no lo suficiente puesto que sus valores también son pequeños. En ambos casos, el número óptimo de grupos puede ser de 4 a 12, así en conjunto con el criterio experto se tomó la decisión de trabajar con cinco grupos.

- Las correlaciones temporales tanto para el método k-medias funcional base como para el método k-medias modificado presentan resultados similares, altos y positivos, pese a que la clasificación de curvas arrojados por el primer algoritmo pareciera ser más compacta; esto debido a la naturaleza del algoritmo.

- Se utilizaron dos índices para medir la correlación espacial entre los grupos formados y dentro de los mismos. Para el primer caso, se empleó el índice global de Moran, que arrojó un valor de 0,127 en el caso del algoritmo k-medias funcional base, indicando una baja correlación, lo que muestra que los grupos podrían estar solapándose. En el caso del algoritmo k-medias funcional espacial, el valor de la correlación fue de $-0{,}64$, indicando que los grupos se encuentran dispersos sobre el área geográfica. Para el segundo caso, se empleó el índice de Geary, el que arroja el valor a través de una comparación de curva a curva dentro de cada grupo formado; en ambos algoritmos empleados se obtuvieron valores cercanos a cero, que indican correlación positiva; no obstante, se puede observar una mínima diferencia que indica una mayor correlación espacial en los grupos obtenidos al aplicar el algoritmo k-medias funcional espacial. En el caso funcional, la distribución espacial de los grupos no es tan diferenciada, puesto que se observó que todos los grupos están distribuidos a lo largo de la zona de estudio de forma regular, y por la naturaleza del índice de Geary, los resultados arrojados indican correlación espacial fuerte.

- Al aplicar el algoritmo k-medias funcional espacial a la base de datos del NDVI, los resultados obtenidos fueron cinco grupos distribuidos latitudinalmente sobre las principales cuencas hidrográficas del Ecuador. Se debe tener en cuenta que pese a que el variograma estimado no presentó una dependencia espacial fuerte, el algoritmo pudo formar grupos homogéneos desde un punto de vista espacial.

- Los resultados obtenidos en cuanto al caso de simulación y aplicación muestran que el método K-medias modificado da buenos resultados y altos porcentajes de correcta clasificación, por lo que puede ser utilizado como herramienta de clasificación para datos funcionales con correlación espacial.

- En los últimos años se han desarrollado tecnologías para recolectar datos, lo que ha posibilitado que se puedan obtener bases de datos de alta dimensionalidad, y trabajar con esta información es de cierta manera difícil; en estas situaciones hacer uso del análisis funcional es clave puesto que al pasar de un campo de dimensión infinita a uno de dimensión finita se puede hacer uso las herramientas de la estadística clásica conocidas. Del mismo modo, considerar la estructura espacial de los datos genera resultados mucho más completos. Por lo tanto, el análisis funcional y la estadística espacial resultan herramientas necesarias en la actualidad.

5.2. Recomendaciones

Las recomendaciones se enfocan en dos aspectos principales; la primera en cuanto a la metodología utilizada y el desempeño del algoritmo K-medias modificado, y la segunda en cuanto a los resultados obtenidos en el caso de aplicación a los datos del NDVI.

- Dado el costo computacional del método de ponderación mediante el variograma multivariado, para su aplicación se recomienda contar con hardware de alto desempeño.

- Se debe considerar la posibilidad de adaptar el código del algoritmo k-medias modificado para su paralelización mediante GPU.

- En futuros estudios se podrían considerar otros tipos de bases de expansión, así como de la implementación de distintos algoritmos con base en la metodología presentada y la posterior creación de un paquete en el CRAN (*Comprehensive R Archive Network*).

- Para la selección del número de grupos, no es suficiente la información que proporcionan los índices SSB y SSW, por lo que es recomendable contar con el criterio experto de un profesional en el campo en el cual se esté desarrollando la investigación.

- Bajo los resultados obtenidos, los futuros estudios que se pueden realizar desde un punto de vista ambiental, son: la caracterización de los grupos obtenidos, encontrar variables medioambientales que estén influenciando la formación de los grupos, conocer si efectivamente la disponibilidad de agua está directamente correlacionada con el NDVI, entre otros.

Bibliografía

[Ambroise & Dang, 1997] Ambroise, Christophe, & Dang, Van Mo. 1997. Clustering of spatial databythe em algorithm. *geoenv i-geostatistics for environmental applications*, **9**(12).

[Andrew B. Lawson, 2002] Andrew B. Lawson, David G.T. Denison. 2002. *Spatial cluster modelling (monographs on statistics and applied probability)*. 1 edn.

[Anil K. Jain, 1988] Anil K. Jain, Richard C. Dubes. 1988. *Algorithms for clustering data*. Prentice Hall Advanced Reference Series : Computer Science. Prentice Hall College Div.

[Arbelaitz *et al.*, 2013] Arbelaitz, Olatz, Gurrutxaga, Ibai, Muguerza, Javier, Pérez, Jesús M., & Perona, Iñigo. 2013. An extensive comparative study of cluster validity indices. *Pattern recognition*, **46**(1), 243–256.

[Bohorquez, 2020] Bohorquez, Martha. 2020. *Estadística espacial y espacio-temporal para campos aleatorios escalares y funcionales*. Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia.

[Bourgault *et al.*, 1992] Bourgault, G., Marcotte, D., & Legendre, P. 1992. The multivariate (co)variogram as a spatial weighting function in classification methods. *Mathematical geology*, **24**, 463–478.

[Celemin, 2009] Celemin, Juan. 2009. Autocorrelación espacial e indicadores locales de asociación espacial: Importancia, estructura y aplicación. **18**(01), 11–31.

[Cesario *et al.*, 2020] Cesario, Eugenio, Vinci, Andrea, & Zhu, Xiaotian. 2020. Hierarchical clustering of spatial urban data. *Pages 223–231 of:* Sergeyev, Yaroslav D., & Kvasov, Dmitri E. (eds), *Numerical computations: Theory and algorithms*. Cham: Springer International Publishing.

[Charu C. Aggarwal, 2013] Charu C. Aggarwal, Chandan K. Reddy. 2013. *Data clustering: Algorithms and applications*. 0 edn. Chapman Hall/CRC Data Mining and Knowledge Discovery Series. Chapman and Hall/CRC.

[Chauhan *et al.*, 2010] Chauhan, Ritu, Kaur, Harleen, & Alam, M.Afshar. 2010. Article:data clustering method for discovering clusters in spatial cancer databases. *International journal of computer applications*, **10**(6), 9–14. Published By Foundation of Computer Science.

[Chen, 2013] Chen, Yanguang. 2013. New approaches for calculating moran's index of spatial autocorrelation. *Plos one*, **8**(07), e68336.

[CNRH, 2002] CNRH, Consejo Nacional de Recursos Hídricos. 2002. *Memoria técnica: División hidrográfica del ecuador, propuesta del cnrh y el grupo interinstitucional para oficializar en el ministerio de relaciones exteriores.*

[Cressie, 1991] Cressie, Noel. 1991. *Statistics for spatial data.* Revised edition edn. Probability and Mathematical Statistics. Wiley-Interscience.

[Cuesta-Albertos & Febrero-Bande, 2010] Cuesta-Albertos, J. A., & Febrero-Bande, M. 2010. A simple multiway anova for functional data. *Test*, **19**, 537–557.

[Cuevas *et al.*, 2004] Cuevas, Antonio, Febrero, Manuel, & Fraiman, Ricardo. 2004. An anova test for functional data. *Computational statistics data analysis*, **47**(1), 111–122.

[Dai & Genton, 2017] Dai, Wenlin, & Genton, Marc. 2017. Multivariate functional data visualization and outlier detection. *Journal of computational and graphical statistics*, **27**(03).

