ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

APLICACIONES DE MODELOS LINEALES DE CORREGIONALIZACIÓN

CLUSTERING PARA DATOS FUNCIONALES ESPACIALMENTE CORRELACIONADOS

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERA MATEMÁTICA

NATALY ALEXANDRA CUICHÁN ORTIZ

nataly.cuichan@epn.edu.ec

DIRECTOR: PEDRO MARTÍN MERINO ROSERO

pedro.merino@epn.edu.ec

DMQ, FEBRERO 2024

CERTIFICACIONES

Yo, NATALY ALEXANDRA CUICHÁN ORTIZ, declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

NATALY ALEXANDRA CUICHÁN ORTIZ

Certifico que el presente trabajo de integración curricular fue desarrollado por NATALY ALEXANDRA CUICHÁN ORTIZ, bajo mi supervisión.

PEDRO MARTÍN MERINO ROSERO
DIRECTOR

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el(los) producto(s) resultante(s) del mismo, es(son) público(s) y estará(n) a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

NATALY ALEXANDRA CUICHÁN ORTIZ

PEDRO MARTÍN MERINO ROSERO

RESUMEN

La clasificación de datos funcionales espacialmente correlacionados se aborda teniendo en cuenta las características específicas de este tipo de datos, lo que facilita la identificación de grupos espacialmente homogéneos en las ubicaciones. La metodología utilizada inicia suavizando los datos observados a través de bases de Fourier. Aprovechando las propiedades de estas bases y utilizando los coeficientes resultantes, se calcula la matriz de disimilitud entre las curvas de datos, incorporándola en el cálculo tanto del semivariograma como del variograma multivariado. Estos resultados se utilizan como insumos para las metodologías que se compararán.

El análisis se centra en ocho estaciones climatológicas de Quito, considerando las variables de temperatura promedio y precipitación total medidas durante las 24 horas de cada día del año 2023. Además, se toman las variables reales registradas en las estaciones climatológicas publicadas en la página de la Secretaría de Ambiente de la ciudad y el resultado de las variables simuladas en el proyecto METEO. Estas variables se compararán mediante tres metodologías de agrupamiento. Este enfoque no solo aborda la clasificación de datos funcionales espacialmente correlacionados, sino que compara las metodologías propuestas con las tradicionales.

Palabras clave: datos funcionales, correlación espacial, bases de Fourier, variogramas, clustering.

ABSTRACT

The classification of spatially correlated functional data is addressed by taking into account the specific characteristics of this type of data, facilitating the identification of spatially homogeneous groups in the locations. The methodology employed begins by smoothing the observed data through Fourier bases. Leveraging the properties of these bases and using the resulting coefficients, the dissimilarity matrix between the data curves is calculated, incorporating it into the computation of both the semivariogram and multivariate variogram. These results serve as inputs for the methodologies that will be compared.

The analysis focuses on eight climatological stations in Quito, considering the variables of average temperature and total precipitation measured over the 24 hours of each day in the year 2023. Additionally, real variables recorded at the climatological stations published on the website of the Environmental Secretary of the city, and the simulated variables from the METEO project, are considered. These variables will be compared using three clustering methodologies. This approach not only addresses the classification of spatially correlated functional data but also compares the proposed methodologies with traditional ones.

Keywords: functional data, spatial correlation, Fourier bases, variograms, clustering.

Índice general

Índice de figuras

Capítulo 1

Descripción del componente desarrollado

1.1. Objetivo general

Diseñar un modelo de correlación espacial para establecer agrupamientos de las variables simuladas en el proyecto METEO [1]

1.2. Objetivos específicos

1. Comparar los resultados obtenidos al aplicar distintos procedimientos de agrupamiento jerárquico para datos funcionales espacialmente correlacionados.

2. Determinar un método adecuado de agrupación considerando datos funcionales espacialmente correlacionados.

[1] El proyecto METEO desarrollado en el Centro de Modelización Matemática MODEMAT, es un sistema de predicción del tiempo que combina la ciencia de la meteorología con el modelo computacional Weather Research & Forecasting Model (WRF), para brindar pronósticos meteorológicos precisos y actualizados en el territorio ecuatoriano https://modemat.epn.edu.ec/es/divulgacion/meteo.

1.3. Alcance

Entender y aplicar técnicas de clustering para datos funcionales espacialmente correlacionados en los datos climáticos de la ciudad de Quito. Este trabajo aplica métodos sofisticados de análisis espacial y clasificación a los conjuntos de datos climáticos recopilados en la región. Como referencia principal, empleamos los métodos y técnicas de [5], asegurándonos de incorporar las particularidades geográficas y climáticas de la zona de Quito.

Para abordar la correlación espacial, primero se extraerán y procesarán los datos simulados desde enero hasta diciembre 2023 generados por el modelo WRF. Consideramos aspectos geográficos como la ubicación. Luego, se utilizarán métodos de clasificación para encontrar patrones climáticos similares en lugares específicos, relacionados con las estaciones de observación meteorológica actuales.

1.4. Marco teórico

1.4.1. Datos funcionales

Los datos funcionales son los conjuntos de observaciones en los que cada elemento de datos es una función completa en lugar de un solo valor. Estos datos muestran las relaciones continuas entre las variables a lo largo de una dimensión específica, como el tiempo o el espacio, en lugar de representar puntos discretos.

En vez de tener una serie de puntos en una línea de tiempo, se tiene toda la curva que describe el comportamiento a lo largo del tiempo. Esto es común en campos como la estadística funcional, que estudia fenómenos cambiantes. Las series temporales de funciones, como patrones de crecimiento, temperatura o precipitación a lo largo del tiempo son ejemplos prácticos [6].

1.4.2. Correlación espacial y Variogramas

La medida de cómo se relacionan o dependen las variaciones espaciales de dos o más variables dentro de un área geográfica específica se conoce como correlación espacial. En otras palabras, la correlación espacial establece cómo las variaciones de una variable en un entorno geográfico determinado están relacionadas espacialmente con las variaciones de otra variable.

En estadística espacial y geoestadística, el *variograma* es una herramienta utilizada para analizar la variabilidad espacial de un fenómeno en función de las distancias entre diferentes ubicaciones. Más precisamente, es una función que describe la medida en que un campo aleatorio depende del espacio. De acuerdo con la definición teórica de varianza, que se basa en el valor esperado de una variable aleatoria, tenemos:

$$\begin{aligned} 2\gamma(h) &= \mathbb{V}(Z(x+h) - Z(x)) \\ &= \mathbb{E}((Z(x+h) - Z(x))^2) - (\mathbb{E}(Z(x+h) - Z(x)))^2 \\ &= \mathbb{E}((Z(x+h) - Z(x))^2) \end{aligned}$$

Donde:

- $Z(x)$ es el valor de la variable en un sitio x.
- $Z(x+h)$ es el valor muestral separado del anterior por la distancia h.

Las propiedades de dependencia espacial en el proceso se describen por la función de semivarianza, que es la mitad del variograma $\gamma(h)$. La función de semivarianza se calcula utilizando el método de momentos cuando se tiene una instancia del fenómeno. El semivariograma experimental, se calcula de esta manera:

$$\overline{\gamma}(h) = \frac{\sum (Z(x+h) - Z(x))^2}{2n}$$

Donde:

- n es el número de parejas que se encuentran separadas por h.

En la práctica, por la irregularidad en el muestreo y en la distancia de los sitios, se establecen los intervalos de distancia $\{[0,h],(h,2h],(2h,3h],...\}$, y el semivariograma experimental se calcula para cada intervalo. La intención es comprender cómo cambia la semivarianza a medida que las ubicaciones de muestreo se encuentran a distancias diferentes [3].

En términos simples, un variograma muestra cómo los valores de una variable específica cambian en función de la distancia entre diferentes lugares. La "nube variográfica", que representa la semivarianza entre pares de puntos en relación con la distancia que los separa, se utiliza para realizar este análisis. Podemos deducir modelos teóricos de semivarianza de este gráfico.

Modelos teóricos de Semivarianza

El semivariograma experimental se calcula solo para distancias promedio específicas, pero proporciona información útil al medir la variabilidad o dispersión de los valores de una variable aleatoria con respecto a su media. Para abordar esto, es fundamental ajustar los modelos teóricos de semivarianza para que generalicen las relaciones observadas desde cualquier punto de vista.

Estos modelos, que pueden ser acotados (por ejemplo, lineal, logarítmico o potencial) o no acotados (por ejemplo, esférico, exponencial o gaussiano), permiten extender los datos del semivariograma experimental a todo el dominio espacial. Este método garantiza una comprensión sólida y una predicción precisa de la variabilidad espacial en ubicaciones no muestreadas. Cada uno de estos modelos tiene tres parámetros comunes, y la forma en que se elijan depende del tipo de variabilidad espacial presente en el área de estudio.

Los parámetros comunes son:

- **Efecto Pepita**: representa una discontinuidad temporal en el inicio. Esta discontinuidad puede ser el resultado de errores de medición en la variable o de la escala de la variable. En ocasiones, la presencia del efecto pepita puede indicar que una parte de la estructura espacial se concentra a distancias inferiores a las observadas, lo que

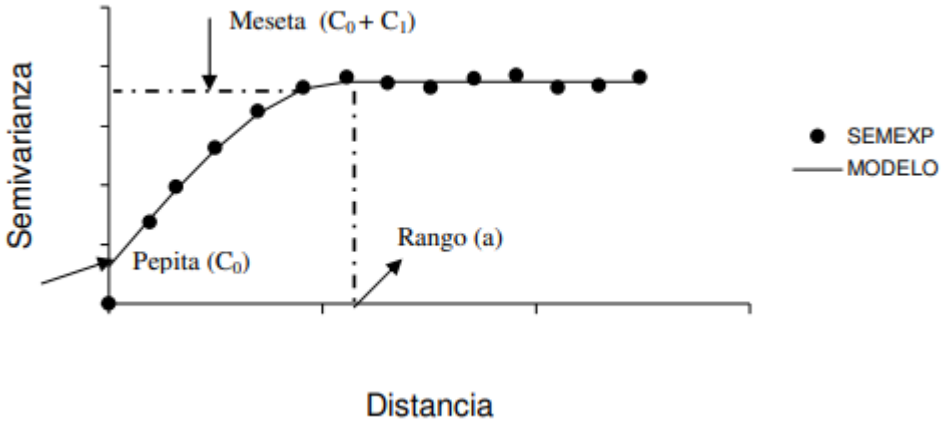

Figura 1.1: Comportamiento típico de un semivariograma acotado con una representación de los parámetros básicos. SEMEXP corresponde al semivariograma experimental y MODELO al ajuste de un modelo teórico [3].

indica la posibilidad de una variabilidad más significativa en escalas más pequeñas de las que se están considerando [3].

- **La Meseta**: cuando el efecto pepita es distinto de cero, representa la cima del semivariograma y se encuentra en el límite cuando la distancia h tiende al infinito. Una meseta infinita indica que la variabilidad sigue aumentando incluso a grandes distancias, mientras que una meseta finita indica que la variabilidad se mantiene constante a medida que aumenta la distancia [3].

