Traballo Fin de Grao

Determinación de la dependencia espacial mediante variogramas

Carlos López Lado

2024-2025

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRADO DE MATEMÁTICAS

Traballo Fin de Grado

Determinación de la dependencia espacial mediante variogramas

Carlos López Lado

Febrero, 2025

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Conocimiento: Estadística e Investigación Operativa
Título: Determinación de la dependencia espacial mediante variogramas
Breve descripción del contenido
El objetivo de este trabajo es aprender a dominar los fundamentos para la inferencia xeoestatística con técnicas Kriging donde la dependencia espacial depende de la función llamada variograma que no es otra cosa que la varianza de la diferencia entre datos a distancia d. El trabajo consistirá en la descripción de las herramientas necesarias para analizar la dependencia espacial, estimar el variograma y elaborar mapas de predicción para una región concreta.
Recomendaciones
Otras observaciones

Índice

Resumen IX

Introducción XI

1. Fundamentos Matemáticos de la Inferencia Geoestadística **1**

1.1. Variable Aleatoria y Función Aleatoria 1

1.2. Valor Regionalizado y Dependencia Espacial 3

1.3. Estacionariedad 5

1.3.1. Tipos de Estacionariedad 6

2. Estimación del Variograma **9**

2.1. Introducción a la Medición de la Dependencia Espacial 9

2.2. Cálculo del Variograma Experimental 11

2.2.1. Factores que Afectan a la Fiabilidad de los Variogramas Experimentales . 14

2.3. El Variograma Teórico y Sus Características 15

2.3.1. Relación entre el Variograma y la Función de Covarianza 17

2.4. Modelos de Variograma Teórico 19

2.4.1. Modelo Efecto-nugget 19

2.4.2. Modelo Esférico 20

2.4.3. Modelo Exponencial 20

2.4.4. Modelo Gaussiano 21

2.5. Modelado del Variograma Teórico en Situaciones Anisotrópicas 22
2.5.1. Anisotropía Geométrica 23
2.5.2. Anisotropía Zonal 25
2.6. Estimación del Variograma: Caso práctico 26
2.6.1. Análisis Exploratorio del Variograma 27
2.6.2. Modelado del Variograma Teórico 32
2.6.3. Anisotropía 35

3. Teoría Geoestadística y Método de Kriging 37

3.1. Kriging Ordinario 38
3.2. Kriging Universal 42
3.3. Kriging Multivariado (Cokriging) 45
3.3.1. Función de Covarianza cruzada y Variograma cruzado 47
3.3.2. Cokriging Ordinario 48

4. Caso Práctico: Aplicación en Datos Reales 51

4.1. Descripción de los Datos 51
4.2. Análisis Exploratorio de los Datos 53
4.2.1. Enfoque del Análisis de los Contaminantes SO_2, NO_x y PM_{10} 54
4.2.2. Análisis Exploratorio de los Contaminantes SO_2, NO_x y PM_{10} 55
4.3. Aplicación de Kriging 57
4.4. Interpretación de Resultados y Conclusión 59

Bibliografía 61

I. Código Fuente 65

I.1. Construcción de la Base de Datos 65
I.2. Análisis Exploratorio de la Base de Datos 67

I.3. Construcción Final de la Base de Datos con Media por Horas del Día 68

I.4. Construcción del Grid de Galicia a partir de un Shapefile 69

I.5. Interpolación Espaciotemporal mediante Kriging 69

II. Mapas **71**

II.1. Gráficas con Grid del Kriging Ordinario 71

II.2. Gráficas Continuas del Kriging Ordinario 72

Resumen

En este trabajo se realiza una introducción a la geoestadística, centrándose especialmente en el concepto de variograma, estructura que cuantifica la dependencia espacial, y el método de interpolación espacial Kriging. Para ello se exponen las bases teóricas de la dependencia espacial como fundamento para el desarrollo del variograma, incluyendo la concepción experimental pero también teórica del mismo, así como los distintos modelos existentes y razones por las que puede no modelizar correctamente la dependencia espacial. A continuación se presenta la teoría detrás del método de interpolación Kriging, junto con las diferentes variantes del modelo: el ordinario, el universal y el multivariante. Finalmente, se presenta un caso práctico que plasma la utilidad de estos conceptos con el fin de modelar, mediante las librerías gstat y sm de R, la interpolación de los contaminantes SO_2, PM_{10} y NO_x en el territorio gallego.

Abstract

This work provides an introduction to geostatistics, focusing particularly on the concept of the variogram, a structure that quantifies spatial dependence, and the Kriging spatial interpolation method. To this end, the theoretical foundations of spatial dependence are presented as the basis for the development of the variogram, including both its experimental and theoretical conception, as well as the different existing models and the reasons why it may fail to properly model spatial dependence. Next, the theory behind the Kriging interpolation method is introduced, along with its different variants: ordinary, universal, and multivariate Kriging. Finally, a practical case is presented to illustrate the usefulness of these concepts, aiming to model the interpolation of the pollutants SO_2, PM_{10} and NO_x in the Galician territory using the R libraries gstat and sm.

Introducción

La geoestadística es la ciencia que, con la combinación de matemática y naturaleza, nos permite encontrar los patrones ocultos en la distribución espacial de diferentes fenómenos físicos. Su aplicación abarca desde la minería, donde estudia la distribución de minerales, hasta la modelización de contaminantes atmosféricos que pueden afectar a la salud de las personas. Nos enseña que, en muchas situaciones, la naturaleza se estructura de forma que los puntos cercanos están conectados entre si, compartiendo características y valores similares.

En ese sentido se desarrolla este trabajo, donde se presentarán primero las herramientas matemáticas sobre las que se sustenta esta rama de la estadística. Siendo la más importante de ellas el concepto del variograma, el cuál da título a este trabajo, y que permite cuantificar la dependencia espacial del fenómeno de estudio. Después de su presentación se aplicarán a un caso práctico concreto, la distribución de contaminantes en el territorio gallego.

El objetivo último es el desarrollo de mapas de predicción de la distribución de la contaminación en la comunidad, que permita entender donde y de que forma se concentran los contaminantes SO_2, PM_{10} y NO_x. Para ello se hará uso de uno de los métodos de interpolación más ampliamente utilizado en este campo, el método desarrollado por G. Matheron en 1963, que recibe el nombre de Kriging.

Este trabajo se compone de cuatro capítulos, tres de ellos enfocados al desarrollo teórico de las herramientas matemáticas necesarias para la obtención de los mapas de contaminación. El primero introduce el concepto de dependencia espacial y sus bases teóricas; el segundo está centrado en el variograma, comprender la teoría detrás de él pero también se estudiarán los problemas que pueden surgir en su modelado y como dichos problemas pueden afectar a nuestras predicciones. Además, al final de dicho capítulo se presentará el cálculo del variograma del caso práctico. El tercer capítulo se enfoca en el desarrollo teórico del Kriging, y de sus diferentes variantes; ordinario, universal y multivariante.

Esta estructura tiene como fin alcanzar una base solida en los fundamentos geoestadísticos para, ya en el último capítulo, presentar a partir de los datos públicos de la estaciones de la comunidad los mapas de distribución de los contaminantes.

Por último, en los dos Anexo, se presenta el código en R utilizado para el análisis de los datos y algunos de los mapas complementarios que no se han podido incluir en el texto principal del trabajo.

Capítulo 1

Fundamentos Matemáticos de la Inferencia Geoestadística

En este primer capítulo se hará una introducción a los fundamentos matemáticos básicos para el desarrollo de la inferencia geoestadística. En este sentido es importante repasar los conceptos de variable aleatoria y función aleatoria de estadística básica junto con los conceptos, más enfocados a la geoestadística, de dependencia espacial y estacionariedad. Se comenzará por los conceptos más básicos antes de adentrarse en los aspectos aplicados a la inferencia espacial.

1.1. Variable Aleatoria y Función Aleatoria

Definición 1.1. Una **variable aleatoria** es una cantidad que toma diferentes valores debido a la variabilidad asociada a un fenómeno aleatorio. Aplicado en el contexto de la geoestadística, este concepto implica que cada punto en el espacio puede tener una cantidad, variable aleatoria, asociada que representa el valor de una propiedad de interés. Matemáticamente $Z(x)$ es una variable aleatoria que define el valor del atributo espacial en el punto x.