[De Bellefon *et al.*, 2018] De Bellefon, Marie-Pierre, Loonis, Vincent, Fontaine, Maëlle, & Costemalle, Vianney. 2018. *Handbook of spatial analysis with r insee-eurostat.*

[Di Blasi *et al.*, 2013] Di Blasi, J.I. Piñeiro, Martínez Torres, J., García Nieto, P.J., Alonso Fernández, J.R., Díaz Muñiz, C., & Taboada, J. 2013. Analysis and detection of outliers in water quality parameters from different automated monitoring stations in the miño river basin (nw spain). *Ecological engineering*, **60**, 60–66.

[Dimitriadou *et al.*, 2002] Dimitriadou, Evgenia, Dolnicar, Sara, & Weingessel, Andreas. 2002. An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, **67**(02), 137–159.

[Emery, 2013] Emery, Xavier. 2013. *Geoestadística.* Departamento de Ingeniería de Minas, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile.

[Febrero *et al.*, 2008] Febrero, Manuel, Galeano, Pedro, & González-Manteiga, Wenceslao. 2008. Outlier detection in functional data by depth measures, with application to identify abnormal nox levels. *Environmetrics*, **19**(4), 331–345.

[Febrero-Bande & de la Fuente, 2012] Febrero-Bande, Manuel, & de la Fuente, Manuel Oviedo. 2012. Statistical computing in functional data analysis: The r package fda.usc. *Journal of statistical software*, **51**(4), 1–28.

[Giraldo *et al.*, 2012] Giraldo, R., Delicado, P., & Mateu, J. 2012. Hierarchical clustering of spatially correlated functional data. *Statistica neerlandica*, **66**(4), 403–421.

[Giraldo, 2008] Giraldo, Ramón. 2008. *Introducción a la geoestadística: Teoria y aplicación.* Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, pg:8-9.

[Gringarten & Deutsch, 1999] Gringarten, Emmanuel, & Deutsch, Clayton V. 1999. Methodology for variogram interpretation and modeling for improved reservoir characterization.

[Górecki & Smaga, 2015] Górecki, Tomasz, & Smaga, Łukasz. 2015. A comparison of tests for the one-way anova problem for functional data. *Computational statistics*, **30**(01).

[Haggarty *et al.*, 2015] Haggarty, R. A., Miller, C. A., & Scott, E. M. 2015. Spatially weighted functional clustering of river network data. *Journal of the royal statistical society. series c (applied statistics)*, **64**(3), 491–506.

[Hall, 2007] Hall, Peter. 2007. F. ferraty and p. vieu, nonparametric functional data analysis: Theory and practice, springer series in statistics, springer, berlin (2006) isbn 978-0-387-30369-7, p. xx, 268pp 29 illus., hardcover. *Computational statistics data analysis*, **51**(9), 4751–4752.

[Hennig *et al.*, 2015] Hennig, C., Meila, M., Murtagh, F., & Rocci, R. 2015. Handbook of cluster analysis.

[Hébrail *et al.*, 2010] Hébrail, Georges, Hugueney, Bernard, Lechevallier, Yves, & Rossi, Fabrice. 2010. Exploratory analysis of functional data via clustering and optimal segmentation. *Neurocomputing*, **73**(03), 1125–1141.

[Jacques, 2014] Jacques, J., Preda C. 2014. Functional data clustering: a survey. *Advances in Data Analysis and Classification*, **8**(3), 24.

[Jain, 2010] Jain, Anil K. 2010. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, **31**(8), 651–666. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR).

[James & Sugar, 2003] James, Gareth, & Sugar, Catherine. 2003. Clustering sparsely sampled functional data. *Jasa. journal of the american statistical association*, **98**(06).

[James Ramsay, 2009] James Ramsay, Giles Hooker, Spencer Graves (auth.). 2009. *Functional data analysis with r and matlab*. 1 edn. Use R. Springer-Verlag New York.

[Jean Paul Chiles, 2012] Jean Paul Chiles, Pierre Delfiner(auth.), Walter A. Shewhart Samuel S. Wilks(eds.). 2012. *Geostatistics: Modeling spatial uncertainty, second edition*. Wiley Series in Probability and Statistics.

[Jiang *et al.*, 2006] Jiang, Zhangyan, Huete, Alfredo R., Chen, Jin, Chen, Yunhao, Li, Jing, Yan, Guangjian, & Zhang, Xiaoyu. 2006. Analysis of ndvi and scaled difference vegetation index retrievals of vegetation fraction. *Remote sensing of environment*, **101**(3), 366–378.

[KheirkhahZarkesh *et al.*, 2014] KheirkhahZarkesh, Mir Masoud, Darvishi, Mehdi, Akbar Abkar, Ali, & Ahmadi, Gholam Reza. 2014. Estimation of rice vegetation indices with multitemporal radar and optic images. *Physical geography research quarterly*, **45**(4), 85–96.

[Kogan *et al.*, 2006] Kogan, Jacob, Nicholas, Charles, & Teboulle, Marc. 2006. *Grouping multidimensional data. recent advances in clustering*.

[Kokoszka, 2017] Kokoszka, Piotr; Reimherr, Matthew. 2017. *Introduction to functional data analysis*. 1st edn. Chapman Hall/CRC texts in statistical science series. Chapman Hall / CRC.

[Kreider, 1971.] Kreider, Donald L. 1971.. *Introducción al análisis lineal /*. Bogotá :: Fondo Educativo Interamericano,.

[Liu *et al.*, 2013] Liu, Yanchi, Li, Zhongmou, Xiong, Hui, Gao, Xuedong, Wu, J., & Wu, Sen. 2013. Understanding and enhancement of internal clustering validation measures. *Ieee transactions on cybernetics*, **43**, 982–994.

[Luz López García *et al.*, 2015] Luz López García, María, García-Ródenas, Ricardo, & González Gómez, Antonia. 2015. K-means algorithms for functional data. *Neurocomputing*, **151**, 231–245.

[Montero *et al.*, 2015] Montero, José, Fernández-Avilés, Gema, & Mateu, Jorge. 2015. *Spatial and spatio-temporal geostatistical modeling and kriging*.

[Montero & Vilar, 2014] Montero, Pablo, & Vilar, José. 2014. Tsclust : An r package for time series clustering. *Journal of statistical software*, **62**(11), 1–43.

[Oliver & Webster, 1989] Oliver, MA, & Webster, R. 1989. A geostatistical basis for spatial weighting in multivariate classification. *Mathematical geology*, **21**(1), 15–35.

[Oyelade *et al.*, 2010] Oyelade, O. J., Oladipupo, O. O., & Obagbuwa, I. C. 2010. *Application of k means clustering algorithm for prediction of students academic performance*.

[P.J. Diggle, 2007] P.J. Diggle, Paulo Justiniano Ribeiro. 2007. *Model-based geostatistics (springer series in statistics)*. 1 edn. Springer Series in Statistics. Springer.

[Pérez, 2018] Pérez, Andrés. 2018. Métodos avanzados de datos funcionales (tesis de maestría). *Universidad de cádiz*.

[Pérez-Ortega *et al.*, 2020] Pérez-Ortega, Joaquín, Almanza-Ortega, Nelva Nely, Vega-Villalobos, Andrea, Pazos-Rangel, Rodolfo, Zavala-Díaz, Crispín, & Martínez-Rebollar, Alicia. 2020. The k-means algorithm evolution. *Chap. 5 of:* Sud, Keshav, Erdogmus, Pakize, & Kadry, Seifedine (eds), *Introduction to data science and machine learning*. Rijeka: IntechOpen.

[Ramirez, 2015] Ramirez, Liliana. 2015. AutocorrelaciÓn espacial: AnalogÍas y diferencias entre el Índice de moran y el Índice getis y ord. 09.

[Ramsay & Silverman, 2001] Ramsay, J.O., & Silverman, B.W. 2001. Functional data analysis. *Pages 5822–5828 of:* Smelser, Neil J., & Baltes, Paul B. (eds), *International encyclopedia of the social behavioral sciences*. Oxford: Pergamon.

[Ramsay J., 2005] Ramsay J., Silverman B.W. 2005. *Functional data analysis*. 2ed edn. Springer Series in Statistics. Springer.