  La meseta es importante porque se puede interpretar en relación con el efecto pepita, que implica errores de medición. El efecto pepita no debería superar el 50 % de la meseta en un modelo que refleje la realidad [3].

- **Rango**: representa la distancia a partir de la cual se considera que dos observaciones son independientes. Se cree que este rango es el área de influencia de la correlación espacial. Algunos modelos de semivariograma no tienen una distancia finita para la que dos observaciones sean completamente independientes. Para aplicar modelos de variograma en el análisis geoestadístico y comprender la dependencia espacial, es necesario comprender el rango [3].

1.4.3. Clustering para datos funcionales espacialmente correlacionados

El análisis de grupos se utiliza ampliamente en una variedad de campos. El agrupamiento jerárquico y la partición son los dos métodos más comunes para clasificar los elementos de una muestra.

El agrupamiento jerárquico se lleva a cabo en una serie de particiones, comenzando con un grupo que incluye a todos los individuos, lo que da como resultado n clústers, cada uno con un solo individuo [2].

Por otro lado, los métodos de partición como k-medias requieren que se especifique el número de clústers para asignar individuos a cada uno de ellos.

El objetivo del trabajo realizado en [5] es ampliar este concepto y utilizar una técnica que nos permita extender el método de agrupamiento clásico hacia el caso de datos funcionales espacialmente correlacionados.

Como primer alcance, se consideran datos funcionales para luego considerar su parte espacial con el objetivo de combinar estas técnicas que inicialmente se han considerado de forma independiente.

En este contexto, siguiendo los lineamientos de [5], suponemos que tenemos una muestra de curvas $X_1(t), ..., X_n(t)$ definidas para $t \in T = [a, b] \subseteq \mathbb{R}$ que pertenecen al espacio de Hilbert, estas son funciones medibles cuadrado integrable definidas en $[a, b]$, es decir:

$$L_2(T) = \{f : T \to \mathbb{R}, \, tal \, que \int_T f(t)^2 \, dt < \infty\}$$

Además, asumimos que estas funciones se pueden representar mediante funciones base de un espacio de dimensión finita, subespacio de $L^2(T)$. Esta representación, tiene la forma:

$$X_i(t) = \sum_{l=1}^{k} a_{il} B_l(t) = a_i^T B(t) \qquad i = 1, ..., n \tag{1.1}$$

Consideramos la norma L_2 entre las curvas $X_i(t)$ y $X_j(t)$:

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

Usando (1,1), con los vectores a_i y a_j coeficientes de las bases para los individuos ith y jth, obtenemos:

$$\begin{aligned} d_{ij} &= \sqrt{\int_{[a,b]} (a_i - a_j)^T B(t) B(t)^T (a_i - a_j) dt} \\ &= \sqrt{\int_{[a,b]} (a_i - a_j)^T W (a_i - a_j) dt}\,, \end{aligned} \tag{1.2}$$

donde:

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

En particular, para cualquier base ortonormal como las bases de Fourier, W corresponde a la matriz identidad. Para otras bases, W debe ser calculada usando algoritmos de integración numérica [5].

Ahora, para tomar en cuenta la estructura espacial de los datos, el agrupamiento permite encontrar clústers considerando sitios contiguos y variables similares. Para el caso descrito consideramos a $(Z(x) = (Z_1(x), ..., Z_m(x)) : x \in \mathbb{D})$ un proceso de m variables espaciales definidas sobre el dominio $\mathbb{D} \subseteq \mathbb{R}^d$.

Para el caso $m = 1$ ponderamos las similitudes (d_{ij}) entre las muestras con lo siguiente:

$$d_{ij}^w = d_{ij} \gamma(h) \tag{1.3}$$

Donde $\gamma(h)$ es el variograma calculado para las distancias entre los sitios i, j [4].

Para el caso $m > 1$ generalizamos la idea usando:

$$d_{ij}^{w} = d_{ij}\Gamma(h), \tag{1.4}$$

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

$$\Gamma(h) = \frac{1}{2}\mathbb{E}(Z(x) - Z(x+h))^{T} M (Z(x) - Z(x+h)). \tag{1.5}$$

Además, siendo M una matriz simétrica definida positiva (que usamos como métrica) y, considerando el caso particular en el que M es la matriz identidad, tenemos:

$$\begin{aligned} \Gamma(h) &= \sum_{l=1}^{m} \frac{1}{2}\mathbb{E}(Z_l(x) - Z_l(x+h))^2 \\ &= \sum_{l=1}^{m} \gamma_{ll}(h). \end{aligned} \tag{1.6}$$

Donde $\gamma_{ll}(h)$ es el variograma de la lth variable [7].

Con el objetivo de llegar al caso de datos funcionales espacialmente correlacionados, como en [5] se procede a unificar los métodos propuestos.

Se considera proceso estacionario aleatorio funcional $\{X_x(t),\ \ x \in \mathbb{D} \subseteq \mathbb{R}^d,\ \ t \in [a,b] \subseteq \mathbb{R}\}$ y una realización de este proceso aleatorio $X_1(t), ..., X_n(t)$ en n sitios con sus respectivas coordenadas $x_1, ..., x_n$, asumimos que estas curvas pertenecen al espacio de Hilbert $L^2([a,b])$.

De la ecuación $(1{,}2)$ notamos que la distancia entre dos curvas puede ser calculada por la distancia entre los coeficientes de las funciones base. De esta manera, el agrupamiento consiste en estimar los variogramas y semivariogramas de los coeficientes, usados para suavizar los datos observados.

Continuando con el método de [5], se asume que la curva para cada muestra en los lugares i, $i = 1, ..., n$, se puede representar por $(1{,}1)$. Los coeficientes de la matriz A están dados por:

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

Esta matriz nace de la realización de una K-variable aleatoria del conjunto $\{A(x) = (A_1(x), ..., A_k(x)) : x \in \mathbb{D} \subseteq \mathbb{R}^d\}$, con la matriz del variograma y crosvariograma de la forma:

$$\Gamma(h) = \begin{pmatrix} \gamma_{11}(h) & \gamma_{12}(h) & ... & \gamma_{1K}(h) \\ \gamma_{21}(h) & \gamma_{22}(h) & ... & \gamma_{2K}(h) \\ \vdots & \vdots & ... & \vdots \\ \gamma_{K1}(h) & \gamma_{K2}(h) & ... & \gamma_{KK}(h) \end{pmatrix}_{(KxK)} , \tag{1.7}$$

donde $\gamma_{lq}(h) = \frac{1}{2}\mathbb{E}(A_l(x_i) - A_q(x_j))^2$, $l, q = 1, ..., K$, $h = \left\| x_i - x_j \right\|$, se propone utilizar el Modelo Lineal de Corregionalización (MLC) para estimar la matriz $(1{,}7)$.

El MLC proporciona un método para modelar los variogramas de dos o más variables. Así, la varianza de cualquier posible combinación lineal de estas variables es siempre positiva. Por tanto, tomando la matriz $(1{,}7)$, se obtiene

$$\begin{aligned} \Gamma(h) &= \sum_{l=1}^{K} \frac{1}{2}\mathbb{E}(A_l(x) - A_l(x+h))^2 \\ &= \sum_{l=1}^{K} \gamma_{ll}(h). \end{aligned} \tag{1.8}$$

Así, llegamos a $(1{,}8)$ y $(1{,}6)$ que coinciden cuando el campo aleatorio multivariable $Z(x)$ es reemplazado por $A(x)$. La distancia espacial ponderada entre dos sitios viene dada por la Ecuación $(1{,}4)$ con d_{ij} calculada con la Ecuación $(1{,}2)$ y $\Gamma(h)$ obtenida a partir de la Ecuación $(1{,}8)$ [5].

Capítulo 2

Metodología

Aplicaremos la descrita en la sección anterior a dos grupos de datos meteorológicos. Para esto, seguiremos los siguientes pasos:

1. Recolección y limpieza de datos de las variables temperatura y precipitación.
2. Implementación de un código en lenguaje R para realizar los cálculos descritos en el marco teórico.
3. Suavizamiento de datos mediante bases de Fourier.
4. Cálculo del semivariograma.
5. Realizar el ajuste del modelo teórico a la nube variográfica.
6. Representar la estructura de agrupación mediante un dendograma y elegir el corte para formar los grupos.
7. Visualizamos e indicamos mediante un mapa los grupos obtenidos.

El primer grupo contiene los registros reales descargados de la página de la secretaría de ambiente del Municipio del Distrito Metropolitano de Quito https://ambiente.quito.gob.ec/, las variables consideradas son la precipitación y la temperatura registrados en la ciudad, esta información se da a nivel de hora por lo que tenemos 24 datos al día.

El segundo grupo de datos fue generado mediante el modelo WRF del proyecto METEO explicado previamente. En particular, tomamos las mismas variables para su comparación y que será descrito en los resultados.

Partiendo del primer grupo de datos reales, mostramos en el mapa las coordenadas de las estaciones climatológicas consideradas:

Figura 2.1: Ubicaciones geográficas de las estaciones climatológicas de Quito - Ecuador.

2.1. Datos Reales estaciones climatológicas

La información considerada en el primer grupo es la temperatura promedio diaria en los meses de enero 2023 hasta diciembre 2023 considerando las estaciones climatológicas San Antonio, Cotocollao, Carapungo, Belisario, Centro, El Camal, Guamaní, Los Chillos y Tumbaco.

2.1.1. Temperatura promedio diaria

Iniciamos suavizando los datos mediante 65 bases de Fourier, obteniendo:

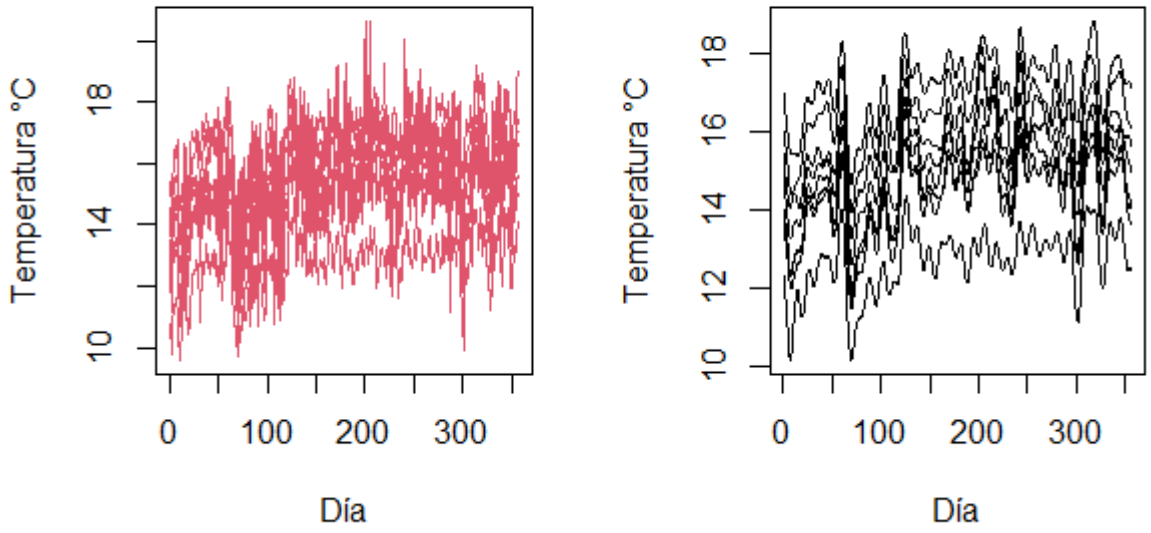

Figura 2.2: Izquierda: Gráfico temperatura promedio diaria. Derecha: Gráfico de curvas suavizadas temperatura promedio diaria.