Ejemplo 1.2. La concentración de ozono a nivel del suelo es un ejemplo de variable aleatoria en el contexto de la geoestadística. El ozono puede tomar diferentes valores en distintas ubicaciones debido a factores aleatorios, como las condiciones meteorológicas y las fuentes de contaminación en la zona. Por lo tanto, si se realiza un estudio sobre la calidad del aire, se puede considerar que en cada ubicación la concentración de ozono es una variable aleatoria $Z(x)$, donde x representa un punto específico en el espacio (Gorai et al., 2015) [9].

El concepto de variable aleatoria permite la construcción del vector aleatorio, que no es más que una colección de variables aleatorias medidas simultáneamente sobre el mismo resultado.

Aplicado a geoestadística sería un vector que proporciona información sobre diferentes variables aleatorias en un punto concreto del terreno.

Ejemplo 1.3. Un ejemplo de vector aleatorio en geoestadística es el que contiene la concentración de clorofila, salinidad y conductividad en el río Ebro. En este caso, en cada punto de muestreo x se define un vector aleatorio que representa los valores de estas tres propiedades:

$$Z(x) = \begin{pmatrix} Z_{\text{clorofila}}(x) \\ Z_{\text{salinidad}}(x) \\ Z_{\text{conductividad}}(x) \end{pmatrix}$$

Donde:

- $Z_{\text{clorofila}}(x)$ representa el valor de la concentración de clorofila en la ubicación x,
- $Z_{\text{salinidad}}(x)$ representa el valor de salinidad en la misma ubicación,
- $Z_{\text{conductividad}}(x)$ representa el valor de conductividad en la misma ubicación.

Este vector aleatorio permitirá analizar conjuntamente las propiedades medidas en cada ubicación, capturando así la relación espacial entre estas variables ambientales (Wackernagel, 2003) [26].

Definición 1.4. Una **función aleatoria** es una extensión del concepto de variable aleatoria. A diferencia de esta, que toma un solo valor en un punto específico, una función aleatoria asigna valores aleatorios a cada punto dentro de un área o intervalo en el espacio. Matemáticamente, si tenemos un dominio $D \in \mathbb{R}^n$, la función aleatoria $Z(x)$ describe el valor aleatorio de una variable en cada punto $x \in D$.

Expresado de forma más rigurosa se puede decir que una función aleatoria $Z : \Omega \to \mathbb{R}$, dado un espacio de probabilidad $(\Omega, \mathbb{A}, \mathbb{P})$, es una función que asocia un número real a cada suceso elemental de Ω, verificando la propiedad de que el conjunto $\{x \in \Omega \mid Z(x) \leq r\} = Z^{-1}((-\infty, r]) \in \mathbb{A}$.

Observación 1.5. En geoestadística, $Z(x)$ se usa tanto para denotar una función aleatoria en el dominio, Definición 1.4, como para representar el valor de una variable aleatoria en una ubicación específica x, Definición 1.1. En general se entiende qué significado de $Z(x)$ se pretende, pero en caso de creerse necesario se esclarecerá.

1.2. Valor Regionalizado y Dependencia Espacial

Aunque se ha estado hablando de variables aleatorias en puntos específicos, es importante recalcar que esto no es necesariamente así. En términos generales, una variable no necesariamente está asociada a un punto específico del espacio. Si se adopta ese enfoque es porque este trabajo se desarrolla en el contexto específico de la geoestadística, donde surge otro concepto clave, el valor regionalizado.

Definición 1.6. El concepto de **valor regionalizado** hace referencia a una propiedad que varía en función de su posición en el espacio. A diferencia de una variable aleatoria, que es el resultado de evaluar una función aleatoria en un punto del espacio, cuando se dice que una propiedad es un valor regionalizado se está diciendo que dicha propiedad no solo varía aleatoriamente, sino que exhibe cierta estructura espacial. Como señala Matheron (1971), "Este término [valor regionalizado] es neutral y puramente descriptivo, y no asume una interpretación probabilística"(p. 5) [14].

Esta noción del valor regionalizado como un término puramente descriptivo puede parecer confusa, sobre todo, cuando al mismo tiempo se habla de variación aleatoria. Sin embargo, aquí reside la clave del concepto. En geoestadística, se considera que las muestras tienen dos aspectos complementarios, uno aleatorio que se manifiesta en variaciones irregulares entre puntos y "otro estructurado que refleja en cierta medida las características estructurales del fenómeno"(Matheron, 1971) [14].

Por lo tanto, tenemos dos aspectos, la regionalidad y la aleatoriedad, para combinar ambos se utiliza la función aleatoria 1.4, como se puede observar en la Figura 2.1. Es esta la encargada de capturar ambas "realidades" a la vez, la realidad del entorno físico que hace que el resultado de la función dependa hasta cierto punto de su ubicación en el espacio, el valor regionalizado. Y la otra "realidad" que consiste en que, aunque los datos estén regionalizados, la muestra no puede modelarse de forma determinista, lo que lleva a asumir un enfoque probabilístico que considere cierta aleatoriedad.

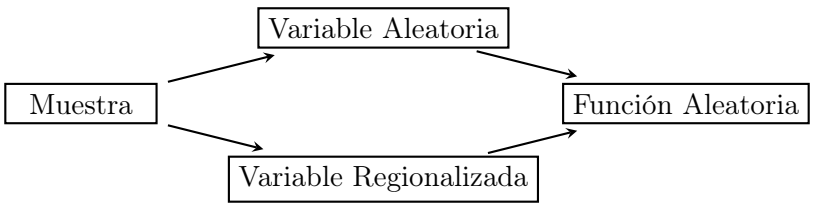


Figura 1.1: Modelo de la función aleatoria, reproducido de *Multivariate Geostatistics: An Introduction with Applications*, por H. Wackernagel, 2003, p. 41 [26].

Observación 1.7. Debido a esta doble naturaleza de la función aleatoria, utilizaremos el término **realización** y se dirá que un valor regionalizado $z(x_0)$ en una ubicación concreta x_0 es la realización de la variable aleatoria $Z(x_0)$, la cual pertenece a una familia infinita de variables aleatorias asociadas a la función aleatoria $Z(x)$.

Observación 1.8. En cuanto al uso de la notación, se denotará la variable aleatoria como $Z(x)$ y su realización como $z(x)$.

Para ejemplificar el concepto se puede suponer una zona de estudio, como un área minera, donde se mida como la concentración de un determinado mineral varía de un punto a otro, sin asumir ninguna causa específica de esta variación ni un componente aleatorio. En la práctica esta variación espacial tiene dos componentes, uno aleatorio, aunque se observen tendencias existe cierta incertidumbre o variabilidad entre la concentración de mineral en un punto y otro. Y un componente de estructura espacial o dependencia, los puntos cercanos tienden a tener una concentración de mineral más relacionada que aquellos distantes. A partir de esta idea surge el concepto de dependencia espacial.

Definición 1.9. La **dependencia espacial** es el principio fundamental de la geoestadística y hace referencia a la correlación entre observaciones en función de su proximidad. Es decir, la dependencia espacial indica que el valor de una variable en una ubicación está influenciado por los valores de esa misma variable en ubicaciones cercanas. Tobler resume este principio en su Primera Ley de la Geografía: "todo está relacionado con todo lo demás, pero las cosas cercanas están más relacionadas [entre sí] que las cosas distantes" (Tobler, 1970) [22]. Esta dependencia espacial es importante para la construcción de modelos de interpolación espacial, como Kriging, que permite estimar valores en ubicaciones no muestreadas basándose en la dependencia espacial entre puntos.

Ejemplo 1.10. El concepto de dependencia espacial aparece en muchas ramas científicas, desde la minería hasta las ciencias sociales. Un ejemplo de dependencia espacial sería la forma en que se concentran los delitos, como robos o asesinatos, en una ciudad. Los crímenes tienden a formar clusters en zonas específicas, creando los llamados puntos calientes. La dependencia espacial en la distribución de los crímenes muestra la tendencia de los puntos calientes a influir en los índices de criminalidad de las áreas circundantes(Wang et al., 2013) [27].

Pero la dependencia espacial no tiene por qué ser directa, puede ser inversa y de diferente “intensidad”, con el objetivo de obtener una definición cuantitativa precisa de la similitud entre valores en función de la distancia surgen diferentes conceptos. Uno de ellos es el **variograma** que, como se verá, utiliza la semivarianza para dar una medida de la intensidad de la dependencia espacial. También surge el concepto de la **autocorrelación espacial** cuyas medidas como el índice de Moran o el coeficiente de Geary utilizan la correlación para evaluar dicha dependencia.

Por último, de la generalización de la dependencia espacial a todo el espacio muestral surge el concepto de continuidad espacial.