[Romano *et al.*, 2011] Romano, Elvira, Verde, Rosanna, & Cozza, Valentina. 2011. Clustering spatial functional data: A method based on a nonparametric variogram estimation. *Pages 339–346 of:* Ingrassia, Salvatore, Rocci, Roberto, & Vichi, Maurizio (eds), *New perspectives in statistical modeling and data analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg.

[Romano *et al.*, 2013] Romano, Elvira, Balzanella, Antonio, & Verde, Rosanna. 2013. *A regionalization method for spatial functional data based on variogram models: An application on environmental data*. Berlin, Heidelberg: Springer Berlin Heidelberg. Pages 99–108.

[Romano *et al.*, 2015] Romano, Elvira, Mateu, Jorge, & Giraldo, Ramon. 2015. On the performance of two clustering methods for spatial functional data. *AStA advances in statistical analysis*, **99**(4), 467–492.

[Romano *et al.*, 2017] Romano, Elvira, Balzanella, Antonio, & Verde, Rosanna. 2017. Spatial variability clustering for spatially dependent functional data. *Statistics and computing*, **27**(3), 645–658.

[Rui Xu, 2008] Rui Xu, Don Wunsch. 2008. *Clustering*. illustrated edition edn. IEEE Press Series on Computational Intelligence. Wiley-IEEE Press.

[Shekhar *et al.*, 2003] Shekhar, Shashi, Zhang, Pusheng, Huang, Yan, & Vatsavai, Ranga Raju. 2003. Trends in spatial data mining. *Data mining: Next generation challenges and future directions*, 357–380.

[Siabato & Guzmán-Manrique, 2019] Siabato, Willington, & Guzmán-Manrique, Jhon. 2019. La autocorrelación espacial y el desarrollo de la geografía cuantitativa. *Cuadernos de geografía: Revista colombiana de geografía*, **28**(1), 1–22.

[Sumathi *et al.*, 2008] Sumathi, N, Geetha, R, & Bama, S Sathiya. 2008. Spatial data mining-techniques trends and its applications. *Journal of computer applications*, **1**(4), 28–30.

[Tarpey & Kinateder, 2003] Tarpey, Thaddeus, & Kinateder, Kimberly. 2003. Clustering functional data. *J. classification*, **20**(05), 093–114.

[Tian, 2010] Tian, Tian Siva. 2010. Functional data analysis in brain imaging studies. *Frontiers in psychology*, **1**, 35.

[Tobler, 1970] Tobler, W. R. 1970. A computer movie simulating urban growth in the detroit region. *Economic geography*, **46**, 234–240.

[Wackernagel, 2003] Wackernagel, Hans. 2003. *Multivariate geostatistics: An introduction with applications*. 3 edn. Springer-Verlag Berlin Heidelberg.

[Wang *et al.*, 2016] Wang, Jane-Ling, Chiou, Jeng-Min, & Müller, Hans-Georg. 2016. Functional data analysis. *Annual review of statistics and its application*, **3**(1), 257–295.

[Yamamoto, 2012] Yamamoto, Michio. 2012. Clustering of functional data in a low-dimensional subspace. *Advances in data analysis and classification*, **6**(3), 219–247.

[Young, 2014] Young, G. Alastair. 2014. Inference for functional data with applications by lajos horváth and piotr kokoszka. *International statistical review*, **82**(1), 155–156.

[Zapata-Rios *et al.*, 2021] Zapata-Rios, Xavier, Lopez-Fabara, Carmen, Navarrete, Abigail, Torres, Sandra, & Flores, Miguel. 2021. Spatiotemporal patterns of burned areas, fire drivers, and fire probability across the equatorial andes. *Journal of mountain science*, **18**(04), 952–972.

[Zhang, 2013] Zhang, Jin-Ting. 2013. *Analysis of variance for functional data*. Chapman Hall/CRC Monographs on Statistics Applied Probability. Chapman Hall / CRC Press.

Apéndice A

Resultados adicionales

A.1. Caso de simulación: 8 grupos

A continuación se presenta de manera gráfica los resultados obtenidos al aumentar el número de grupos, teniendo en cuenta que los datos se simularon siguiendo el modelo 4.1.

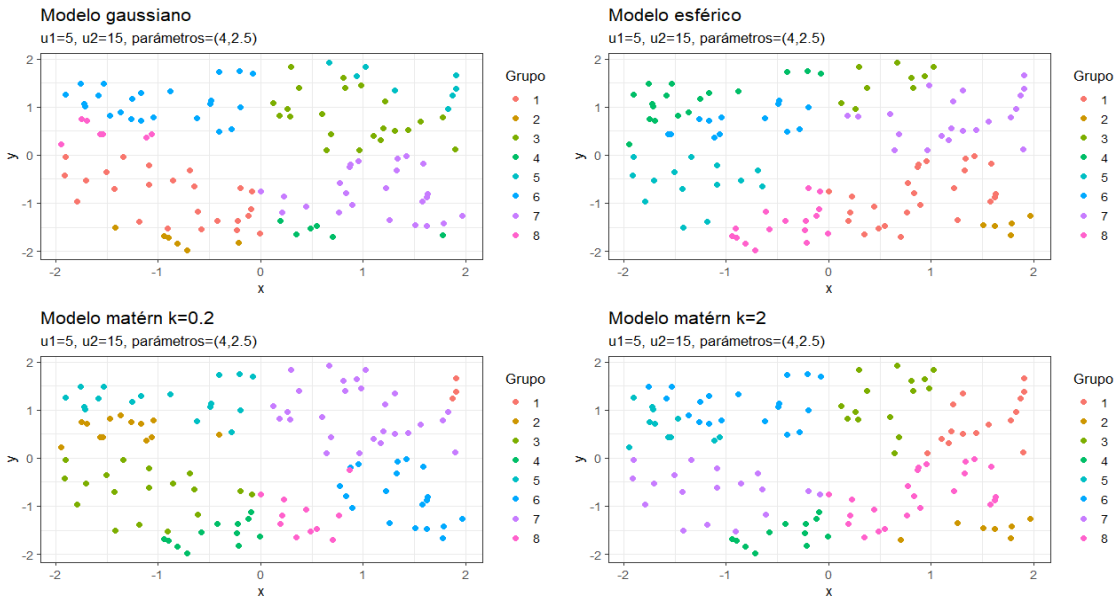


Figura A.1: Clasificación WTV: media funcional $\mu_1(t) = 5,\ \mu_2(t) = 15$.

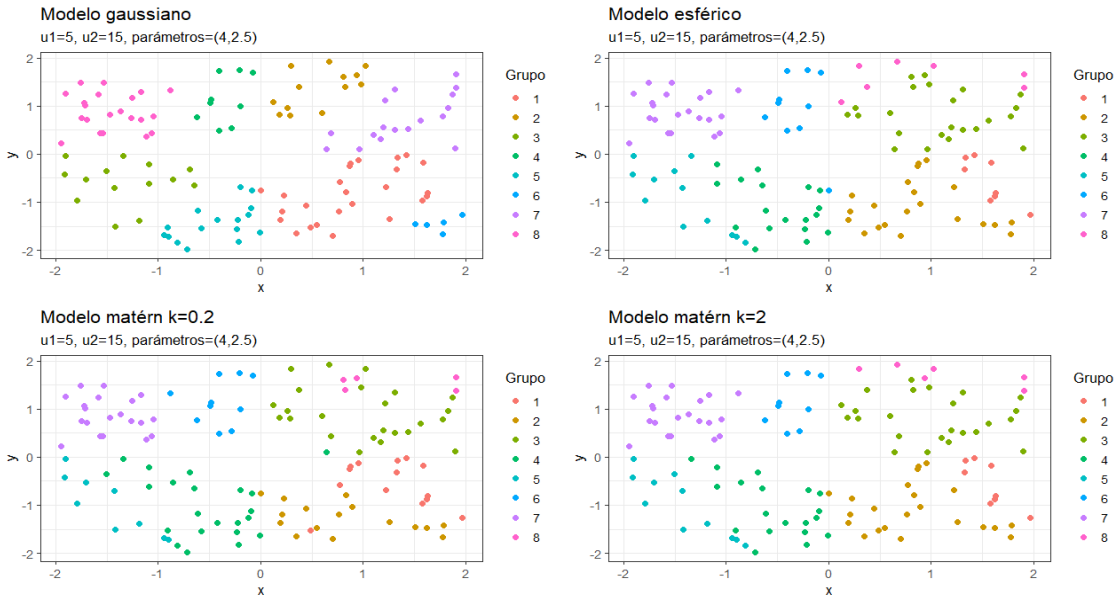


Figura A.2: Clasificación WMV: media funcional $\mu_1(t) = 5,\ \mu_2(t) = 15$.