Tomando en cuenta los datos de temperaturas promedio registradas en estaciones climatológicas, se procede a la selección de pares de ubicaciones dentro del conjunto de datos disponible. En cada par de ubicaciones, se realiza el cálculo de la diferencia entre los valores de la variable, posteriormente se agrupan los pares de ubicaciones según la distancia entre ellas. La siguiente etapa implica la representación gráfica de la semivarianza en función de la distancia, también conocida como lag, mediante la creación de un diagrama variográfico. Este diagrama proporciona una visualización clave de cómo varía la diferencia de temperaturas en relación con las distancias espaciales, permitiendo identificar patrones y estructuras de variabilidad que son fundamentales en el análisis espacial y geoestadístico.

Continuando con la metodología, definimos el modelo de variograma que se ajusta a nuestro caso, calculamos los parámetros: nugget igual a cero, la silla es 2784. Se puede ajustar un modelo matemático al variograma para obtener una descripción más suave de la variabilidad espacial, eligiendo según el gráfico el variograma esférico:

Se ajusta también el modelo lineal de corregionalización utilizando

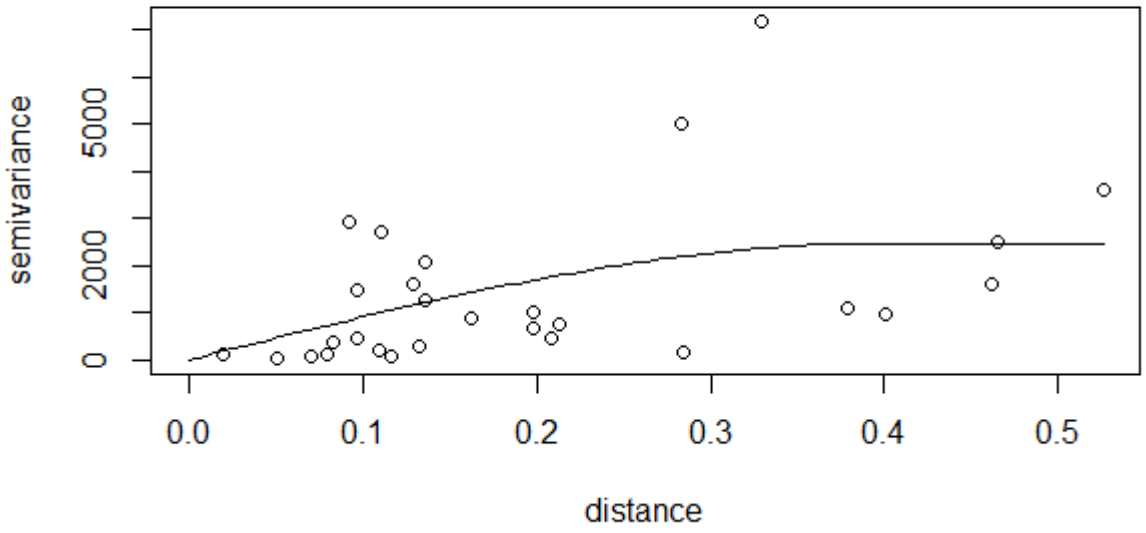

Figura 2.3: Modelo esférico ajustado a la nube variográfica para temperatura promedio real en Quito.

el campo aleatorio multivariable constituido por los coeficientes de las funciones de la base de Fourier utilizadas para suavizar los residuos. Con base al LMC ajustado, se estimó el variograma multivariado mediante la Ecuación (1,5).

Para visualizar las diferencias entre los métodos convencionales y los propuestos se realizaron tres análisis jerárquicos de clústeres para datos funcionales. El primero se fundamentó en la matriz de distancias euclidianas entre los coeficientes de las funciones de base de Fourier utilizadas para suavizar los datos de temperatura con (1,2), mientras que los restantes se llevaron a cabo al ponderar esta matriz de disimilitud con el variograma y el variograma multivariado obtenido con los coeficientes de la base de Fourier utilizados para suavizar las funciones. Así llegamos a los dendogramas en los que determinaremos las agrupaciones de las estaciones climatológicas y se analizarán en las conclusiones

2.1.2. Precipitación total diaria

La información considerada en el segundo grupo es la precipitación total diaria en los meses de octubre 2022 hasta octubre 2023, es decir que se ha sumado los datos para tener la totalidad de la precipitación diariamente, las mismas estaciones climatológicas que se consideran son las mismas. Realizamos el suavizamiento mediante bases de Fourier, obteniendo:

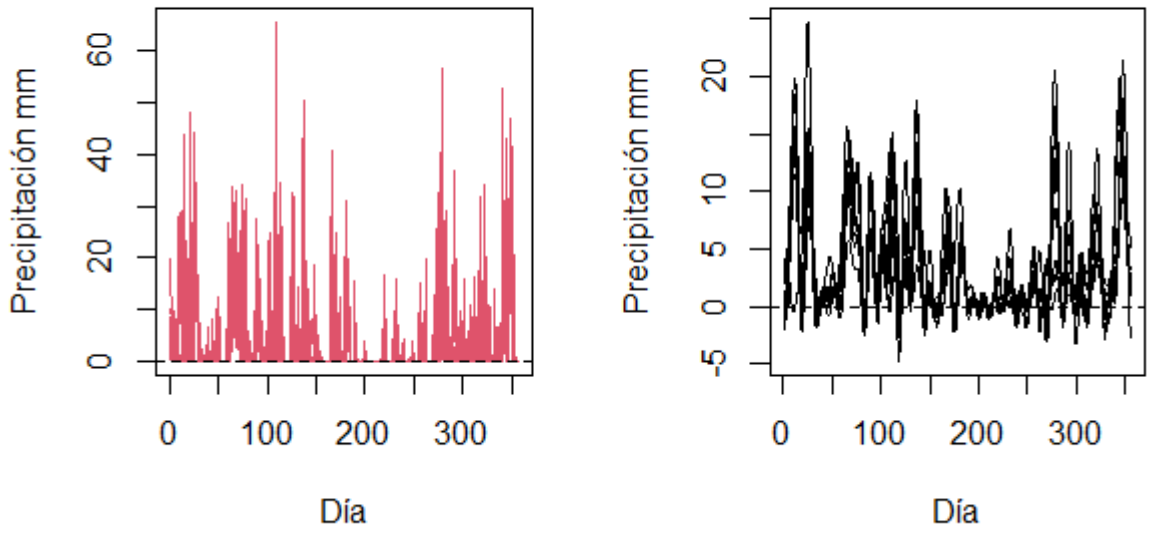

Figura 2.4: Izquierda: Gráfico precipitación total diaria. Derecha: Gráfico de curvas suavizadas precipitación total diaria

Procedemos a definir el variograma para determinar la variabilidad espacial de nuestros datos y obtenemos la siguiente nube variográfica:

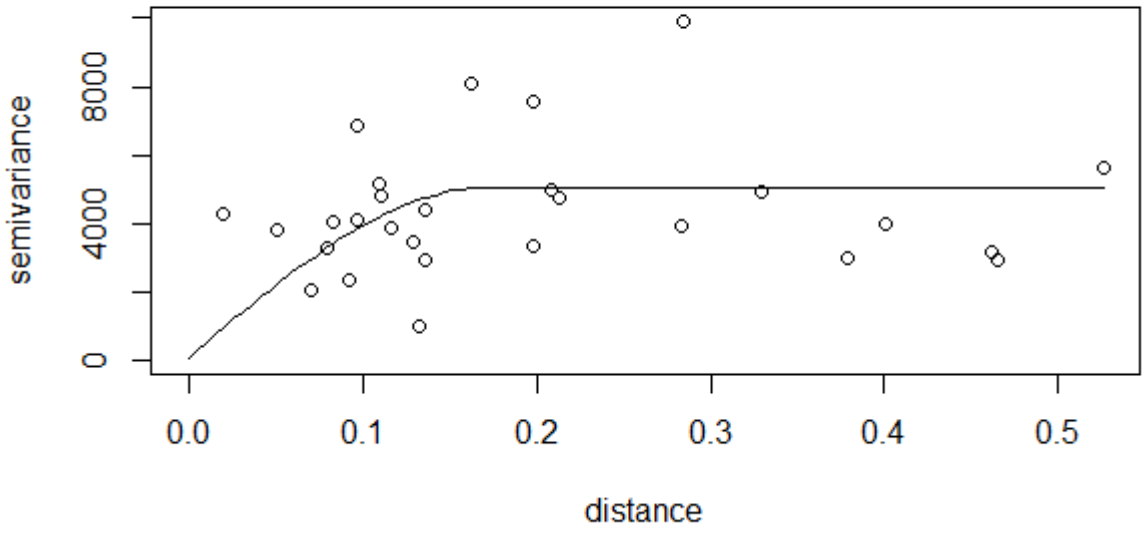

Figura 2.5: Modelo esférico ajustado a la nube variográfica para precipitación total diaria en Quito

Y mediante el MLC obtenemos uno de los resultados esperados y los comparamos con las dos opciones explicadas.

2.2. Datos simulados estaciones climatológicas

El sistema de pronóstico meteorológico para todo el territorio ecuatoriano (METEO) se fundamenta en una modelización matemática de la atmósfera y en el desarrollo de métodos computacionales avanzados para la resolución numérica del modelo correspondiente. El objetivo principal es generar pronósticos de precipitación, temperatura y nubosidad con un alto nivel de confiabilidad, centrándose en un período de 48 horas. Estos modelos incorporan características específicas del territorio ecuatoriano, como la vegetación, latitud y topografía, entre otras consideraciones.

En este proyecto, se emplean modelos especializados, como el modelo WRF (Weather Research and Forecast) y el modelo COSMO del sistema meteorológico alemán, para llevar a cabo las predicciones. Además, se ha dedicado esfuerzo a la visualización efectiva de la información pronosti-

cada. Los resultados del proyecto se detallan en cuatro secciones:

- **Mallado**: Describe la estructura de la malla utilizada en la resolución numérica de las ecuaciones de la atmósfera, que es esencial para obtener precisión de los pronósticos.
- **Asimilación de datos**: Enfatiza la importancia de integrar datos observacionales en tiempo real en el modelo para mejorar la calidad de las predicciones.
- **Post-procesamiento**: Detalla los procedimientos realizados después de la obtención de los resultados brutos del modelo para mejorar su interpretación y utilidad.
- **Laboratorio Nacional de Cálculo Científico**: El modelo WRF se resuelve dos veces al día en el supercomputador del Laboratorio Nacional de Cálculo Científico en la implementación y desarrollo del sistema. Por la gran cantidad de variables y el tipo de mallado, es necesario que el modelo WRF sea resuelto con técnicas de paralelización y supercómputo.