Definición 1.11. La **continuidad espacial** describe la manera en la que los valores de una variable cambian gradualmente en el espacio. En todo el espacio los valores de una determinada variable suelen estar rodeados por valores similares y es que "cuando vemos un valor extremo solitario en un mapa, nuestra intuición nos avisa de que puede haber un error, porque señala una relación inusual con los otros valores"(Isaaks & Srivastava, 1989, p.51) [11]. La existencia de continuidad espacial es clave para la construcción del variograma pues asegura que los valores de la variable de interés no cambian de forma abrupta, lo que permite utilizar puntos cercanos para estimar puntos desconocidos.

1.3. Estacionariedad

Para finalizar esta introducción se definirá el concepto de estacionariedad con el fin de entender su relevancia en geoestadística.

Definición 1.12. La **estacionariedad** en un proceso espacial implica que "las características de la función aleatoria se mantienen constantes cuando se mueve un determinado conjunto de n puntos de una parte de la región a otra"(Wackernagel, 2003, p. 43) [26], equivalentemente se dice que dicha función es invariante por traslaciones. En el contexto de la geoestadística, las propiedades invariantes bajo estacionariedad son las estadísticas de la función, como la media y la varianza.

La estacionariedad es una condición sumamente importante en el estudio de la geoestadística porque permite simplificar la modelización espacial. Como se ha visto, para obtener una medida de la dependencia espacial se usará el variograma, que utiliza la semivarianza para dar una medida de la intensidad de esta dependencia. Bajo la hipótesis de estacionariedad, la semivarianza depende únicamente de la distancia entre puntos independientemente de su ubicación en el espacio, lo que provoca que la dependencia espacial sea homogénea y hace mucho más fácil el cálculo del variograma y la estimación de puntos desconocidos mediante la técnica de interpolación Kriging.

Hasta aquí puede parecer todo perfecto, pues la hipótesis de estacionariedad facilita los cálculos y, a priori, parece que mejora la estimación de puntos desconocidos. Sin embargo la suposición de estacionariedad no es acertada porque no se ajusta a la realidad en la gran mayoría de los casos. En geoestadística pocos fenómenos son realmente estacionarios por la presencia de variaciones espaciales o de tendencias. En este sentido, y como la hipótesis de estacionariedad es

tan útil, en la práctica se busca aproximar la realidad a una forma de estacionariedad, surgiendo diferentes "grados" o tipos de estacionariedad.

1.3.1. Tipos de Estacionariedad

Definición 1.13. El tipo de estacionariedad más restrictiva es la **estacionariedad estricta**, que se caracteriza porque la traslación de la configuración de un conjunto de puntos en cualquier dirección no cambia la distribución múltiple. Expresado matemáticamente:

$$F_{x_1,...,x_n}(z_1,...,z_n) = F_{x_1+h,...,x_n+h}(z_1,...,z_n)$$

Donde $x_1,...,x_n$ es un conjunto de n puntos (con n un número arbitrario) y h un vector de traslación.

La estacionariedad estricta es la más rigurosa y por lo tanto la que más se acerca a la idea de estacionariedad mencionada en el punto anterior, lo que la hace muy poco común en geoestadística. Esta idea se parece a la uniformidad en escala y perfección descrita por Serres en su análisis del sistema de Leibniz: "Es en todas partes y siempre como aquí, en todas partes y siempre como en casa, con ciertos grados de magnitud y perfección"(Serres, 1968, p. 39) [21]. Las cosas no cambian fundamentalmente entre un punto y otro en el espacio, pero la clave está, como menciona Wackernagel [26], en la restricción "con cierto grado". Será esta sutil diferencia la que debilita el concepto de estacionariedad estricta y lleve al estudio de otros grados de estacionariedad.

Observación 1.14. Es importante recalcar que la estacionariedad es una propiedad que afecta a la función aleatoria y no al valor regionalizado, aunque se abuse de la notación y se diga que un valor regionalizado es estacionario, lo que se pretende decir con esto es que esos valores se consideran realizaciones de una función aleatoria estacionaria.

Definición 1.15. Una estrategia menos restrictiva consistiría en considerar como estacionarios solo los dos primeros momentos de la variable, a este tipo de estacionariedad se la conoce como **estacionariedad de segundo orden** o **débil**. Bajo esta suposición se tendrá que tanto la esperanza como la covarianza son invariantes por traslaciones, siendo h un vector y x y $x+h$ dos puntos del dominio:

$$\mathbb{E}[Z(x+h)] = \mathbb{E}[Z(x)], \tag{1.1}$$

$$\text{cov}[Z(x+h), Z(x)] = C(h) \tag{1.2}$$

La media $\mathbb{E}[Z(x+h)] = m$ es constante en todos los puntos del dominio y la covarianza depende solo de la "separación" definida por el vector h.

Observación 1.16. Es un buen momento para explicar la importancia de la media en un punto. Puede parecer contraintuitivo tomar la media del valor de la función aleatoria en un punto x, pues podría parecer que al ser un valor puntual, sobre todo cuando se están modelizando fenómenos físicos, el valor debería ser único. Pero es necesario recordar que en geoestadística se asume que $Z(x)$ no tiene un valor fijo sino que es una variable aleatoria. Aunque en un solo punto x solo se pueda observar un valor de $Z(x)$, el modelo probabilístico permite inferir cómo podría variar $Z(x)$ bajo distintas realizaciones.

Definición 1.17. El tipo de estacionariedad menos restrictiva es la **estacionariedad intrínseca**, de la cual surgirá la noción de variograma. Esta estrategia consiste en asumir que por cada vector h el incremento, es decir, la diferencia entre los valores de pares de puntos x y $x + h$, definido matemáticamente como:

$$Y_h(x) = Z(x + h) - Z(x) \tag{1.3}$$

es una función aleatoria estacionaria de segundo orden.

El variograma teórico $\gamma(h)$ se define precisamente a partir de esta hipótesis intrínseca que se basa, como menciona Wackernagel (2003, p. 51) [26], en dos suposiciones sobre estos incrementos(1.3):

1. Su media $m(h)$, llamada **deriva**, es invariante para cualquier traslación de un vector h en el espacio. Además se supone que la deriva será 0 independientemente de la posición de h.
2. La varianza de los incrementos tiene un valor finito $2\gamma(h)$ dependiendo de la longitud y de la orientación dadas por el vector h, pero no por su posición en el dominio.

Expresado de forma matemática, un modelo intrínseco con dos puntos x y $x + h$ con deriva constante cero sería de la forma:

$$\mathbb{E}[Z(x + h) - Z(x)] = 0, \tag{1.4}$$

$$\mathbb{E}[(Z(x + h) - Z(x))^2] = 2\gamma(h) \tag{1.5}$$

Se llega así a la definición teórica del variograma:

$$\gamma(h) = \frac{1}{2}\mathbb{E}[(Z(x + h) - Z(x))^2] \tag{1.6}$$

En el siguiente capítulo se estudiará su estimación y cómo hallar esta estructura tan importante para la geoestadística de forma experimental.

Observación 1.18. Es importante recalcar que, aunque exista esperanza y varianza en estos incrementos, esto no implica que existan los momentos de la función aleatoria. Una función aleatoria intrínseca puede tener una varianza infinita a pesar de que la varianza de los incrementos sea finita para cualquier vector h.

Capítulo 2

Estimación del Variograma

Antes de adentrarse en la definición formal del variograma, su cálculo y los diferentes modelos de variogramas, es importante hacer un análisis introductorio de otras dos medidas de la dependencia espacial que ayudarán a comprender la relevancia del variograma.

2.1. Introducción a la Medición de la Dependencia Espacial

Definición 2.1. La definición de la **función de covarianza** $C(h)$ surge a partir de la hipótesis de estacionariedad de segundo orden, es decir, bajo la hipótesis de que tanto la esperanza como la covarianza son invariantes por traslaciones. De esta forma se tendrá que, como $\mathbb{E}[Z(x)] = m$, entonces la función de covarianza:

$$C(h) = \mathbb{E}[(Z(x) - m) \cdot (Z(x+h) - m)] = \mathbb{E}[Z(x)Z(x+h)] - m^2 \tag{2.1}$$

La función de covarianza cumple dos propiedades:

1. Es un función par, $C(-h) = C(h)$.
2. Es una función acotada que cumple $|C(h)| \leq C(0) = var(Z(x))$.

Definición 2.2. El **correlograma** o **función de correlación** es una función estrechamente relacionada con la función de covarianza y se puede interpretar como el coeficiente de correlación entre $Z(x)$ y $Z(x+h)$. Se expresa como:

$$\rho(h) = \frac{C(h)}{C(0)} \tag{2.2}$$

Resulta obvio que dicha función está acotada entre -1 y 1.