En las figuras anteriores se puede apreciar que el algoritmo k-medias modificado logra clasificar los grupos adicionales dentro de cada cuadrante. Este comportamiento se debe a la manera en que los datos son generados, pues los grupos existentes dentro de cada cuadrante tienen la misma media funcional.

A.2. Caso de simulación: $\sigma^2 = 6$

A continuación se presentan los resultados obtenidos para el caso de simulación considerando un mayor valor para el parámetro σ^2.

- **Escenario 1:** $\sigma^2 = 6$, $\phi = 1{,}8$

Modelo	Método	$\mu_1(t) = 5$ $\mu_2(t) = 6$	$\mu_1(t) = 5$ $\mu_2(t) = 10$	$\mu_1(t) = 5$ $\mu_2(t) = 15$
Gaussiano	WTV	91,36	96,34	96,55
	WMV	92,16	95,76	96,42
Esférico	WTV	78,41	91,26	91,59
	WMV	81,57	92,09	92,76
Matérn	WTV	86,70	93,15	92,08
$\kappa = 0,2$	WMV	89,18	91,89	88,01
Matérn	WTV	98,04	98,75	98,86
$\kappa = 2$	WMV	98,09	96,55	91,74

Cuadro A.1: PCC: Escenario 1.

- **Escenario 2:** $\sigma^2 = 6$, $\phi = 2{,}5$

Modelo	Método	$\mu_1(t) = 5$ $\mu_2(t) = 6$	$\mu_1(t) = 5$ $\mu_2(t) = 10$	$\mu_1(t) = 5$ $\mu_2(t) = 15$
Gaussiano	WTV	95,44	98,29	98,77
	WMV	96,01	98,20	98,75
Esférico	WTV	86,34	94,82	95,13
	WMV	88,46	94,45	93,90
Matérn	WTV	89,19	93,29	92,18
$\kappa = 0,2$	WMV	92,01	93,44	87,65
Matérn	WTV	98,98	99,47	99,25
$\kappa = 2$	WMV	99,00	96,31	86,96

Cuadro A.2: PCC: Escenario 2.

En las tablas, A.1 y A.2, se evidencia un comportamiento similar al que se obtuvo en las tablas de la sección 4.1 en cuanto a los valores de PCC, a pesar de que se haya incrementado el valor del parámetro σ^2. Esto ocurre debido a que el parámetro σ^2 afecta la variabilidad de las curvas; bajo los escenarios propuestos y la metodología planteada la clasificación es obtenida considerando principalmente la dependencia espacial.

Apéndice B

ANEXOS

B.1. Código del algoritmo k-medias funcional espacial

B.1.1. Librerías

```
library(sp)
library(operators)
library(fda.usc)
library(mvtnorm)
library(sp)
library(gstat)
library(geoR)
library(ggplot2)
library(plyr)
library(parallel)
library(data.table)
library(geofd)
library(TSclust)
library(rgl)
library(fdANOVA)
library(npsp)
library(rnaturalearth)
library(rnaturalearthdata)
library(rgeos)
library(plotly)
library(fdaoutlier)
library(lattice)
library(tidyr)
library(ggpubr)
library(grid)
library(gridExtra)
library(tidyverse)
library(patchwork)
```

B.1.2. Distancia L^2

```
Ml2_dist<-function(data,nbasis=65){
  nbasis<-nbasis
  n <-  dim(data)[1]
  s <-  dim(data)[2]
  argvals<-seq(1,n, by=1)
  range  <- c(1,n)
  period <- n
  basis <- create.fourier.basis(range, nbasis, period)
  datafd <- Data2fd(data,argvals,basis)

  L2norm<-matrix(0,nrow=s,ncol=s)
  coef<-datafd$coef
  M<-fourierpen(basis,Lfdobj=0)
  for (i in 1:(s-1)){
    coef.i<-coef[,i]
    for (j in (i+1):s){
      coef.j<-coef[,j]
      L2norm[i,j]<-t(coef.i-coef.j)%*%M%*%(coef.i-coef.j)
      L2norm[j,i]<-L2norm[i,j]
    }
  }

  names<-names(data.frame(data))
  dimnames(L2norm)<-list(names,names)
  return(L2norm)
}
```

B.1.3. Asignación de centroides iniciales

```
kmeans.center.iniL2<-function (fdataobj,Vmdist=FALSE,coord=NULL, ncl = 2,
                               cov.model="spherical", Kappa=NULL,
                               multivgm=multivgm, method = "exact",
                               max.iter = 10000, max.comb = 1e+06,
                               par.metric = NULL, ...){
  if (!is.fdata(fdataobj))
    fdataobj = fdata(fdataobj)
  if (is.null(par.metric))
    par.metric = list()
  par.metric$fdata1 <- fdataobj
  z <- fdataobj[["data"]]
  if(!isTRUE(Vmdist) && is.null(coord)){
    mdist<-Ml2_dist(t(z),nbasis = 65)
  }
  else{
    mdist<-MVari(t(z),coord,nugget.fix=NULL, max.dist.variogram=NULL,
                 cov.model = cov.model,Kappa = Kappa,multivgm = multivgm)
    print("entro al variograma con el cov.model")
```

```
  print(cov.model)
}

tt <- fdataobj[["argvals"]]
rtt <- fdataobj[["rangeval"]]
names <- fdataobj[["names"]]
nr = nrow(fdataobj)
nc = ncol(fdataobj)
if (is.vector(ncl)) {
  len.ncl = length(ncl)
  ngroups = ncl
  max.combn <- choose(nr, ngroups)
  if (len.ncl == 1) {
    ind = 1
    if (method == "exact") {
      if (max.combn > max.comb)
        warning(paste0(max.combn, " samples are required,
        it has been limited to a random sample of size ",
                       max.comb))
      method = "sample"
    }
    if (method == "sample") {
      max.iter <- min(max.iter, max.combn)
      vec <- array(NA, dim = c(ngroups, max.iter))
      vec.d <- rep(NA, nc)
      for (i in 1:max.iter) {
        vec[, i] <- sample(1:nr, ngroups, replace = FALSE)
        vec.d[i] <- sum(mdist[vec[, i], vec[, i]])
      }
      ind.max <- which.max(vec.d)
      lxm <- vec[, ind.max]
    }
    else if (method == "exact") {
      co <- combn(1:nr, ngroups)
      nco <- ncol(co)
      vec <- rep(NA, nco)
      for (i in 1:nco) {
        vec[i] <- sum(mdist[co[, i], co[, i]])
      }
      max.vec <- which.max(vec)
      lxm <- co[, max.vec]
    }
    else stop("Center initialization method unknown")
    xm = z[lxm, ]
  }
  else stop("Argument 'ncl' is expected the number of groups to detect")
}
else stop("Argument 'ncl' is expected the number of groups to detect")
```

```
  d = rbind(mdist, mdist[lxm, ])
  centers = fdata(xm, tt, rtt, names)
  out <- list(centers = centers, lcenters = lxm, z.dist = mdist,
              fdataobj = fdataobj)
  class(out) <- "kmeans.fd"
  return(invisible(out))
}
```