En conjunto, este enfoque integrado y detallado busca proporcionar pronósticos meteorológicos confiables y específicos para las condiciones ecuatorianas [1]. Es así que el segundo grupo considera la temperatura y la precipitación, simuladas mediante lo descrito, en los meses de enero 2023 hasta diciembre 2023 considerando las 24 horas para cada día del año en las estaciones climatológicas San Antonio, Cotocollao, Carapungo, Belisario, Centro, El Camal, Guamaní, Los Chillos y Tumbaco.

2.2.1. Temperatura promedio diaria

La data simulada de temperatura se promedia por día e ingresa al proceso para suavizar los puntos, el proceso para suavizarlos es expresarlo mediante bases de fourier del que requerimos los coeficientes. Graficamos las funciones suavizadas:

Procedemos a definir el variograma para determinar la variabilidad espacial de nuestros datos y obtenemos la siguiente nube variográfica:

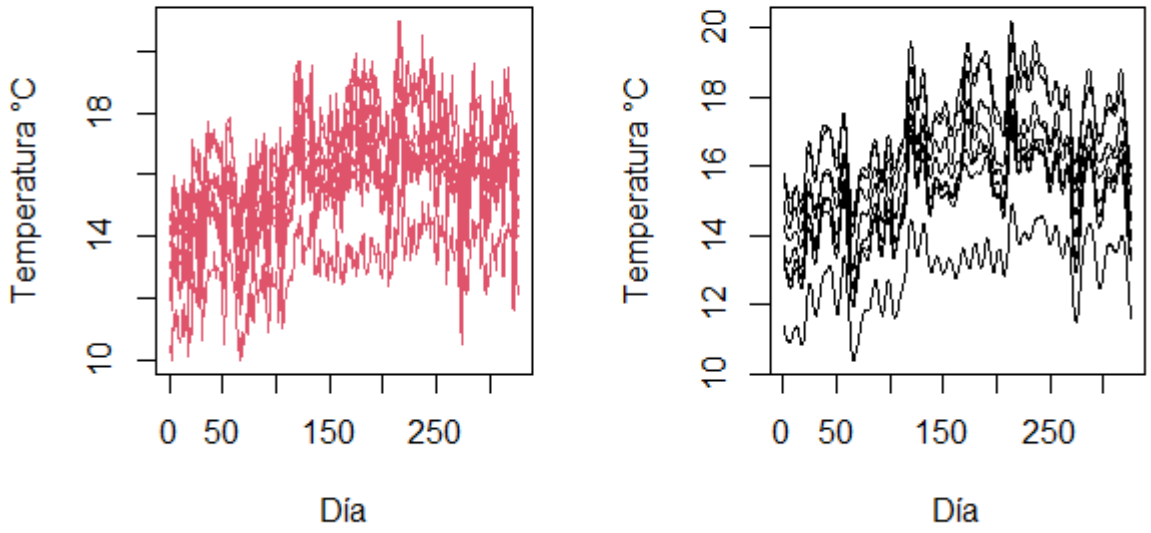

Figura 2.6: Izquierda: Gráfico temperatura promedio diaria simulada. Derecha: Gráfico de curvas suavizadas temperatura promedio diaria simulada.

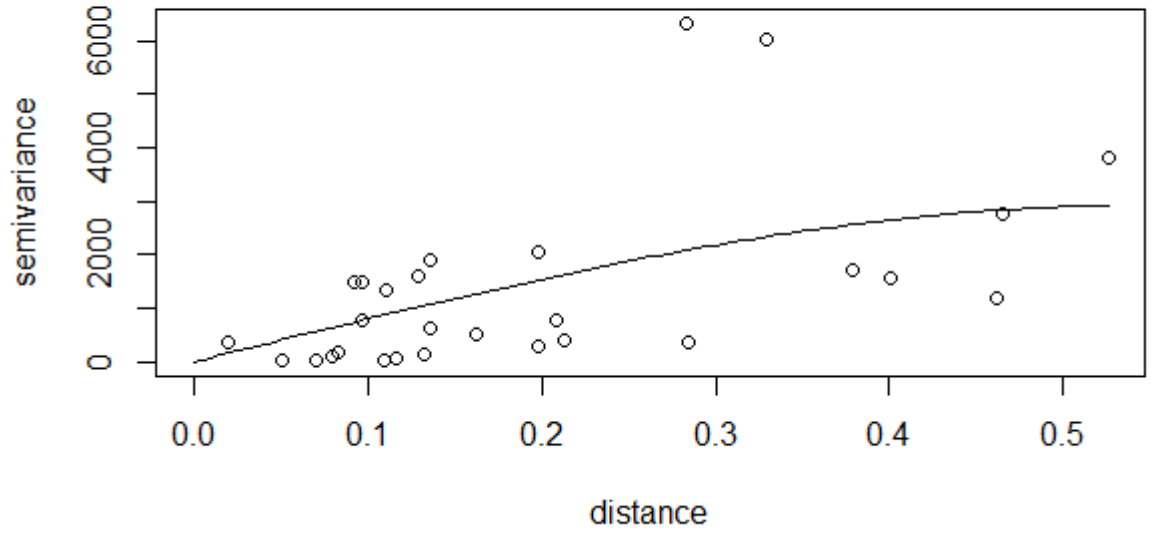

Figura 2.7: Modelo esférico ajustado a la nube variográfica para temperatura promedio diaria simulada en Quito.

Y mediante el MLC obtenemos uno de los inputs para comparar los métodos propuestos.

2.2.2. Precipitación total diaria

La temperatura simulada se suma a diario y luego se incorpora al proceso de suavizado de puntos. En este proceso, utilizamos las bases de Fourier para suavizar los datos, y es esencial obtener los coeficientes correspondientes. Finalmente, graficamos las funciones suavizadas:

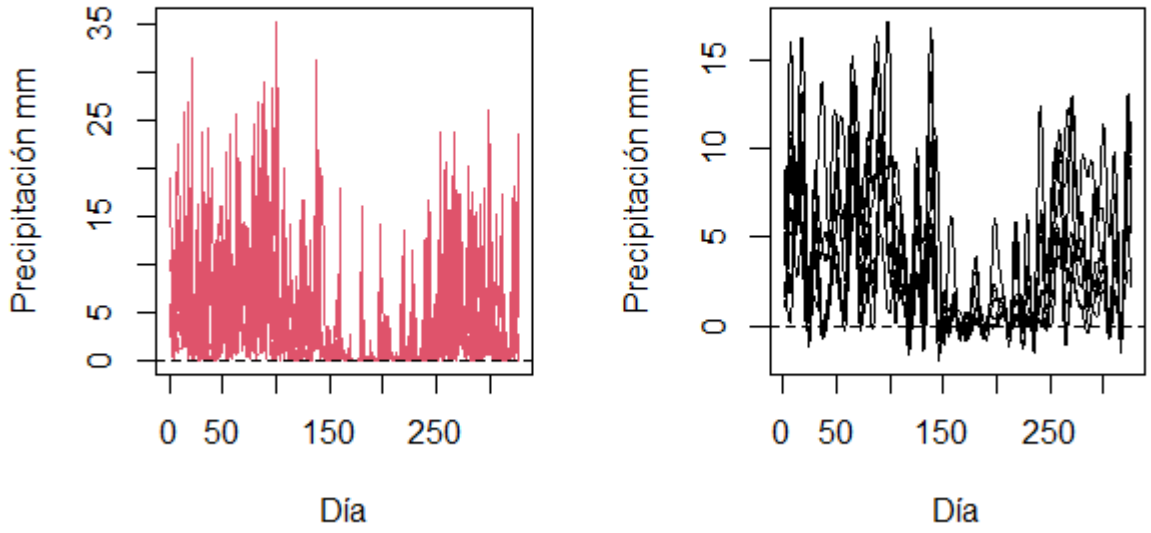

Figura 2.8: Izquierda: Gráfico precipitación total diaria simulada. Derecha: Gráfico de curvas suavizadas precipitación total diaria simulada.

Continuamos definiendo el variograma para analizar la variabilidad espacial de nuestros datos y obtenemos la nube variográfica resultante:

Utilizando el Método Lineal de Corregionalización (MLC), obtenemos uno de los resultados para realizar la comparación con los otros métodos propuestos.

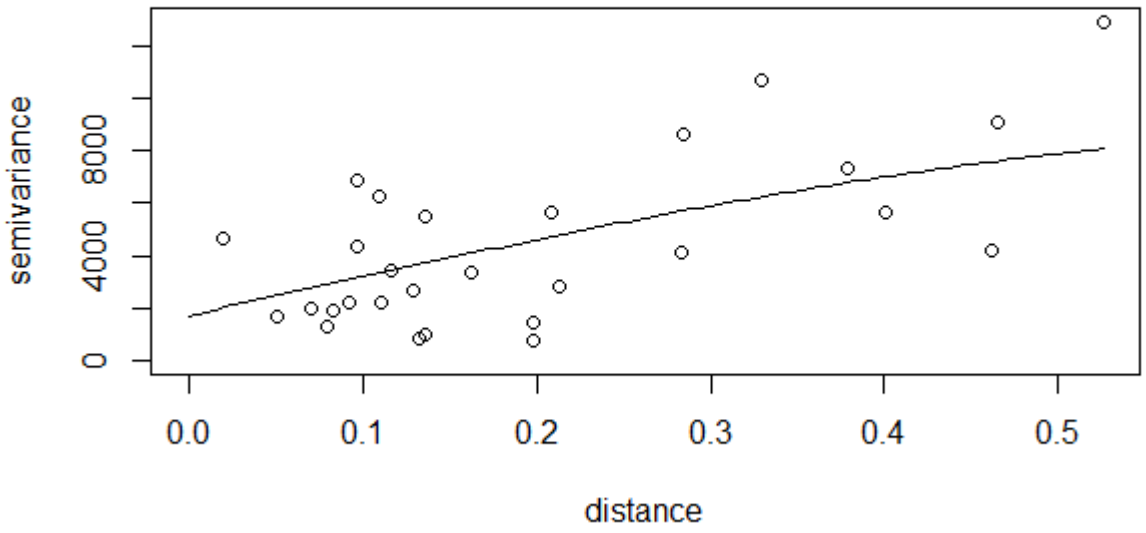

Figura 2.9: Modelo esférico ajustado a la nube variográfica para precipitación total diaria simulada en Quito.

Capítulo 3

Resultados, conclusiones y recomendaciones

En esta sección se presentan los resultados de ejecutar el código computacional que implementa el conglomerado de las variables de precipitación y temperatura de la ciudad de Quito utilizando datos de las estaciones meteorológicas y de las simulaciones realizadas en el modelo WRF. Basados en los resultados obtenidos, expondremos las conclusiones del método de corregionalización espacial.