Es importante entender que tanto el correlograma como la función de covarianza muestran cómo la correlación evoluciona con la separación o *Lag* definida por el vector h. Pero lo mejor para afianzar estos conceptos será poner un ejemplo práctico del cálculo del correlograma y de la función de covarianza.

Ejemplo 2.3. Para este ejemplo se supondrá que se mide la concentración de un elemento específico, el porcentaje de arena de grano grueso, en una región muestral. Para simplificarlo se asumirá que la región es lineal y las mediciones de la concentración de arena se hacen en un punto cada cinco metros a lo largo de una recta, dando como resultado el siguiente gráfico.

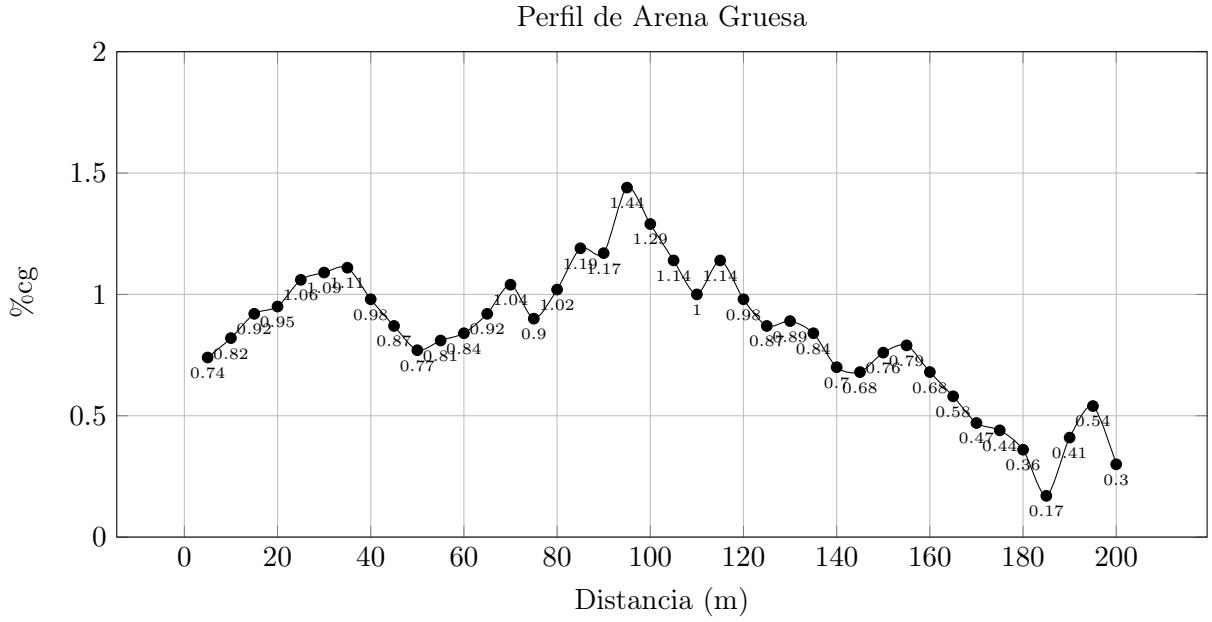


Figura 2.1: Ejemplo extraído del video *What the Heck is a Variogram?* [6]

Con estos datos se puede comenzar a estudiar la dependencia espacial, para ello se observa como varía el porcentaje de arena dependiendo de la separación entre puntos o *Lags*. Imagínese que se toma *Lag 1*, un único grado de separación, que equivale a cinco metros en este ejemplo y *Lag 2*, dos grados de separación, que equivale a diez metros. El objetivo es comprender como se comportan los puntos dependiendo de la distancia a la que se sitúan unos de otros, de esta forma obtenemos las siguientes Figuras 2.2 y 2.3, donde se le llama inicio al punto inicial x y final al punto $x + h$ con la distancia h dependiente del *Lag* del gráfico.

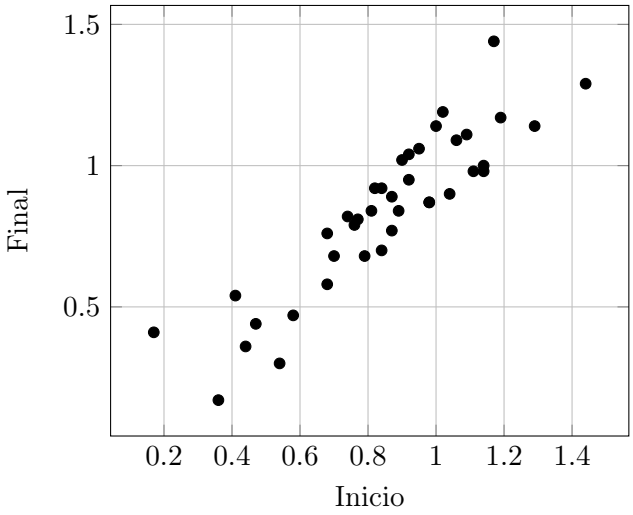


Figura 2.2: Gráfico *Lag 1*

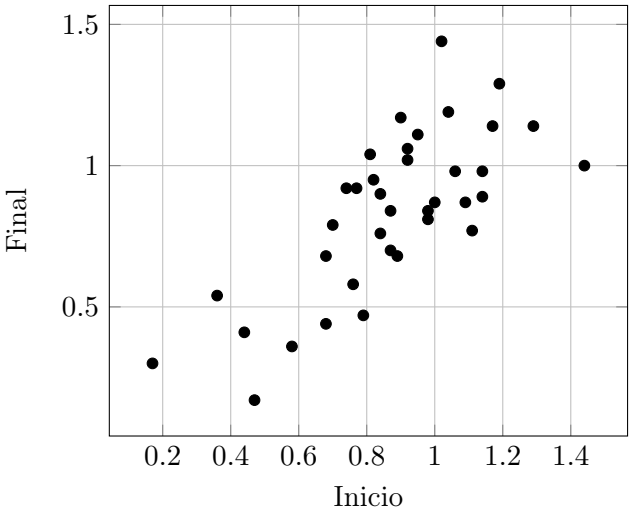


Figura 2.3: Gráfico *Lag 2*

De forma intuitiva a partir de los gráficos de dispersión ya se puede visualizar que a menor separación, *Lag 1*, mayor correlación y a mayor separación, *Lag 2*, menor correlación. Si se hallan los coeficientes de correlación asociados a cada separación, los cálculos se omiten en este trabajo, esta idea intuitiva queda claramente expresada de forma matemática. La Tabla 2.1 refleja el correlograma asociado a la muestra del ejemplo.

Separación	**Lag 1**	**Lag 2**	**Lag 3**	**Lag 4**	**Lag 5**
Coef. Corr.	0.8953	0.8195	0.7419	0.6453	0.5511

Cuadro 2.1: Valores de coeficientes de correlación para diferentes *Lags*.

2.2. Cálculo del Variograma Experimental

Antes de definir y exponer de forma más rigurosa el variograma experimental, es importante recalcar que en geoestadística no siempre tenemos regiones lineales como en el Ejemplo 2.3, en la mayoría de casos estamos en espacios, es decir, en superficies de 2 dimensiones donde las muestras no se toman en una línea recta.

Se explicará ahora la idea intuitiva detrás del variograma experimental. Como el lector puede recordar del Ejemplo 2.3, si se supone que se han filtrado las observaciones en función de su separación, *Lag*, se puede obtener un diagrama de dispersión que representa la relación entre valores medidos a una distancia fija. En dicho diagrama, como el de la Figura 2.4, se tendrá que si las observaciones con esa determinada separación son similares entonces el diagrama de dispersión tendrá la mayoría de los puntos próximos a la recta $Z(x) = Z(x + h)$. Se puede

entender entonces que cuanto más alejado esté un punto de dicha recta, mayor será la variabilidad entre observaciones o, lo que es lo mismo, menor dependencia espacial. Siendo el caso de mayor dependencia espacial posible en el que los puntos del diagrama de dispersión se sitúan sobre la recta $Z(x) = Z(x+h)$ pues implica que las observaciones a esa distancia son exactamente iguales.

A partir de esta idea, y con la ayuda de la Figura 2.4 se podrá entender de dónde surge el concepto de variograma experimental.