B.1.4. Cálculo de matriz de distancia ponderada

```
MVari<-function(data,coord,nugget.fix=NULL, max.dist.variogram=NULL,
                multivgm=multivgm,nbasis=65,cov.model="spherical",
                Kappa=NULL){
  dia=1:length(data[,1])
  fourier.basis=create.fourier.basis(
    rangeval = range(dia), nbasis = nbasis
  )
  temp2fd=Data2fd(argvals = dia,
                  y=data,
                  basisobj = fourier.basis)

  M2d=dist(t(temp2fd$coefs))^2

  Eu.d <-as.matrix(dist(coord,method="euclidian"))
  if(isFALSE(multivgm)){

    if (is.null(max.dist.variogram))
      if(is.null(Kappa)){
      data=cbind(coord,value=apply(temp2fd$coefs,2,sum))
      coordinates(data) = ~x+y
      emp.trace.vari=variogram(value~1, data,dX=0,cutoff=3)
      plot(emp.trace.vari)
      sigma2.0=quantile(emp.trace.vari$gamma,0.75)
      phi.0=quantile(emp.trace.vari$dist,0.2)
      nt=mean(emp.trace.vari$gamma)/4
      trace.vari = fit.variogram(emp.trace.vari, vgm(sigma2.0, cov.model,
                                      phi.0,nt))
      print("con vgm")
      g=plot(emp.trace.vari,trace.vari)
      show(g)
    }
    else{
      data=cbind(coord,value=apply(temp2fd$coefs,2,mean))
      coordinates(data) = ~x+y
      emp.trace.vari=variogram(value~1, data,dX=0,cutoff=3.5)
      plot(emp.trace.vari)
      sigma2.0=quantile(emp.trace.vari$gamma,0.75)
      phi.0=quantile(emp.trace.vari$dist,0.2)
```

```
      nt=mean(emp.trace.vari$gamma)/4
      trace.vari = fit.variogram(emp.trace.vari, vgm(psill =  sigma2.0,
                                                     model =  cov.model,
                                                     range = phi.0,
                                                     nugget = nt,
                                                     kappa = Kappa),
                                 fit.kappa = TRUE)
      g=plot(emp.trace.vari,trace.vari)
      show(g)
      print("matern con kappa")
    }
    sigma2=trace.vari$psill[2]
    nugget=trace.vari$psill[1]
    tra.vari.mat <- sigma2+nugget - cov.spatial(Eu.d,
                                        cov.model=tolower(trace.vari$model[2]),
                                        cov.pars=c(sigma2,trace.vari$range[2]),
                                        kappa=trace.vari$kappa[2])

    weig.mat<-sqrt(as.matrix(M2d))*tra.vari.mat
  }
  else{
    multvariogram=multiv(Eu.d,coord,temp2fd$coefs,cov.model,3,Kappa)

    L2norm<-M2d
    weig.mat<-sqrt(as.matrix(L2norm))*multvariogram
    print("matriz multivariograma")
  }
  return(weig.mat)
}
```

B.1.5. Actualización de centroides

```
kmeans.centers.update=function(out,group
                               ,dfunc=func.trim.FM,
                               ,par.dfunc=list(trim=0.05)
                               ,...){
  if (class(out)!="kmeans.fd")
    stop("Error: incorrect input data")
  z = out$fdataobj[["data"]]
  tt = out$fdataobj[["argvals"]]
  rtt <- out$fdataobj[["rangeval"]]
  names = out$fdataobj[["names"]]
  mdist = out$z.dist
  centers = out$centers
  xm = centers[["data"]]
  nr = nrow(z)
  nc = ncol(z)
  grupo = group
```

```
  ngroups = length(unique(group))
  d = out$d
  ncl = nrow(xm)
  for (j in 1:ngroups){
    jgrupo <- grupo==j
    dm=z[jgrupo,]
    ind=which(jgrupo)
    if (is.vector(dm) || nrow(dm)==1) {
      k=j
      stat=dm
    }
    else  {
      par.dfunc$fdataobj<-centers
      par.dfunc$fdataobj$data<-dm
      stat=do.call(dfunc,par.dfunc)
    }
    if (is.fdata(stat)) xm[j,]=stat[["data"]]
    else  xm[j,]=stat
  }
  centers$data=xm

  row.names(centers$data) <- paste("center ",1:ngroups,sep="")

  return(list("centers"=centers,"cluster"=grupo))
}
```

B.1.6. Asignación de curvas a los grupos

```
kmeans.assig.groups=function(out,...){
  if (!is.null(out$lcenters))
    lxm=out$lcenters else   lxm=NULL
    nr = nrow(out$fdataobj)
    nc = ncol(out$fdataobj)
    xm = out$centers[["data"]]

    grupo = rep(0,nr)
    d = out$d

    ngroups=nrow(d)-nrow(out$fdataobj[["data"]])

    grupo <- apply(d[(nr+1):(nr+ngroups),],2,which.min)
    return(list("centers"=out$centers,"cluster"=grupo))

}
```

B.1.7. k-medias funcional espacial

```
kmeans.fdas<-function (fdataobj, ncl = 2,Vmdist=FALSE,coord=NULL,
                       cov.model="spherical", Kappa=NULL,
                       multivgm=multivgm, dfunc = func.trim.FM,
                       max.iter = 10000, par.metric = NULL,
                       par.dfunc = list(trim = 0.05),
                       method = "sample", cluster.size = 1,...){
  if (!is.fdata(fdataobj))
    fdataobj = fdata(fdataobj)
  nas1 <- is.na(fdataobj)
  if (any(nas1))
    stop("fdataobj contain ", sum(nas1), " curves with some NA value \n")
  z <- fdataobj[["data"]]
  tt <- fdataobj[["argvals"]]
  rtt <- fdataobj[["rangeval"]]
  nr = nrow(z)
  nc = ncol(z)
  if (is.vector(ncl)) {
    len.ncl = length(ncl)
    if (len.ncl == 1) {
      par.ini <- list()
      par.ini$fdataobj = fdataobj
      par.ini$method = method
      par.ini$ncl = ncl
      par.ini$draw = draw
      par.ini$max.comb = 1e+06
      par.ini$max.iter = max.iter
      par.ini$Vmdist = Vmdist
      par.ini$coord = coord
      par.ini$cov.model=cov.model
      par.ini$Kappa=Kappa
      par.ini$multivgm=multivgm
      if (!is.null(par.metric))
        par.ini$par.metric <- par.metric
      par.ini$... <- par.metric
      out1 = do.call(kmeans.center.iniL2, par.ini)
      lxm <- out1$lcenters
      out1$d = rbind(out1$z.dist, out1$z.dist[lxm, ])
      print("entro al que debe")
    }
    else {
      ngroups <- length(ncl)
      lxm <- ncl
      xm <- z[lxm, ]
      out1 <- list()
      out1$fdataobj <- fdataobj
      out1$ncl <- len.ncl
```

```
    if (is.null(par.metric))
      par.metric <- list(p = 2, w = 1)
    par.metric$fdata1 <- fdataobj
    if(!isTRUE(Vmdist) && is.null(coord))      {
      mdist<-Ml2_dist(t(z))
    }
    else{
      mdist<-MVari(t(z),coord,nugget.fix=NULL,
      max.dist.variogram=NULL,cov.model = cov.model)
      print("entro al cov.model")
    }
    out1$z.dist <- mdist
    out1$d <- rbind(mdist, mdist[lxm, ])
    out1$centers <- fdataobj[ncl, ]
    out1$lcenters <- ncl
    class(out1) <- "kmeans.fd"
  }
}
else if (is.fdata(ncl)) {
  lxm = NULL
  xm = ncl[["data"]]
  if (is.null(par.metric))
    par.metric = list(p = 2, w = 1)
  par.metric$fdata1 <- fdataobj

  if(!isTRUE(Vmdist) && is.null(coord)){
    mdist<-Ml2_dist(z)
  }
  else{
    mdist<-MVari(t(z),coord,nugget.fix=NULL, max.dist.variogram=NULL)
  }
  par.metric2 <- par.metric
  par.metric2$fdata2 <- ncl

  if(!isTRUE(Vmdist) && is.null(coord)){
    mdist<-Ml2_dist(z)
  }
  else{
    mdist<-MVari(t(z),coord,nugget.fix=NULL, max.dist.variogram=NULL)
  }
  out1 = list()
  out1$fdataobj <- fdataobj
  out1$centers = ncl
  out1$lcenters <- NULL
  ngroups = nrow(ncl)
  ncl = nrow(ncl)
  out1$d = rbind(mdist, t(mdist2))
  class(out1) = "kmeans.fd"
```

```
  }
  ngroups = nrow(out1$centers[["data"]])
  a = 0
  aa <- i <- 1
  same_centers = FALSE
  if (is.null(colnames(out1$d))){
    cnames <- colnames(out1$d) <- 1:NCOL(out1$d)}
  else{ cnames <- colnames(out1$d)}
  while ((i < max.iter) && (!same_centers)) {
    iterar <- FALSE
    out3 = kmeans.assig.groups(out1, draw = draw)
    names(out3$cluster) <- cnames
    tab <- table(out3$cluster)
    imin <- which.min(tab)[1]
    if (cluster.size > tab[imin]) {
      warning(paste0(" One of the clusters only has ",
                     tab[imin], " curves and the minimum cluster size is ",
                     cluster.size, ".\n The cluster is completed with the
                     closest curves of the other clusters."))
      iclust <- out3$cluster == imin
      dist.aux <- out1$d[imin, ]
      icambios <- as.numeric(names(sort(out1$d[imin, ])[1:cluster.size]))
      out1$d[imin, icambios] <- 0
      out1$z.dist <- out1$d
      out3$cluster[icambios] <- imin
      out2 <- out3
      out1$cluster <- out3$cluster
      par.dfunc$fdataobj <- fdataobj[c(icambios)]
      out1$centers[imin] = do.call(dfunc, par.dfunc)
      iterar <- TRUE
      i = i + 1
      print("entro al out3")
    }
    out2 = kmeans.centers.update(out1, group = out3$cluster,
                                 dfunc = dfunc, par.dfunc = par.dfunc)
    if (!iterar) {
      same_centers <- out2$centers$data == out3$centers$data
      out1$centers <- out2$centers
      i = i + 1
    }
  }

  out <- list(cluster = out2$cluster, centers = out2$centers)
  return(out)
}
```