Las mediciones utilizadas en este análisis se tomaron de la secretaría de ambiente de Quito y corresponden a la temperatura promedio y la precipitación total diaria de la ciudad. Por otra parte, los datos simulados por el modelo WRF corresponden al año 2023 de donde se extrajeron las variables de temperatura y precipitación.

Graficamos los datos reales de la secretaría de ambiente y los datos simulados de las dos variables:

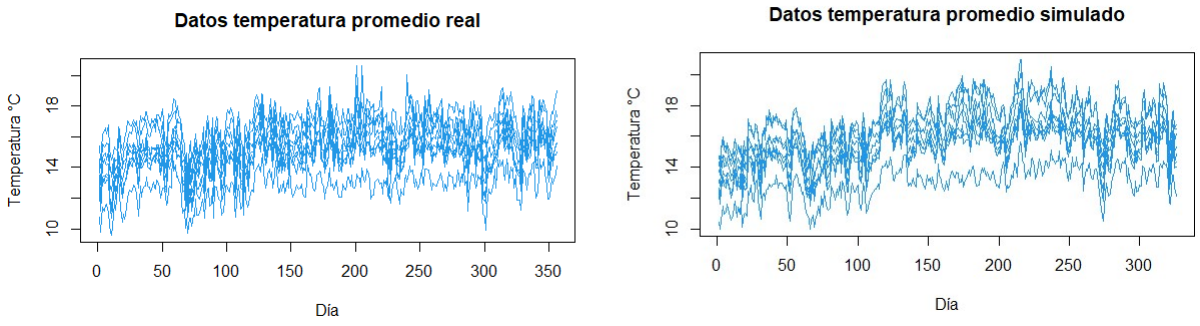

Figura 3.1: Izquierda: Datos de temperatura promedio real diaria. Derecha: Datos de temperatura promedio simulados diaria.

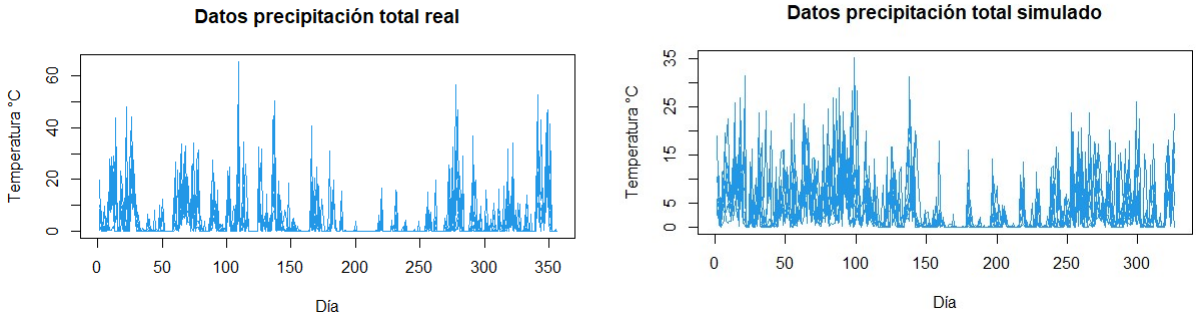

Figura 3.2: Izquierda: Datos de precipitación total real diario. Derecha: Datos de precipitación total simulados diario.

Donde podemos notar las diferencias entre los datos reales y los datos simulados en el mismo periodo de tiempo.

3.1. Resultados

Para los casos de temperatura y precipitación reales y simulados se aplicaron tres procedimientos de agrupamiento jerárquico: el clásico, el procedimiento en el que se asignan pesos a las observaciones en función del variograma y en función del variograma multivariado. Los resultados obtenidos con los enfoques que consideran la dependencia espacial se compararán con el método convencional.

Para visualizar los resultados de mejor manera, se presenta el dendograma y el mapa geográfico en el que se distinguen los grupos que se han formado para cada procedimiento.

3.1.1. Temperatura promedio diaria con datos reales

Para el caso de temperatura promedio diaria, formamos 4 grupos en cada procedimiento. Ejemplificamos el método clásico mediante el dendograma y el mapa en el que se notan los grupos. El primer grupo consta de la estación Guamaní, el segundo de Los Chillos y Tumbaco, el tercer grupo consta de las estaciones Cotocollao, Belisario y El Centro, el último grupo tiene las estaciones Carapungo y San Antonio:

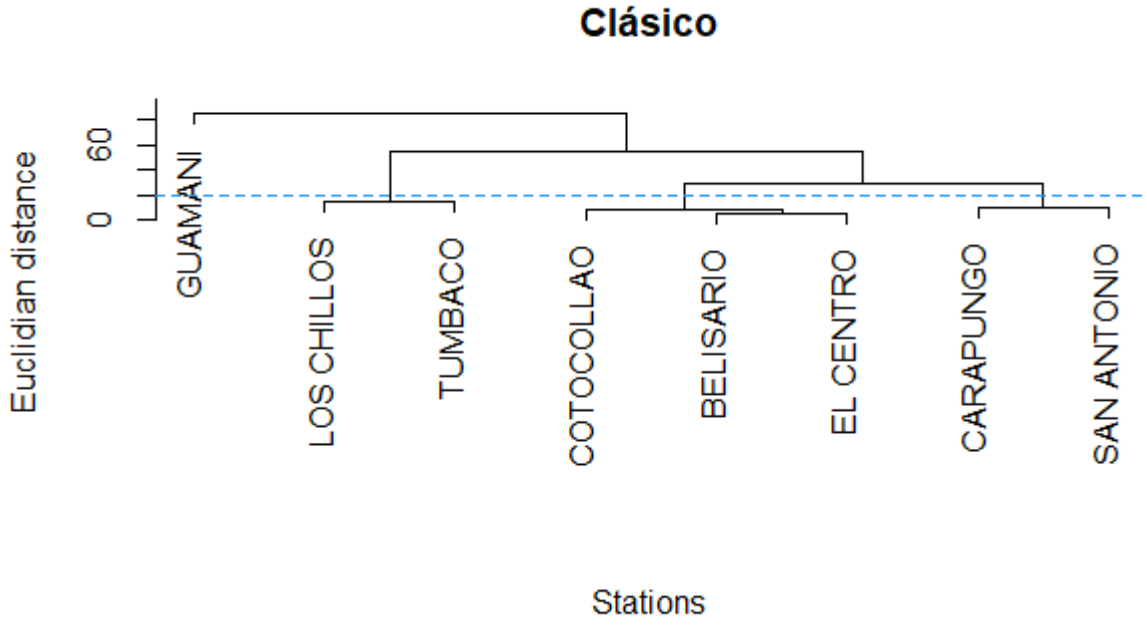

Figura 3.3: Dendograma resultado del agrupamiento por método clásico.

Veamos mediante el mapa los grupos de estaciones que se forman por el corte que hicimos en el dendograma:

Figura 3.4: Mapa en el que se muestran los grupos formados mediante el método clásico para la temperatura promedio diaria real en Quito.

En el caso en que se pondera por el variograma, tenemos el dendograma y el mapa en el que se pueden observar los grupos obtenidos. El primer grupo consta de la estación Guamaní, el segundo de Los Chillos y Tumbaco, el tercer grupo consta de las estaciones Belisario y El Centro, el último grupo tiene las estaciones Carapungo, Cotocollao y San Antonio:

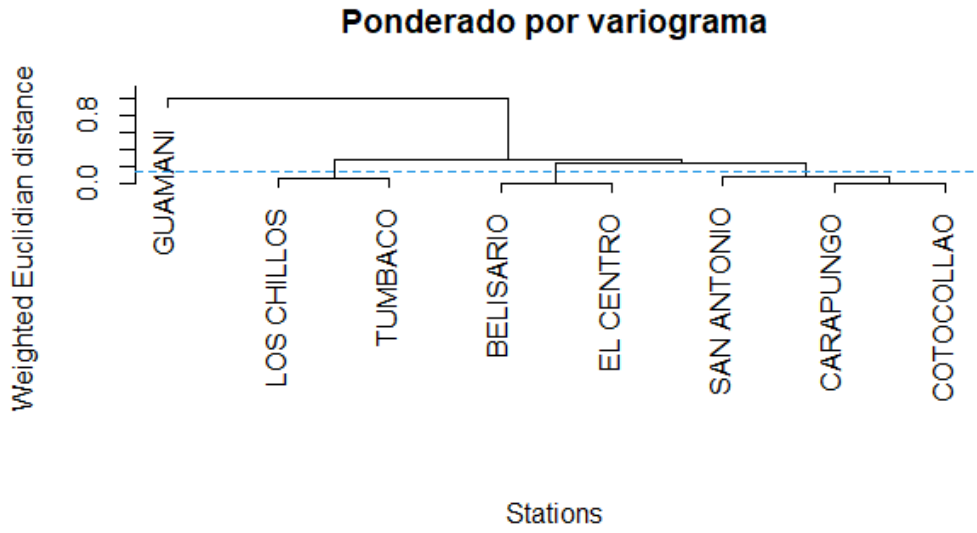

Figura 3.5: Dendograma resultado del agrupamiento por el segundo método.

Veamos mediante el mapa los grupos de estaciones que se forman por el corte que hicimos en el dendograma:

Figura 3.6: Mapa en el que se muestran los grupos formados mediante el segundo método para la temperatura promedio diaria real en Quito.

En el tercer caso, se pondera por el variograma multivariado. El dendograma y el mapa obtenidos son los siguientes:

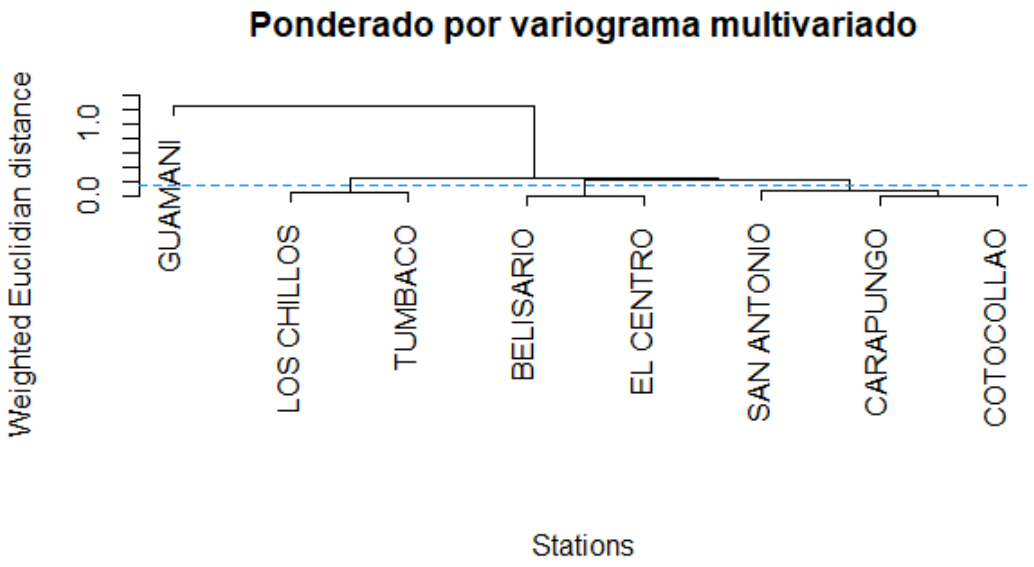

Figura 3.7: Dendograma resultado del agrupamiento por el tercer método.