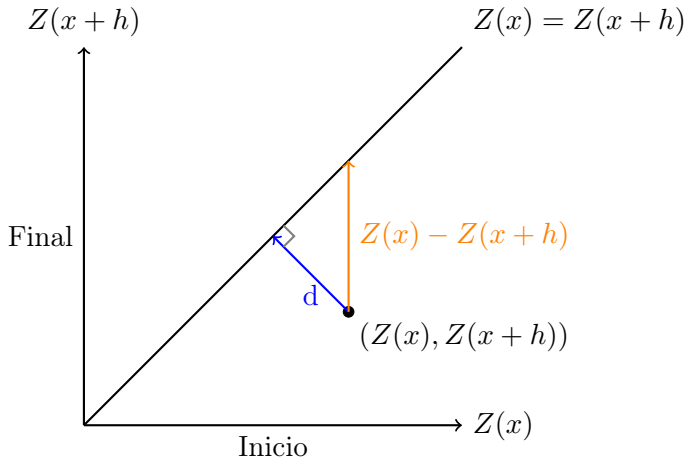


Figura 2.4: Visualización geométrica de la variabilidad espacial entre puntos $Z(x)$ y $Z(x+h)$.

La idea intuitiva del variograma experimental surge del cálculo de la distancia entre el punto $(Z(x), Z(x+h))$ y la recta $Z(x) = Z(x+h)$, para ello basta recordar la relación geométrica entre catetos e hipotenusa. Se sabe que este triángulo es rectángulo isósceles, por lo tanto se cumplirá que $Cateto_1 = Cateto_2$ y por Pitágoras $2d^2 = (Z(x) - Z(x+h))^2$ de donde se concluye que la distancia del punto a la recta es:

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

Como se puede observar, al igual que la función de covarianza, esta distancia es simétrica respecto de h, $d^2(-h) = d^2(h)$, por no depender del signo del vector h al ser una cantidad elevada al cuadrado.

Definición 2.4. Una representación típica de estas distancias, d^2, consiste en graficarlas respecto de la separación o *Lag* definida por el vector h, a este tipo de gráficos se les conoce como **nube de variograma**.

Como se puede observar en la Figura 2.5, este gráfico por sí solo ya es una herramienta muy potente para entender cómo se comporta la dependencia espacial con la separación, pues ano-

malías y heterogeneidades pueden ser fácilmente detectadas simplemente observando si existen grandes desigualdades en pequeñas distancias.

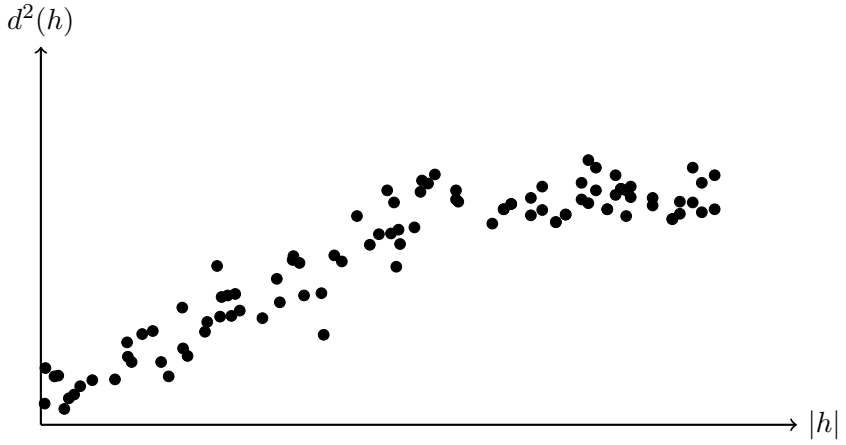


Figura 2.5: Ejemplo de Nube de Variograma.

Definición 2.5. Ahora bien, el **variograma experimental** es la media de las distancias para una determinada separación o *Lag*. Por lo tanto, si tomamos todos los pares de observaciones que se encuentran a dicha distancia fijada h_k, n_c, esta media de las diferencias respecto a una determinada distancia es el valor del **variograma experimental**. Matemáticamente se expresa como:

$$\gamma^*(h_k) = \frac{1}{2n_c}\sum_{\alpha=1}^{n_c}(Z(x_\alpha) - Z(x_\alpha + h_k))^2 \tag{2.4}$$

Observación 2.6. El *Lag*, o vector de separación, no siempre es constante, ya que los puntos de muestreo rara vez se encuentran exactamente a una distancia específica. Por esa razón, con el fin de manejar dicha variabilidad, se agrupan los vectores de separación en intervalos o clases de distancia, $\mathfrak{H}_k$, a las que se asocian los vectores h cuyas longitudes se encuentran dentro de un rango determinado, por lo que el variograma experimental queda expresado como:

$$\gamma^*(\mathfrak{H}_k) = \frac{1}{2n_c}\sum_{\alpha=1}^{n_c}(Z(x_\alpha) - Z(x_\alpha + h))^2 \text{ con } h \in \mathfrak{H}_k \tag{2.5}$$

En la práctica, el variograma experimental se calcula generalmente utilizando vectores h con una longitud inferior a la mitad del diámetro de la región. Para los pares de muestras con vectores h de una longitud casi igual al diámetro de la región, las muestras correspondientes se encuentran cerca del borde lo que provoca que no tengan contribución de muestras ubicadas en el centro de la región y, por lo tanto, no son representativas del conjunto completo de datos.

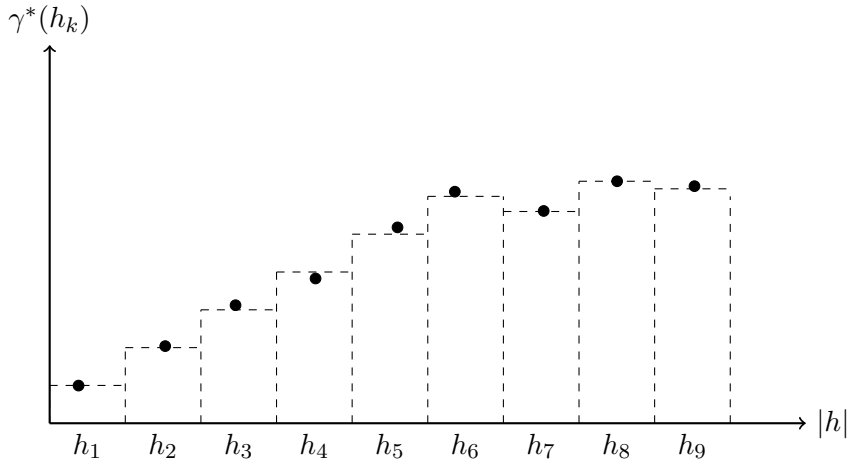


Figura 2.6: Ejemplo de variograma experimental obtenido calculando γ^* para cada clase $\mathfrak{H}_k$.

Un ejemplo de un variograma experimental obtenido para una secuencia de clases $\mathfrak{H}_k$ es el de la Figura 2.6.

Definición 2.7. El **variograma regional** es el variograma experimental ideal calculado cuando el dominio D es perfectamente conocido. Se define como:

$$\gamma_R(h) = \frac{1}{2|D \cap D_{-h}|} \int_{D \cap D_{-h}} [z(x+h) - z(x)]^2 dx, \tag{2.6}$$

donde $D \cap D_{-h}$ es la intersección del dominio D con su traslación, y $|D \cap D_{-h}|$ es su medida (longitud, área o volumen). Este variograma carece de dimensión probabilística y representa una magnitud física independiente de interpretaciones estadísticas o probabilísticas [12].

2.2.1. Factores que Afectan a la Fiabilidad de los Variogramas Experimentales

Es importante entender que factores afectan a la fiabilidad de los variogramas experimentales, porque a partir de ellos se construirá el variograma teórico que se usará para la interpolación espacial, lo que proporciona una idea de posibles limitaciones a la hora de su aplicación en el Capítulo 4.

El primero de los factores que pueden afectar a la fiabilidad es, como en todo modelo estadístico, el tamaño muestral. Si no se tienen suficientes datos con una densidad y con un intervalo de separación adecuado, el resultado puede ser un variograma con poca precisión. En este sentido, el intervalo del muestreo dependerá de la escala del fenómeno físico que se esté estudiando.

Otro factor relevante es la construcción de las clases $\mathfrak{H}_k$. Cuando tenemos datos a intervalos irregulares debemos agrupar las comparaciones por distancias, pero para ello se debe elegir tanto

la longitud h_k como los límites de dichas clases, su "ancho", dentro de las cuales se promediará el variograma experimental. Elegir este ancho es importante, si se generan muchas clases poco anchas entonces se tendrán muchas estimaciones de γ^* lo que, como indica Oliver y Webster (2015), "puede llevar a un variograma "ruidoso" porque las semivarianzas se calculan a partir de pocas comparaciones" [17]. Sin embargo, si las clases son anchas podría acabar por tenerse pocas estimaciones como para revelar la forma del variograma, por eso "la elección [de clases] es un compromiso y no debe automatizarse"(Oliver & Webster, 2015, p. 24) [17].