B.1.7.1. Cálculo del variograma multivariado

```
multiv<-function(Eu.d,coord,data,cov.model="Gau",max.dist=3,Kappa=NULL){
  k=dim(data)[1]
  n=nrow(coord)
  varis=matrix(0,nc=n,nr=n)
  for (i in 1:k) {
    dat=cbind(coord,value=data[i,])
    coordinates(dat) = ~x+y
    emp.trace.vari=variogram(value~1, dat,dX=0,cutoff=max.dist)

    sigma2.0=quantile(emp.trace.vari$gamma,0.75)
    phi.0=quantile(emp.trace.vari$dist,0.2)
    nt=mean(emp.trace.vari$gamma)/4

    if(is.null(Kappa)){
      trace.vari = fit.variogram(emp.trace.vari, vgm(sigma2.0, cov.model,
                                                     phi.0,nt))
    }
    else{
      trace.vari = fit.variogram(emp.trace.vari, vgm(psill = sigma2.0,model =
                                                     cov.model,range =  phi.0,
                                                     nugget = nt,kappa = Kappa),
                                 fit.kappa = TRUE)
    }
    sigma2=trace.vari$psill[2]
    nugget=trace.vari$psill[1]
    tra.vari.mat <- sigma2+nugget - cov.spatial(Eu.d,
                                        cov.model=tolower(trace.vari$model[2]),
                                        cov.pars=c(sigma2,trace.vari$range[2]),
                                        kappa=trace.vari$kappa[2])

    varis=varis+tra.vari.mat
  }

  return(varis)
}
```

B.2. Selección del número de grupos

B.2.1. Índice SSB

```
# a es el resultado de aplicar la funcion kmeans.fdas
SSBindex1<-function(a){
  centros<-t(a$centers$data)
  xmean<-func.mean(fdata(t(data))) ##estaba con t
  centros<-cbind(centros,t(xmean$data))
```

```
  distancias<-Ml2_dist(centros,nbasis = 65)
  njs<-data.frame(table(a$cluster))
  n1<-length(distancias[,1])
  dnj<-sum(distancias[n1,1:(n1-1)]*njs$Freq)/length(data)
  return(dnj)
}
```

B.2.2. Índice SSW

```
# a es el resultado de aplicar la funcion kmeans.fdas y data es la original
SSWindex<-function(a,data){
  s1<-c()
  for(i in 1:max(a$cluster)){
    b<-which(a$cluster==i)
    c<-data[,b]
    cj<-a$centers$data[i,]
    c<-cbind(c,(cj))

    distancias<-Ml2_dist(c,5)
    n1<-length(distancias[,1])
    dnj<-sum(distancias[n1,1:(n1-1)])
    s1[i]<-dnj
  }
  ss<-sum(s1)/length(data)
  return(ss)
}
```

B.2.3. Gráfico de los índices SSB y SSW

```
testf<-function(ncl=5){
  a<-kmeans.fdas(fdataobj = t(data),ncl = ncl,Vmdist = TRUE,coord = coord,
                 cov.model = cov.model,Kappa = Kappa,multivgm = multivgm)
  ssb<-SSBindex1(a)
  ssw<-SSWindex(a,(data))

  r=cbind(ssb,ssw)
  res=rbind(data.frame(),r)
}
##
data=t(data) #ubicaciones en filas y dias en columnas
cl <- parallel::makeCluster(detectCores())
 #functions.R contiene las funciones de seccion A.1.
clusterEvalQ(cl, source("functions.R"))
ncl=seq(2,20)
clusterExport(cl = cl, varlist = c("data","coord","cov.model",
                                   "Kappa","multivgm","ncl"))
```

```
system.time({
  results <- parallel::parLapply(cl,ncl,testf)
})

res=do.call(rbind,results)

parallel::stopCluster(cl)

ggplot(data.frame(x=c(2:20),y=res$ssb),aes(x=x,y=y))+
  geom_line(size=2,col="blue")+geom_point(size=4,col="red")+
  theme_bw()+
  scale_x_continuous(breaks = c(2:20),labels = c(2:20))+
  labs(x="Cluster",y="SSB Index",
       title = "SSB Index")+
  theme(axis.text=element_text(size=15),
        title = element_text(size=15,face='bold'))
```

B.3. Validación de grupos

B.3.1. Correlación temporal

```
corrtemporder1 <- function (x, y) {
  p <- length(x)
  sum((x[2:p] - x[1:(p - 1)]) *
        (y[2:p] - y[1:(p - 1)]))/(sqrt(sum((x[2:p] -x[1:(p - 1)])^2)) *
                                    sqrt(sum((y[2:p] - y[1:(p - 1)])^2)))
}
```

B.3.2. Correlación temporal entre centroide y miembros

```
# a es el resultado de aplicar la funcion kmeans.fdas
tcorc<-function(a){
  cort<-c()
  centros<-a$centers$data
  for (i in 1:length(centros[,1])) {
    b<-which(a$cluster==i)
    c<-data[,b]
    centros1<-t(c)
    corrc<-c()
    for (j in 1:length(centros1[,1])) {
      corrc[j]<-as.numeric(corrtemporder1(centros[i,],centros1[j,]))
    }

    cort[i]<-mean(corrc)
  }

  return(cort)
}
```

B.3.3. Correlación temporal entre miembros

```
dcorcurvas<-function(curvas){
  corr<-matrix(NA,nrow = length(curvas[,1]),ncol = length(curvas[,1]))
  for (i in 1:length(curvas[,1])) {
    for (j in 1:length(curvas[,1])) {
      corr[i,j]<-corrtemporder1(curvas[i,],curvas[j,])
    }
  }
  return(corr)
}

# a es el resultado de aplicar la funcion kmeans.fdas
correm<-function(a){
  res=c()
  for (i in 1:length(a$centers$data[,1])) {
    b<-which(a$cluster==i)
    c<-data[,b]
    curvas<-t(c)
    m=round(dcorcurvas(curvas),2)
    res[i]=mean(m)
  }

  return(res)
}
```