Veamos mediante el mapa los grupos de estaciones que se forman por el corte que hicimos en el dendograma:

Figura 3.8: Mapa en el que se muestran los grupos formados mediante el tercer método para la temperatura promedio diaria real en Quito.

Notemos las diferencias entre los métodos propuestos; el método clásico, al no considerar la georeferenciación, crea grupos con estaciones en el centro de la ciudad y en el norte (Belisario, El Centro y Cotocollao) como uno solo. En contraste, el segundo y tercer caso toman en cuenta la ubicación de las estaciones y por tanto, se aprecia que agrupan los puntos del sur (Guamaní), los puntos de los valles (Los Chillos, Tumbaco), las estaciones del centro (El Centro, Belisario) y las estaciones del norte (San Antonio, Carapungo y Cotocollao).

Así notamos la diferencia de tomar la parte geográfica como parte del proceso de clustering.

3.1.2. Precipitación total diaria con datos reales

Implementamos los tres métodos de agrupamiento jerárquico para la variable de precipitación medida en las estaciones mencionadas. En este caso, es de interés la precipitación acumulada. Nuevamente, los casos considerados son: el enfoque convencional, el que asigna pesos a las

observaciones según el variograma, y el que asigna pesos a las observaciones según el variograma multivariado. Para una representación visual más efectiva de los datos, se grafica el dendrograma y el mapa que ilustra los grupos formados, nuevamente con el objetivo es comparar los resultados obtenidos mediante el método tradicional con aquellos generados por los enfoques que incorporan la parte espacial.

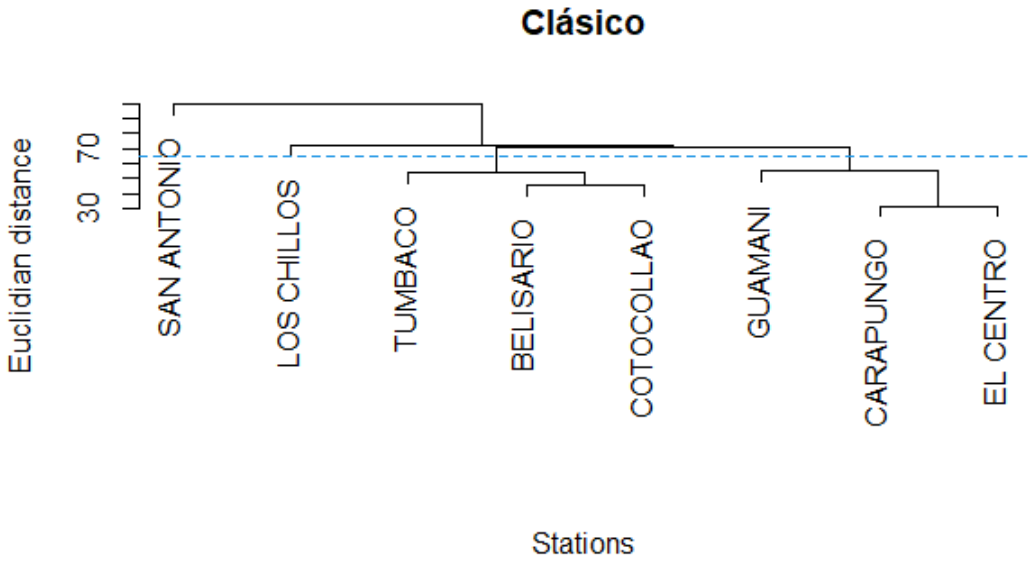

Figura 3.9: Dendograma resultado del agrupamiento por método clásico.

Mediante mapa se nota claramente que existen cruces geográficos en los grupos ya que este método no toma en cuenta la ubicación espacial:

Figura 3.10: Mapa en el que se muestran los grupos formados mediante el método clásico para la precipitación total diaria real en Quito.

En el segundo caso, tenemos el dendograma y el mapa en el que se notan los grupos considerando la ubicación de las estaciones.

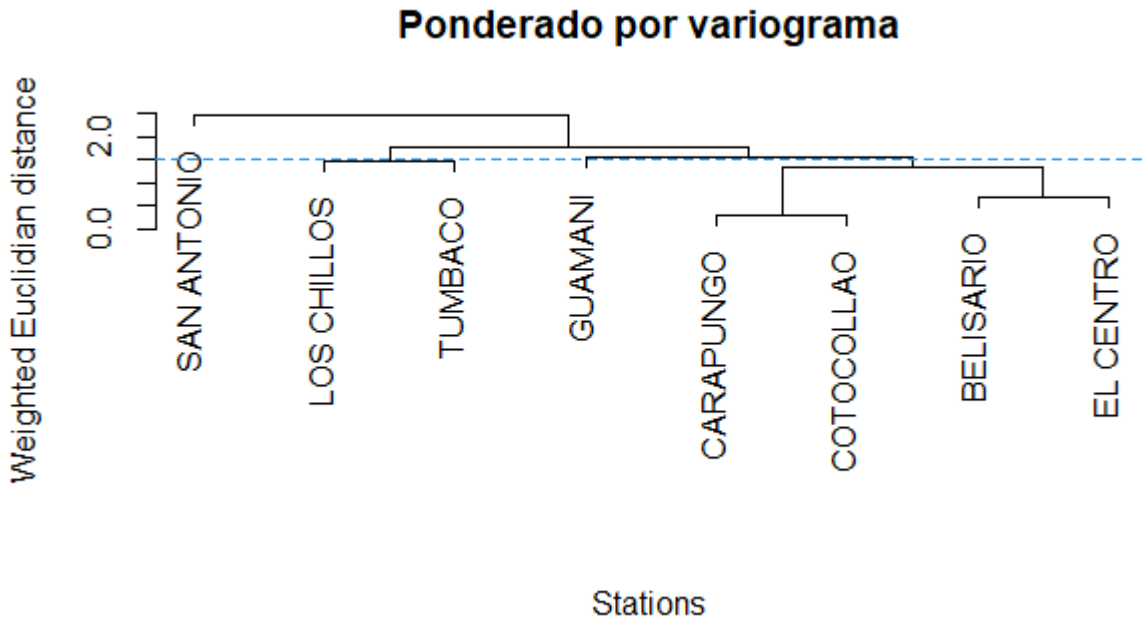

Figura 3.11: Dendograma resultado del agrupamiento por el segundo método.

Mediante el mapa notamos la diferencia con el método clásico ya que

los grupos consideran la georeferenciación de las estaciones climatológicas:

Figura 3.12: Mapa en el que se muestran los grupos formados mediante el segundo método para la precipitación total diaria real en Quito.

Vamos al tercer caso, tenemos el dendograma y el mapa en el que se notan los grupos considerando las ubicaciones:

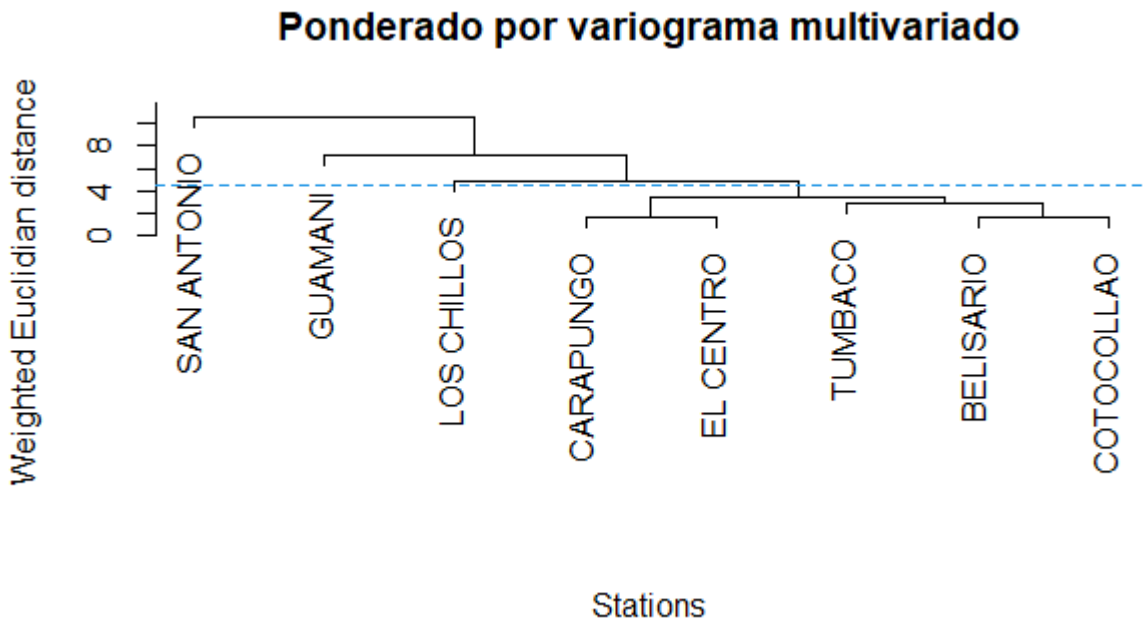

Figura 3.13: Dendograma resultado del agrupamiento por el tercer método.

En este caso, se ha elegido 4 grupos con el objetivo de comparar todos los métodos. Si notamos el dendograma podríamos cortar los grupos a un mayor número en el que se nota cómo varían los casos al considerar las ubicaciones:

Figura 3.14: Mapa en el que se muestran los grupos formados mediante el tercer método para la precipitación total diaria real en Quito.

En el caso de la variable de precipitación total, las diferencias son más evidentes. En el primer escenario, la estación San Antonio constituye el primer grupo, seguida por la estación Los Chillos en el segundo grupo. El tercer grupo engloba varias estaciones climatológicas, como Tumbaco, Belisario y Cotocollao, mientras que el cuarto grupo incluye a las estaciones Carapungo, Belisario y Guamaní. Visualmente, se aprecia cómo los puntos no tienen en cuenta la ubicación geográfica, en marcado contraste con los métodos que sí consideran este aspecto. Estas diferencias subrayan la importancia de abordar la componente espacial al analizar estos datos.