Por último se tratarán los casos, un poco más complejos, de comportamiento anisotrópico y la existencia de tendencias en los datos.

Definición 2.8. El **comportamiento anisotrópico** surge cuando los cálculos experimentales revelan una variación en el comportamiento del variograma experimental de una dirección a otra. Las funciones habituales que ajustan el variograma teórico se definen para el caso isotrópico y, por lo tanto, no sirven en el caso de situaciones anisotrópicas; por esa razón, se necesitan transformaciones que permitan su uso. En muchos casos, una transformación lineal simple de las coordenadas espaciales es suficiente para convertir una situación anisotrópica en isotrópica.

Un problema también complejo es la existencia de tendencias. Cuando existe una tendencia el variograma experimental deja de ser útil porque, aunque se puede calcular mediante la Ecuación 2.5, no se cumple la suposición de estacionariedad. Para lidiar con este problema se utilizarán otras técnicas como el Kriging Universal, que se tratará en la Subsección 3.2.

2.3. El Variograma Teórico y Sus Características

Definición 2.9. El variograma experimental es reemplazado por el **variograma teórico** en la Figura 2.7, para ello se genera una función teórica del variograma ajustándola a la secuencia de las diferencias medias correspondientes a la Figura 2.6. La importancia del variograma teórico no reside en lo bueno que sea el ajuste que se haga respecto del variograma experimental, la información más importante que aporta es qué tipo de continuidad asume para la variable regionalizada y la hipótesis de estacionariedad asociada a la función aleatoria. Serán estas suposiciones las que guíen la selección de la función de variograma adecuada, por esa razón Wackernagel dice que "el ajuste se realiza a ojo porque generalmente no es tan relevante qué tan bien la función de variograma se ajusta a la secuencia de puntos"(Wackernagel, 2003, p. 59) [26].

Esto coincide con lo señalado por Chiles y Delfiner, quienes afirman que, "desde un punto de vista estadístico, el variograma teórico [...] es una propiedad del proceso permanente que subyace a las observaciones, más que una característica de las fluctuaciones incidentales de la muestra. Es lo esencial en vez de lo anecdótico" (Chiles & Delfiner, 1999, p. 38) [2]. Se discutirá el ajuste

del variograma teórico con más profundidad en la Sección 2.4.

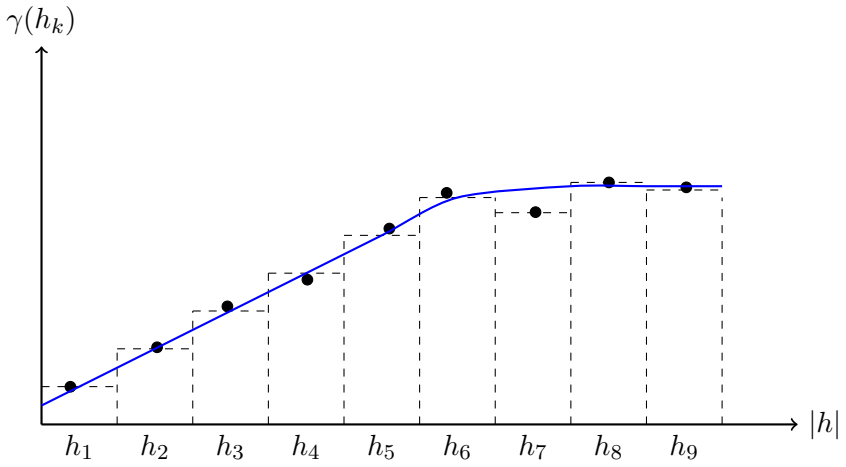


Figura 2.7: Ejemplo de variograma teórico.

El comportamiento del variograma teórico cerca del origen, $h = 0$, es muy importante porque indica el tipo de continuidad de los valores regionalizados. En el caso de la Figura 2.7 es muy claro que es discontinua en el origen, lo que se conoce como efecto nugget.

Definición 2.10. El **efecto nugget** surge cuando el valor del variograma experimental en $h = 0$ es distinto de cero, aunque el valor teórico en dicho punto debería ser 0. Factores como errores en el muestreo y la variabilidad a pequeña escala causan que los valores de las muestras a distancias muy pequeñas sean bastante diferentes, provocando esa discontinuidad en el origen.

Aparte del efecto nugget, otras tres características importantes del variograma son el rango, el umbral y la meseta.

Definición 2.11. Como podemos ver en la Figura 2.7 a medida que aumenta la separación entre pares de puntos, es decir, conforme nos movemos en h, el valor del variograma aumenta para, llegada una determinada distancia, alcanzar una **meseta**. La distancia a la que el variograma alcanza dicha meseta es el **rango**, en la Figura 2.7 el rango se logra en h_6.

Definición 2.12. El **umbral** no es más que el valor que alcanza el variograma, $\gamma(h)$, en la meseta.

La relación entre el efecto nugget y el umbral se denomina efecto nugget relativo y generalmente se expresa en porcentaje.

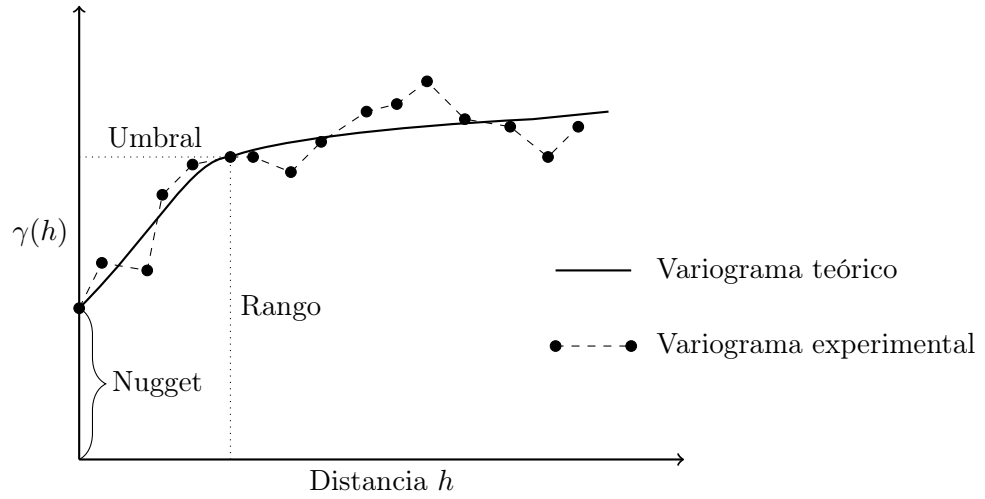


Figura 2.8: Esquema reproducido de Loots et al. *Spawning distribution of North Sea plaice and whiting from 1980 to 2007* (2010) [13].

2.3.1. Relación entre el Variograma y la Función de Covarianza

Como ya se vio en la Sección 1.3.1, se consideran dos clases de funciones aleatorias, funciones aleatorias estacionarias y funciones aleatorias intrínsecas. En general cuando se mencionan las funciones aleatorias estacionarias se hace referencia a la estacionariedad de segundo orden. Es importante recordar esto porque la relación entre el variograma, $\gamma(h)$, y la función de covarianza, $C(h)$ depende del tipo de estacionariedad que se asuma que cumplen los datos.

En el caso en que el proceso aleatorio $Z = \{Z(x) : x \in \mathbb{R}^d\}$ sea estacionario, la covarianza, $C(h)$, depende únicamente del vector de separación h y no de la ubicación de los puntos x y $x+h$. Bajo esta suposición, la relación entre el variograma y la función de covarianza es la siguiente:

$$\gamma(h) = C(0) - C(h), \quad h \in \mathbb{R}^d \tag{2.7}$$

Ahora bien, es importante recalcar que dicha relación 2.7 solo se cumplirá si el variograma está acotado. Si el variograma no está acotado dicha relación se puede mantener de forma local para un determinado r positivo.

$$\gamma(h) = C(0) - C(h), \quad h \in \mathbb{R}^d, |h| \leq r, \tag{2.8}$$

Entonces se dice que $C(h)$ es una **función de covarianza estacionaria localmente equivalente**, esta forma de covarianza "tiene interés tanto teórico como computacional, porque permiten una predicción óptima y una simulación rápida" (Gneiting, Sasvári, & Schlather, 2000) [8]. En este caso se puede interpretar el variograma $\gamma(h)$ como una estructura que mide la pérdida

de similitud entre valores conforme aumenta la distancia entre ellos. Para ilustrar mejor este concepto, consideremos el siguiente ejemplo:

Ejemplo 2.13. Imagínese que se quiere estudiar los patrones de lluvia en una región montañosa. Esta región tiene valles, picos y otro tipo de accidentes geográficos que afectan a la distribución de las precipitaciones, lo que hace que las propiedades estadísticas de la lluvia no sean las mismas en toda la región. Ahora bien, si se toma un área más pequeña, un determinado valle o montaña, se puede asumir que en dicha subregión la cantidad de lluvia es una variable estacionaria. En este caso, la función de covarianza estacionaria localmente equivalente permitiría modelar la cantidad de precipitación dentro de esta subregión de manera precisa.