B.3.4. Índice de Moran

```
MoranGfda<-function(b,weig.mat,data,coord,tipo="entre"){
  centros<-b$centers
  n<-length(b$centers$data[,1])
  if(tipo=="entre"){
      xmean<-func.mean(fdata(t(data)))
      ss<-c()
      s2<-0
      for (i in 1:n) {
        s=0
        for (j in 1:n) {
          num<-(centros[i,]-xmean$data)*(centros[j,]-xmean$data)

          integral<-int.simpson(fdata(num$data))
          numerador<-weig.mat[i,j]*integral
          s=s+numerador
      }

        ss[i]<-s
        den<-(centros[i,]-xmean$data)^2
        integral2<-int.simpson(fdata(den$data))
        s2<-s2+integral2
```

```
      }
      IMoran<-(n*sum(ss))/(sum(sum(weig.mat))*s2)
      }
  else{
    IMoran<-c()
    for (i in 1:length(centros$data[,1])) {
      centroide<-centros$data[i,]
      miembros<-which(b$cluster==i)
      miembros<-data[,miembros]
      ss<-c()
      s2<-0
      #Adya
      coords<-coord[which(b$cluster==i),1:2]
      Eu.d <-as.matrix(dist(coords,method="euclidian"))
      weig.mat<-1/Eu.d
      diag(weig.mat)<-0

      for (j in 1:length(miembros[1,])) {
        s=0
        for (k in 1:length(miembros[1,])) {
          num<-(miembros[,j]-centroide)*(miembros[,k]-centroide)

          integral<-int.simpson(fdata(num))
          numerador<-weig.mat[j,k]*integral
          s=s+numerador

        }
          ss[j]<-s
          den<-(miembros[,j]-centroide)^2
          integral2<-int.simpson(fdata(den))
          s2<-s2+integral2
      }
      IMoran[i]<-(n*sum(ss))/s2
    }

  }


  return(IMoran)
}
```

B.3.5. Índice de Geary

```
GearyGfda<-function(b,weig.mat,data,coord,tipo="entre"){
  centros<-b$centers
  n<-length(b$centers$data[,1])
  if(tipo=="entre"){
    xmean<-func.mean(fdata(t(data)))
```

```
  ss<-c()
  s2<-0
  for (i in 1:n) {
    s=0
    for (j in 1:n) {
      num<-(centros[i,]-centros[j,])^2

      integral<-int.simpson(fdata(num$data))
      numerador<-weig.mat[i,j]*integral
      s=s+numerador
    }

    ss[i]<-s
    den<-(centros[i,]-xmean$data)^2
    integral2<-int.simpson(fdata(den$data))
    s2<-s2+integral2
  }
  IGeary<-((n-1)*sum(ss))/(2*sum(weig.mat)*s2)
}
else{
  IGeary<-c()
  sss<-c()
  for (i in 1:length(centros$data[,1])) {
    centroide<-centros$data[i,]
    miembros<-which(b$cluster==i)
    miembros<-data[,miembros]
    ss<-c()
    s2<-0
    #Adya
    coords<-coord[which(b$cluster==i),1:2]
    Eu.d <-as.matrix(dist(coords,method="euclidian"))
    weig.mat<-1/Eu.d
    diag(weig.mat)<-0

    for (j in 1:length(miembros[1,])) {
      s=0
      for (k in 1:length(miembros[1,])) {
        num<-(miembros[,j]- miembros[,k])^2

        integral<-int.simpson(fdata(num))
        numerador<-weig.mat[j,k]*integral
        s=s+numerador

      }
      ss[j]<-s
      den<-(miembros[,j]-centroide)^2
      integral2<-int.simpson(fdata(den))
      s2<-s2+integral2
```

```
      }
      sss[i]<-sum(ss)
      IGeary[i]<-((n-1)*sss[i])/(2*s2*sum(weig.mat))
    }

  }


  return(IGeary)
}
```

B.4. Simulación

B.4.1. Porcentaje de correcta clasificación

```
metod=function(datat,metodo="tvm"){

  metodo=ifelse(metodo=="tvm","mtv","tvm")
  nsin=unique(datat$simu)

  res=list()
  for (i in nsin) {
    z0=datat %>% dplyr::filter(simu==i) %>% dplyr::select(-metodo)
    names(z0)[1]="Grupo"

    z1=z0 %>%
      dplyr::group_by(simu,Grupo,cuad) %>% dplyr::summarise(n=n()) %>%
      mutate(perc=round(n/30*100,2)) %>%
      dplyr::group_by(simu,cuad) %>% dplyr::summarise(perc=max(perc))

    z1.1=z0 %>%
      dplyr::group_by(simu,Grupo,cuad) %>% dplyr::summarise(n=n()) %>%
      mutate(perc=round(n/30*100,2)) %>%
      dplyr::group_by(cuad)

    z1.2=merge(z1,z1.1,by=c("cuad","perc"))
    z1.2=z1.2[,-c(3,4)]
    z1.2=z1.2 %>% distinct(cuad,.keep_all = TRUE)

    z1.2.1=z1.2 %>% group_by(Grupo) %>% summarise(perc=max(perc))
    z1.2.1=merge(z1.2.1,z1.2,
    by=c("Grupo","perc")) %>% distinct(Grupo,.keep_all = TRUE)

    inde=z1.2.1$cuad
    indg=z1.2.1$Grupo

    ind1=which(1:4 %!in% inde)
    ind2=which(1:4 %!in% indg)
```

```
    if(sum(1:4 %in% z1.2$Grupo)<4){
      for (j in ind2) {
        w=z1.1 %>% dplyr::filter(Grupo==j)
        w=w[,-1]
        w=w %>% filter(cuad %in% ind1)
        ind1.1=which(w$perc==max(w$perc))
        w=w[ind1.1[1],]
        ind1=ind1[-which(ind1==w$cuad)]
        w=w[,c(2,4,1,3)]
        ind2.1=which(w$cuad==z1.2$cuad)
        z1.2[ind2.1,]=w
      }
      res[[i]]=z1.2[,1:2]
    }
    else{
      z1.2= z1.1 %>% dplyr::group_by(cuad) %>% dplyr::summarise(perc=max(perc))
      res[[i]]=z1.2
    }
  }

  return(res)
}
```

B.4.1.1. Generación de datos

```
Datsim=function(Eu.d,cov.model="Exp",cov.pars=c(0.5,1.5),Kappa=NULL,
                media=c(5,6)){
  if(is.null(Kappa)){
    cov.model=tolower(cov.model)
    Sigma1=cov.spatial(Eu.d,cov.model = cov.model,cov.pars = cov.pars)
    sim1=rmvnorm(365,mean=c(rep(media[1],30),rep(media[2],30),
                            rep(media[2],30),rep(media[1],30)), sigma=Sigma1)
  }
  else{
    cov.model=tolower(cov.model)
    Sigma1=cov.spatial(Eu.d,cov.model = cov.model,cov.pars = cov.pars,
                          kappa = Kappa)
    sim1=rmvnorm(365,mean=c(rep(media[1],30),rep(media[2],30),
                            rep(media[2],30),rep(media[1],30)), sigma=Sigma1)
  }
  return(sim1)
}
```