3.1.3. Temperatura promedio diaria con datos simulados

Nuevamente consideramos el enfoque convencional, el que asigna pesos a las observaciones según el variograma, y el que asigna pesos según

el variograma multivariado. Con el propósito de lograr una representación visual más efectiva de los datos, se genera tanto el dendrograma como el mapa que visualiza los grupos formados. Iniciamos con el método clásico:

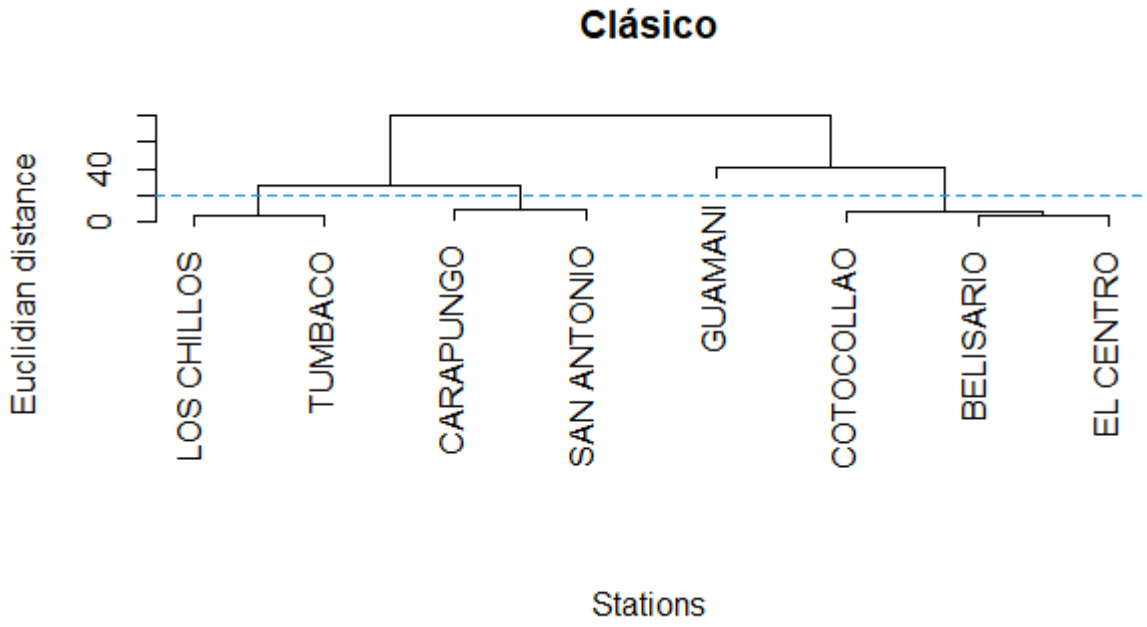

Figura 3.15: Dendograma resultado del agrupamiento por el método clásico.

En los datos simulados vemos una agrupación acorde a las ubicaciones geográficas a pesar de no tomarlas en cuenta para el primer caso:

Figura 3.16: Mapa en el que se muestran los grupos formados mediante el método clásico para la temperatura promedio diaria simulada en Quito.

En el segundo caso, se presenta tanto el dendrograma como el mapa, donde se evidencian los grupos tomando en consideración la ubicación geográfica de las estaciones.

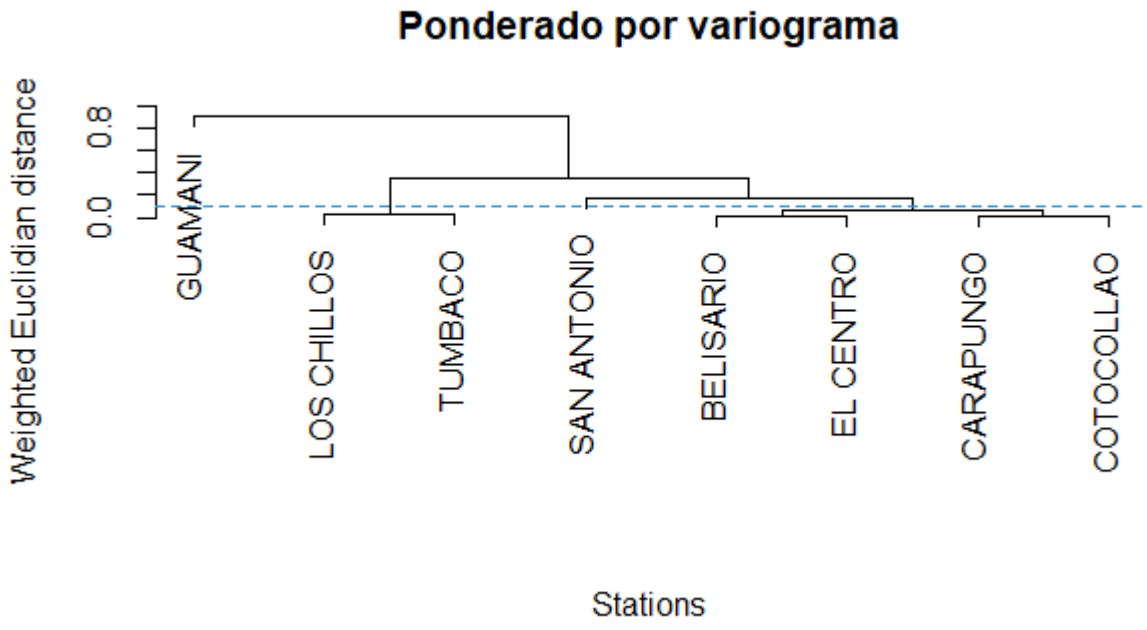

Figura 3.17: Dendograma resultado del agrupamiento por el segundo método.

Este enfoque permite una visualización más completa y explícita de la estructura de agrupamiento, revelando cómo la proximidad espacial influye en la formación de grupos.

Figura 3.18: Mapa en el que se muestran los grupos formados mediante el segundo método para la temperatura promedio diaria simulada en Quito.

Vamos al tercer caso, donde tenemos el dendograma y el mapa en el que se notan los grupos considerando las ubicaciones:

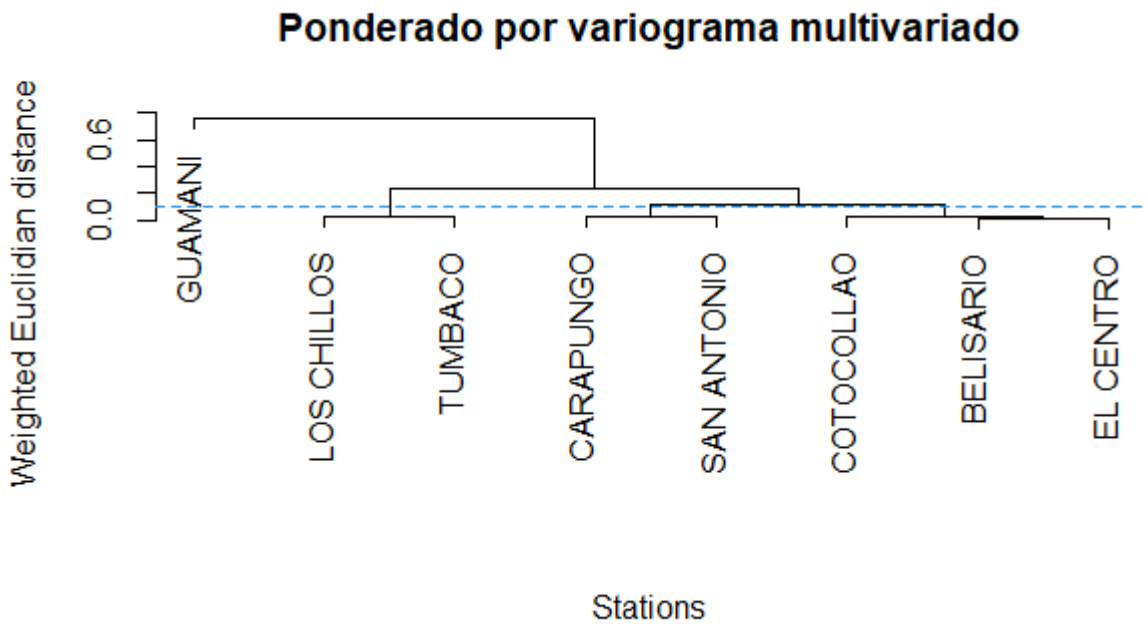

Figura 3.19: Dendograma resultado del agrupamiento por el tercer método.

Figura 3.20: Mapa en el que se muestran los grupos formados mediante el tercer método para la temperatura promedio diaria simulada en Quito.

En los datos simulados de temperatura promedio diaria no notamos grandes diferencias entre los métodos ya que esta variable está muy bien marcada entre los sectores considerados dentro de Quito.

3.1.4. Precipitación total diaria con datos simulados

En el último grupo que tenemos aplicamos los tres métodos de agrupamiento jerárquico que ya hemos mencionado. Se construyen tanto el dendrograma como el mapa en el que observamos los grupos formados. Una vez más, el propósito es comparar los resultados obtenidos a través de los tres métodos. Mediante el método clásico obtenemos lo siguiente:

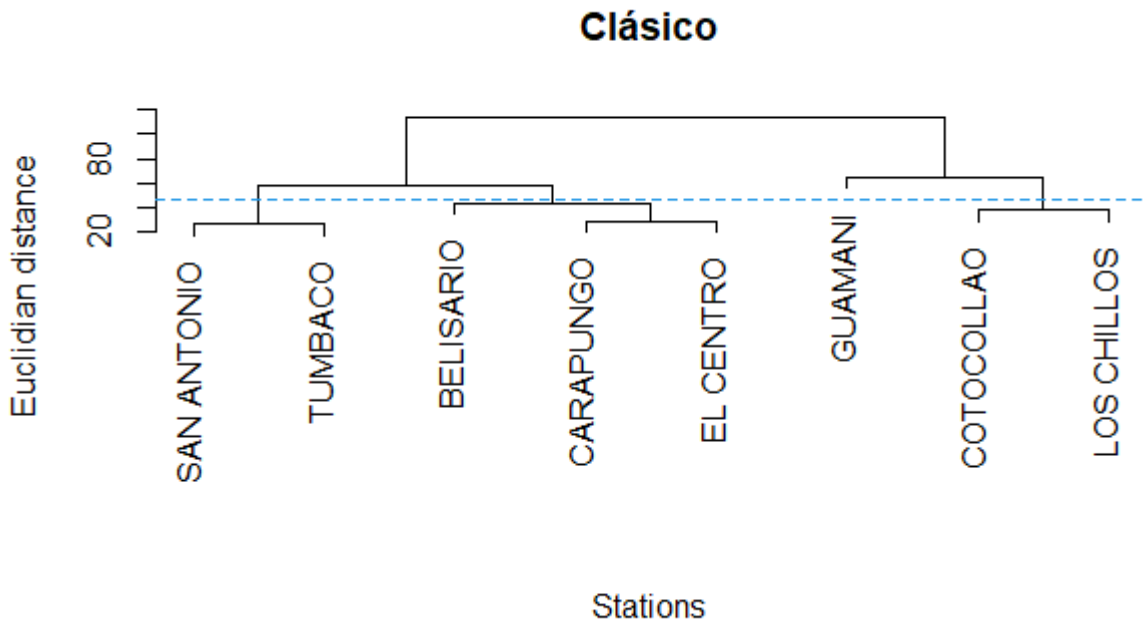

Figura 3.21: Dendograma resultado del agrupamiento por el método clásico.

A través del mapa, es evidente que este método no considera la ubicación espacial. Los grupos formados no reflejan patrones geográficos discernibles, lo que muestra que la información de proximidad no se ha incorporado en el proceso de agrupamiento, por lo que continuamos con los métodos que si consideran la ubicación geográfica de las estaciones. Esto lo notamos en el grupo uno (Tumbaco, San Antonio), grupo dos (Belisario, Carapungo y El Centro) y cuatro (Cotocollao y Los Chillos) que están dispersos por la ciudad chocando unos con otros.