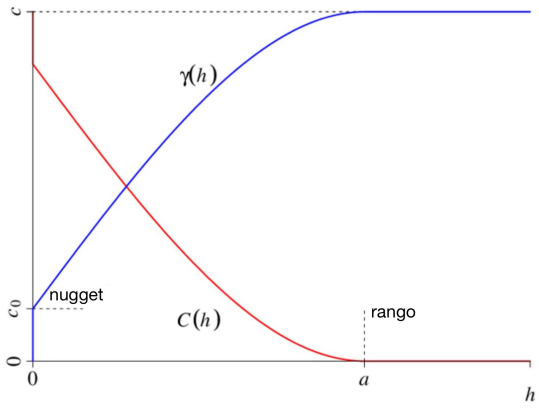


Figura 2.9: Esquema del variograma y función de covarianza de un proceso estacionario de segundo orden. Reproducido de Raimon Tolosana Delgado, *Geostatistics for constrained variables: positive data, compositions and probabilities. Applications to environmental hazard monitoring* (2005) [23].

Se verá ahora el caso de estacionariedad intrínseca, aunque en este caso no es necesario definir la función de covarianza, si el variograma está acotado por un valor finito, existe $\gamma(\infty)$, dicha función de covarianza se puede definir como:

$$C(h) = \gamma(\infty) - \gamma(h) \tag{2.9}$$

En este caso el variograma no representa la pérdida de similitud en términos de una covarianza explícita (porque, para empezar, no existe una función de covarianza bien definida), sino que describe la variabilidad promedio de las diferencias entre los valores del proceso separados por una distancia h.

2.4. Modelos de Variograma Teórico

Los modelos de variograma para situaciones anisotrópicas se derivan de los modelos isotrópicos, por esa razón se empezará con la exposición de estos últimos. Entre los modelos más comunes del variograma teórico para el caso isotrópico se incluye el modelo esférico, el exponencial y el Gaussiano.

Se empleará la relación, $\gamma(h) = C(0) - C(h)$, presentada en la Subsección 2.3.1 para representar gráficamente los modelos de variograma teórico a partir de las funciones de covarianza.

2.4.1. Modelo Efecto-nugget

Este modelo representa una discontinuidad en el origen del variograma teórico, $\gamma(h = 0) > 0$. Este comportamiento aparece en situaciones donde los errores de medición o la variabilidad a pequeña escala son importantes.

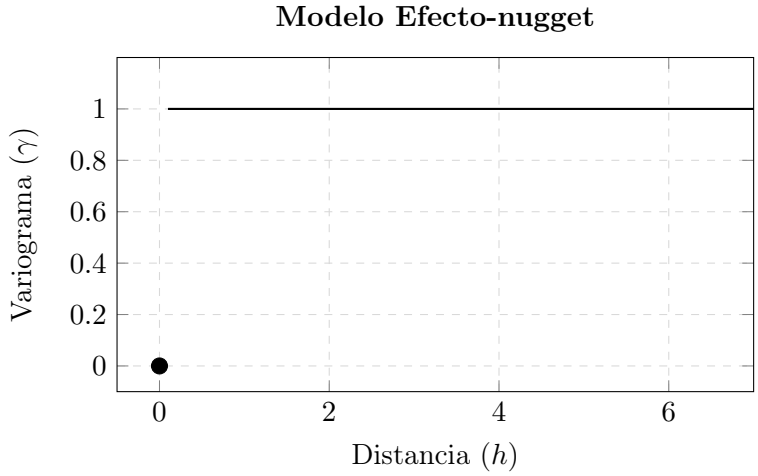


Figura 2.10: Un modelo de variograma efecto-nugget, su valor es cero en el origen y 1 en cualquier otro lugar (Wackernagel, 2003, p.58) [26].

Aunque no se suele utilizar como único modelo, el efecto nugget suele incluirse en otros modelos teóricos (esférico, exponencial o gaussiano) para reflejar esta propiedad de los datos. Expresado en forma de función de covarianza $C(h)$:

$$C_{\text{nug}}(h) = \begin{cases} 0 & \text{si } |h| = 0, \\ b & \text{si } |h| > 0. \end{cases}$$

Donde b es un valor positivo y h es la distancia.

2.4.2. Modelo Esférico

Este modelo representa un variograma que aumenta rápidamente al principio, reflejando una fuerte correlación entre puntos cercanos, y luego se aplana al alcanzar el rango, indicando la falta de correlación para los puntos que están más lejos de esa distancia. Su fórmula es:

$$C_{\text{sph}}(h) = \begin{cases} b\left(1 - \frac{3|h|}{2a} + \frac{1}{2}\frac{|h|^3}{a^3}\right), & \text{para } 0 \leq |h| \leq a, \\ 0, & \text{para } |h| > a. \end{cases} \tag{2.10}$$

Donde h es la distancia, a es el rango y b es el parámetro que representa el máximo valor de la covarianza. El modelo esférico es especialmente útil para describir fenómenos espaciales donde la correlación disminuye rápidamente y luego se estabiliza, como es el caso de suelos, recursos naturales o propiedades geológicas.

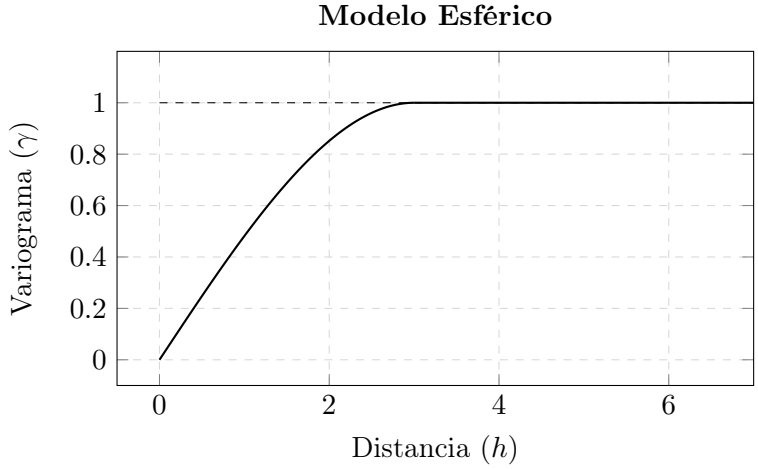


Figura 2.11: Variograma esférico con meseta $b = 1$ y rango $a = 3$, definido como $\gamma(h) = C(0) - C(h)$ a partir de la función de covarianza del modelo esférico (2.10)(Wackernagel, 2003, p.60) [26].

Se puede considerar que el Modelo Efecto-nugget es un caso especial de la función de covarianza esférica con un rango infinitamente pequeño. Aún así, "hay una importante diferencia entre los dos modelos, $C_{\text{nug}}(h)$ describe un fenómeno discontinuo [...] mientras que $C_{\text{sph}}(h)$ representa un fenómeno continuo" (Wackernagel, 2003, p.58) [26].

2.4.3. Modelo Exponencial

El modelo exponencial representa, como su mismo nombre indica, un crecimiento exponencial y asintótico hacia la meseta b, sin alcanzar un rango definido. Este modelo es adecuado para describir estructuras espaciales con correlación a corta distancia. Para un valor de $|h| = 3a$ el

variograma se aproxima al 95 % al valor de la meseta, a esta distancia se le conoce como el **rango práctico** del modelo exponencial. La fórmula de este modelo es:

$$C_{\exp}(h) = b \exp\left(-\frac{|h|}{a}\right) \quad \text{con } a, b > 0. \tag{2.11}$$

El modelo exponencial es útil en casos donde se espera que la correlación espacial disminuya rápidamente pero sin un rango claramente definido. Se aplica en estudios medioambientales, como análisis de calidad del aire o del agua.