B.4.2. Proceso de simulación

```
# functions.R contiene las funciones de la sección A.2.
source("functions.R")
```

```
set.seed(123)
c1=c(runif(30,-2,-0.5),runif(30,-2,-0.5), runif(30,0.5,2), runif(30,0.5,2))
c2=c(runif(30,-2,-0.5),runif(30,0.5,2), runif(30,-2,-0.5), runif(30,0.5,2))

coord=data.frame(cbind(c1,c2))
names(coord)=c("x","y")
Eu.d <-as.matrix(dist(coord,method="euclidian"))

## generación de simulaciones

simus=function(Eu.d=NULL,nsim=10,cov.model="exponential",cov.pars=c(0.5,1.5),
               Kappa=NULL,media=c(5,15),coord=NULL){

  res=list()
  datalist=list()
  for (i in 1:nsim) {
      datalist[[i]]=t(Datsim(Eu.d,cov.model = cov.model,
      cov.pars = cov.pars,media = media,Kappa = Kappa))
  }
  cl <- parallel::makeCluster(detectCores())
  clusterEvalQ(cl, source("functions.R"))

  a<- parallel::parLapply(cl,datalist,kmeans.fdas,ncl = 4,Vmdist=TRUE,
                          coord=coord,cov.model=cov.model,Kappa=Kappa,
                          multivgm=FALSE)

  b<- parallel::parLapply(cl,datalist,kmeans.fdas,ncl = 4,Vmdist=TRUE,
                          coord=coord,cov.model=cov.model,Kappa=Kappa,
                          multivgm=TRUE)

  parallel::stopCluster(cl)
  for (i in 1:nsim) {
    c=cbind(tvm=a[[i]]$cluster,mtv=b[[i]]$cluster)
    c=data.frame(c)
    c$simu=i
    c$x=coord$c1
    c$y=coord$c2
    c$cuad=c(rep(3,30),rep(2,30),rep(4,30),rep(1,30))
    c=arrange(c,tvm)

    res[[i]]=c
  }
  datat=do.call(rbind,res)

  z1=metod(datat,metodo = "tvm")
  z1=do.call(rbind,z1) %>% summarise(tvm=mean(perc))
  z2=metod(datat,metodo = "mtv")
```

```
  z2=do.call(rbind,z2) %>% summarise(mtv=mean(perc))

  return(data.frame(z1,z2))
}

system.time(
  k1<- simus(Eu.d,nsim = 100,cov.model = "Mat",cov.pars = c(4,2.5),
             Kappa = 2,media = c(5,15),coord = coord)
)
k1
```

B.5. Aplicación

```
# carga de datos
load("./1_Datos_MF/NDVI_cluster_2.RData")
load("./1_Datos_MF/Ubicacion_NDVI_cluster.RData")

data=Data_cluster
coord=ubicacion

ubicacionrnd<-ubicacion
ubicacionrnd$ind=row.names(ubicacionrnd)
Data_cluster$ind=row.names(Data_cluster)

# emparejamiento de coordenadas y curvas
data1=merge(ubicacionrnd,Data_cluster,by="ind")
data1$ind=as.numeric(data1$ind)

# selección de datos y coordenadas
data=data1[,4:368]
coord=data1[,2:3]

set.seed(666)
data2=data1[sample(nrow(data1),10000,replace=FALSE),]

# base provisional
data=data2[,4:368]
coord=data2[,2:3]

# depuración de atípicos
data=t(data)
out=msplot(t(data))

# base final
data=data[,-out$outliers]
coord=coord[-out$outliers,]
```

```
# verificación de tendencia
datap=cbind(coord,valor=apply(data, 2,sum))
plot(as.geodata(datap),trend="cte")

# verificación de isotropía
emp.trace.vari1<-variog(coords=coord, data=apply(data,2,sum),
                        option="bin",message=FALSE,dir=0,
                        tolerance = pi/8,trend = "2nd",max.dist = 2)
emp.trace.vari1=data.frame(u=emp.trace.vari1$u,v=emp.trace.vari1$v)

emp.trace.vari2<-variog(coords=coord, data=apply(data,2,sum),
                        option="bin",message=FALSE,dir=pi/2,
                        tolerance = pi/8,trend="2nd",max.dist = 2)
emp.trace.vari2=data.frame(u=emp.trace.vari2$u,v=emp.trace.vari2$v)

emp.trace.vari3<-variog(coords=coord, data=apply(data,2,sum),
                        option="bin",message=FALSE,dir=pi/4,
                        tolerance = pi/8,trend="2nd",max.dist = 2)
emp.trace.vari3=data.frame(u=emp.trace.vari2$u,v=emp.trace.vari2$v)

emp.trace.vari4<-variog(coords=coord, data=apply(data,2,sum),
                        option="bin",message=FALSE,dir=3*pi/4,
                        tolerance = pi/8,trend="2nd",max.dist = 2)
emp.trace.vari4=data.frame(u=emp.trace.vari3$u,v=emp.trace.vari3$v)

# gráficos
iso1=ggplot(emp.trace.vari1)+
  geom_point(aes(x=u,y=v),size=3)+
  ylim(0,450)+theme_bw()+
  theme(axis.text=element_text(size=15),
        title = element_text(size=15,face='bold'))+
  labs(title="Ángulo=0°",x="Distancia",y="Varianza")

iso2=ggplot(emp.trace.vari2)+
  geom_point(aes(x=u,y=v),size=3)+
  ylim(0,450)+theme_bw()+
  theme(axis.text=element_text(size=15),
        title = element_text(size=15,face='bold'))+
  labs(title="Ángulo=45°",x="Distancia",y="Varianza")

iso3=ggplot(emp.trace.vari3)+
  geom_point(aes(x=u,y=v),size=3)+
  ylim(0,450)+theme_bw()+
  theme(axis.text=element_text(size=15),
        title = element_text(size=15,face='bold'))+
  labs(title="Ángulo=90°",x="Distancia",y="Varianza")

iso4=ggplot(emp.trace.vari4)+
```

```
  geom_point(aes(x=u,y=v),size=3)+
  ylim(0,450)+theme_bw()+
  theme(axis.text=element_text(size=15),
        title = element_text(size=15,face='bold'))+
  labs(title="Ángulo=135°",x="Distancia",y="Varianza")

grid.arrange(arrangeGrob(iso1,iso2,iso3,iso4,ncol=2,nrow = 2))

# estimación del variograma
emp.trace.vari<-variog(coords=coord, data=apply(data,2,sum),option="bin",
                       message=FALSE,trend = "2nd",max.dist = 2)
plot(emp.trace.vari)
eyefit(emp.trace.vari)

# aplicación del algoritmo k-medias funcional espacial
b=kmeans.fdas(fdataobj = t(data),ncl = 5,Vmdist = TRUE,coord = coord,
              cov.model = "powered.exponential",Kappa = 1.5,multivgm = FALSE)

# gráfico de los resultados sobre el mapa hidrográfico
shp=readOGR(dsn = "./data/ShapesRellenos/Shapefiles/
demarcacion_hidrografica_250k_2014_geo.shp")
spdf <- spTransform(shp, CRS("+proj=longlat +datum=WGS84"))
spdf<- tidy(spdf)
spdf=spdf %>% filter(long>=-81)

dh_simbol=as.character(shp@data[["dh_simbol"]])
dh_simbol=data.frame(dh_simbol,id=c(0:8))
spdf=merge(spdf,dh_simbol,by="id")

names(coord)[3]="Grupo"
names(spdf)[8]="Cuenca"

# data para nombres de cuencas
cuenca=c("DHESMERALDAS","DHGUAYAS","DHJUBONES","DHMANABÍ","DHMIRA","DHNAPO",
         "DHPASTAZA","DHPUYANGO-CATAMAYO","DHSANTIAGO")
x=c(-79.3,-79.8,-79.7,-80.3,-78,-77.5,-77,-79.7,-78)
y=c(0.5,-1.8,-3.3,-1,0.7,-0.5,-2,-4.1,-2.7)
dfc=data.frame(cuenca,x,y)

ggplot() +
  geom_point(data = coord,aes(x,y,color=Grupo),alpha=1,size=3)+
  geom_polygon(data = spdf, aes( x = long, y = lat, group = group,fill=Cuenca),
               alpha=0.2,color="black")+
  scale_fill_viridis_d(option = "turbo",direction = -1,begin = 0,end = 1)+
  labs(x="",y="",title = "Clasificación: Demarcación Hidrográfica")+theme_bw()+
  theme(axis.text=element_text(size=15),
        panel.grid = element_blank())+
  geom_label(data=dfc,aes(label=cuenca,x=x,y=y,fontface=2))+
```

```
  scale_y_continuous(breaks = c(2,1,0,-1,-2,-3,-4),labels = c(2,1,0,-1,-2,-3,-4))
```