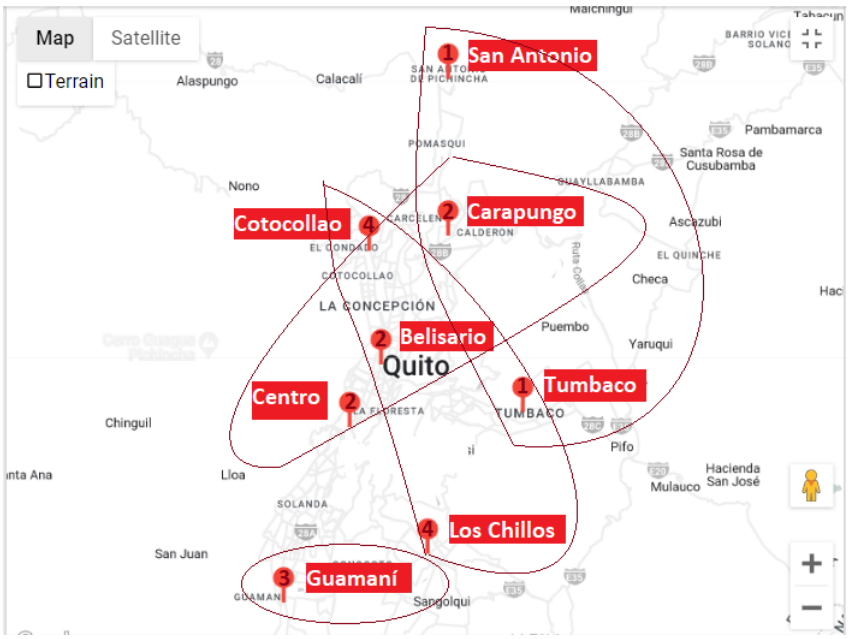

Figura 3.22: Mapa en el que se muestran los grupos formados mediante el método clásico para la precipitación total diaria simulada en Quito.

Continuando con el segundo método:

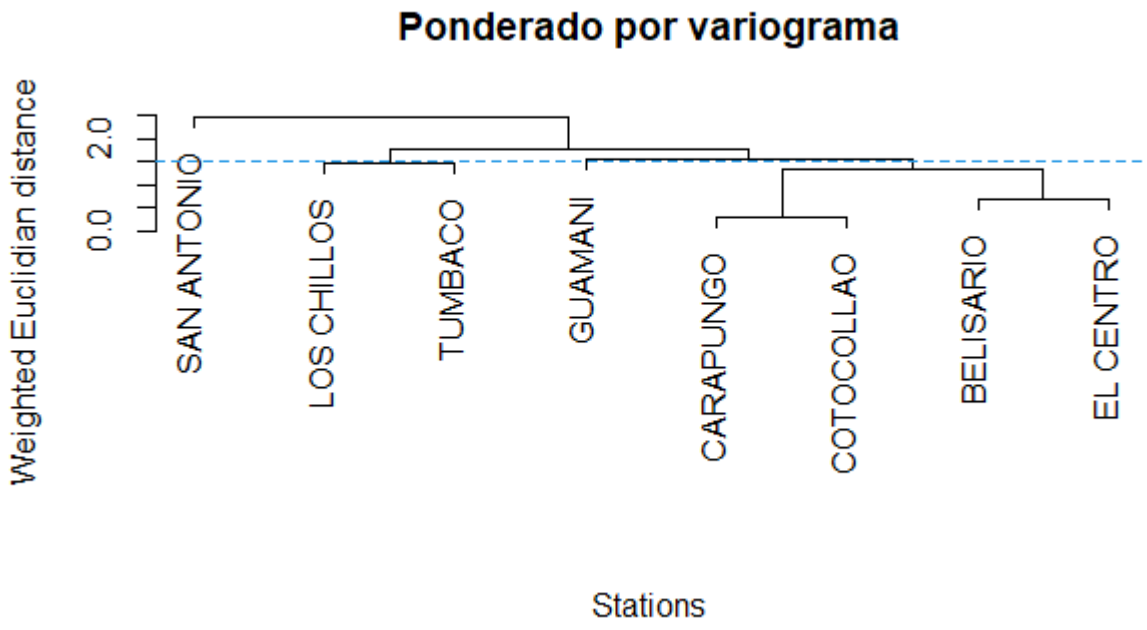

Figura 3.23: Dendograma resultado del agrupamiento por el segundo método.

A través del mapa, se evidencia una diferencia con el método clásico, ya que los grupos formados ahora tienen en cuenta la georeferenciación.

Esta consideración espacial en el proceso de agrupamiento ha resultado en una distribución de grupos que refleja de mejor manera las relaciones geográficas entre las estaciones.

Figura 3.24: Mapa en el que se muestran los grupos formados mediante el segundo método para la precipitación total diaria simulada en Quito.

En el tercer caso, tenemos el dendograma y el mapa:

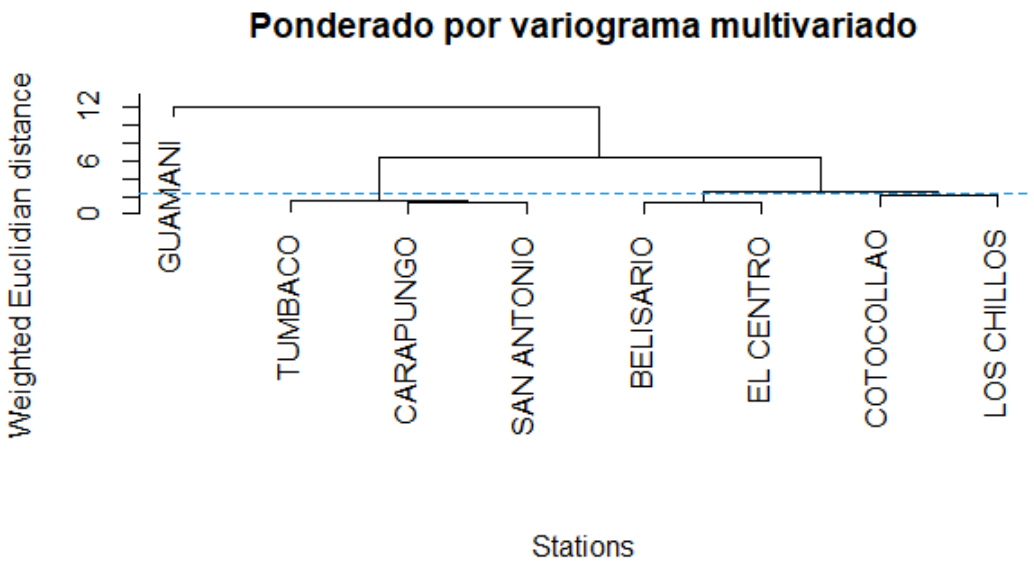

Figura 3.25: Dendograma resultado del agrupamiento por el tercer método.

Se ha elegido cuatro grupos en el corte del dendograma con el objetivo de compararlos con los grupos formados por los otros métodos. En este caso tenemos al primer grupo formado por las estaciones Cotocollao y Los Chillos que chocan con el grupo formado por el grupo de El Centro y Belisario, únicamente por el número de grupos que hemos elegido, pues si elegimos bajar el corte en el dendograma separamos el primer grupo y claramente se nota la diferencia entre el método clásico y el método que se sugiere al tener observaciones con correlación espacial.

Figura 3.26: Mapa en el que se muestran los grupos formados mediante el segundo tercer para la precipitación total diaria simulada en Quito.

3.2. Conclusiones y recomendaciones

3.2.1. Conclusiones

1. Se ha utilizado 3 metodologías para agrupar las estaciones que miden algunas variables climatológicas en la ciudad de Quito, de los que hemos tomado la temperatura promedio y la precipitación total diario del año 2023. Estos datos fueron suavizados, obteniendo el input requerido, los datos funcionales con correlación espacial.

2. Se considera la ubicación geográfica de las estaciones climatológicas

dentro de las metodologías propuestas en [5], destacando su utilidad como herramienta de clasificación y su rendimiento positivo cuando es necesario considerar la correlación espacial en el análisis de agrupamiento en las variables simuladas por el proyecto METEO, cumpliendo así con el objetivo general.

3. Un caso interesante es el de la precipitación total diaria con datos simulados, que si bien los métodos que añaden la referencia espacial mejora los grupos según su ubicación, la elección del número de grupos impacta, como observamos en la figura 3.26, en el cruce de grupos. Si el corte en el dendograma se realiza más abajo, con el objetivo de tener 5 grupos, las estaciones Cotocollao y Los Chillos pueden separarse obteniendo un resultado óptimo en el ejercicio.

4. Las alternativas que se han propuesto al presentar los resultados para las variables elegidas han demostrado un mejor comportamiento que el enfoque clásico de agrupamiento de datos funcionales, especialmente cuando existen diferencias significativas entre las curvas medias de los sitios y correlación espacial.

5. El tercer método que asigna pesos a las observaciones mediante el variograma multivariado se considera el más adecuado para los casos y variables presentados en este documento, resta elegir el número de grupos adecuado según lo que se requiera.

3.2.2. Recomendaciones

Esta metodología puede ser empleada en la elaboración de informes sobre variables climatológicas para crear informes sobre grupos que comparten características y proximidad geográfica. A modo de ejemplo, consideremos un informe elaborado por METEO

Figura 3.27: Grupo seleccionado empíricamente para reportar la probabilidad de lluvia acumulada en sectores de Quito.

La selección de los sectores presentados en la imagen ha sido realizada de manera empírica; no obstante, se sugiere que la metodología propuesta podría aplicarse para crear estos grupos mediante criterios técnicos.

Además, se recomienda ampliar este ejercicio a otras variables meteorológicas, como el viento, la radiación solar, la humedad relativa, radiación ultravioleta, entre otras. Estas variables, al igual que la temperatura y la precipitación, pueden obtenerse de la Secretaría del Ambiente de Quito o extender la premisa a otras ciudades y países.

Referencias bibliográficas

[1] *Sistema de pronóstico del clima y el tiempo para el territorio ecuatoriano: modelización numérica y estadística*. MeteoEcuador.

[2] M. Leese y D. Stahl B. Everitt, S. Landau. *Cluster Analysis*. Wiley, 2011.

[3] R. Giraldo. *Introducción a la Geoestadística*. Universidad Nacional de Colombia.

[4] M. Oliver and R. Webster. *A geostatistical basis for spatial weighting in multivariate classification*. Mathematical Geology, Vol. 21, nº 1, pp. 15-35, 1989.

[5] P. Delicado y J. Mateu R. Giraldo. *Hierarchical clustering of spatially correlated functional data*. Statistica Neerlandica, vol. 66, nº 4, pp. 403-421, 2012.

[6] J. O. Ramsay y B. Silverman. *Functional Data Analysis*. New: Springer Science+Business Media, Inc., 2005.

[7] G. Bourgault y D. Marcotte. *Multivariable Variogram and Its Application to the Linear Model of Coregionalization*. Mathematical Geology, vol. 23, nº 7, pp. 899-928, 1991, 1991.