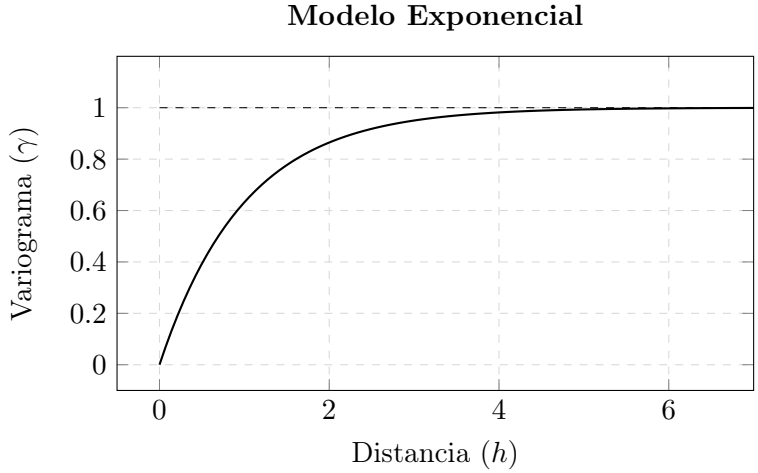


Figura 2.12: Representación del modelo de variograma exponencial. La curva crece asintóticamente hacia la meseta $b = 1$, el parámetro de rango es $a = 1$ con un rango práctico de $|h| = 3$. Esto implica que el modelo se ha acercado a la meseta en un 95 % (Wackernagel, 2003, p.59) [26].

Observación 2.14. Es importante recalcar que, aunque en el modelo esférico a representa el valor del rango, en el caso del modelo exponencial y también en el caso del modelo gaussiano, a representa un parámetro de escala, dado que estos modelos no tienen un rango definido.

2.4.4. Modelo Gaussiano

El modelo gaussiano se caracteriza por un variograma que crece de manera suave hacia la meseta b con un comportamiento regular, como se puede observar en la Figura 2.13. Su variograma aumenta lentamente al principio y luego de manera más rápida antes de alcanzar la meseta. Esto lo hace útil para describir fenómenos con alta correlación en distancias cortas y transiciones graduales hacia la independencia espacial.

La fórmula del modelo gaussiano con parámetro de escala $a > 0$ se define a partir de la

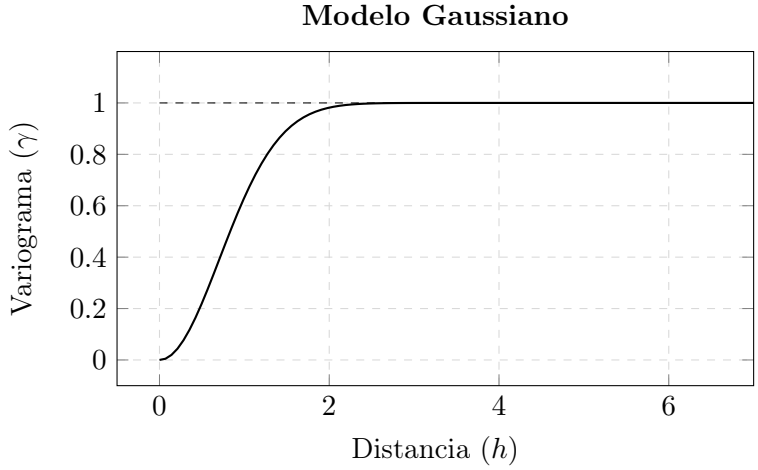


Figura 2.13: Representación del modelo de variograma Gaussiano. La curva crece suavemente hacia la meseta $b = 1$. El rango práctico está definido por la función de covarianza (2.12) con un parámetro de escala $a > 0$(Chiles & Delfiner, 1999, p. 83) [2].

siguiente función de covarianza:

$$C_{\text{Gauss}}(h) = \exp\left(-\frac{h^2}{a^2}\right) \tag{2.12}$$

Este modelo, al igual que en el caso del modelo exponencial, tiene un rango práctico. En este caso, el variograma se aproxima al 95 % de la meseta para una distancia $h = 1.73a$.

Este modelo se utiliza en meteorología, geociencias y batimetría, estudio de las superficies del fondo marino. Es importante mencionar que aparte de los modelos aquí expuestos existen muchos más modelos, tanto derivados de los modelos presentados, como modelos diferentes como el modelo K-Bessel, Cauchy Generalizado, $|h|^\alpha$... El ajuste del variograma teórico es una materia realmente profunda y con muchas ramificaciones, todas las cuales no se pueden tratar en este trabajo.

2.5. Modelado del Variograma Teórico en Situaciones Anisotrópicas

Primero, recuérdese en qué consiste una situación anisotrópica, mencionada en la Subsección 2.2.1. El comportamiento anisotrópico no es más que la diferencia en la variación espacial según la dirección. En ese caso, la dependencia espacial no se comporta de la misma manera en todas las direcciones, lo que provoca que sea necesario realizar transformaciones en las coordenadas. A continuación, se expondrán dos tipos de anisotropías.

2.5.1. Anisotropía Geométrica

La idea intuitiva detrás del modelo de variograma teórico para el caso de anisotropía geométrica surge de representar el variograma experimental trazando un mapa de contornos del variograma como función de un vector h.

Las isolíneas, por definición, representan curvas que conectan puntos con un mismo valor. Son parecidas a las curvas de nivel de un mapa topográfico que conectan los puntos que están a una misma altura. Pero, en este caso, las isolíneas unen los puntos con un mismo valor del variograma experimental, es decir, con una misma dependencia espacial.

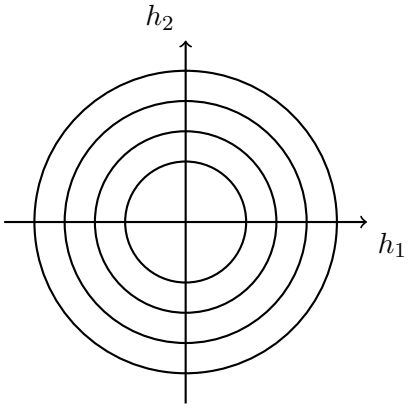


Figura 2.14: Representación de una situación isotrópica con isolíneas en un espacio de dos dimensiones.

La situación ideal es en la que las isolíneas son circulares alrededor del origen. En este caso, se estaría ante una situación isotrópica, Figura 2.14, pues el variograma solo dependería de la distancia respecto del origen y no de su dirección.

La definición general para que un variograma en $\mathbb{R}^n$ represente la anisotropía geométrica es de la forma:

$$\gamma(h) = \gamma_0 \left(\sqrt{\mathbf{h'Qh}} \right) \tag{2.13}$$

Donde γ_0 es el modelo isotrópico, $\mathbf{Q}$ es una matriz definida positiva $n \times n$ que se encarga de caracterizar la estructura de la anisotropía en el espacio $\mathbb{R}^n$ y $\mathbf{h}$ es un vector que representa las componentes $\{h_i : i = i, ..., n\}$ del vector distancia h.

El conjunto de autovalores y autovectores asociados a la matriz $\mathbf{Q}$ definen un nuevo sistema de coordenadas ortogonales donde los autovalores b_i representan la "intensidad" de la anisotropía en cada dirección, y los autovectores v_i representan las direcciones principales.

Por lo tanto, se realizan dos transformaciones simultáneas. Un cambio de coordenadas, donde

las direcciones principales (autovectores de $\mathbf{Q}$) se convierten en los ejes del nuevo sistema de coordenadas, "alineando" el espacio con los ejes principales de anisotropía. Y un escalado de las coordenadas tras el cambio de sistema, las nuevas coordenadas $\widetilde{h_i}$ todavía reflejan la anisotropía, basta tomar $\hat{h_i} = b_i\widetilde{h_i}$ para restaurar la isotropía. De forma más intuitiva:

$$\mathbf{h'Qh} = \mathbf{h'}\left(\sum_{i=1}^{n} b_i^2 v_i v_i'\right)\mathbf{h} = \sum_{i=1}^{n} b_i^2(\mathbf{h'}v_i)^2 = \sum_{i=1}^{n} b_i^2\widetilde{h_i}^2 = \sum_{i=1}^{n}\hat{h}_i^2.$$

Teniendo en cuenta que $\mathbf{Q}$ es definida positiva, se tendrá que $\mathbf{Q} = \mathbf{A'A}$ por lo tanto, se puede expresar el variograma como:

$$\gamma(h) = \gamma_0\left(\sqrt{\mathbf{h'Qh}}\right) = \gamma_0\left(\sqrt{\mathbf{h'A'Ah}}\right) = \gamma_0\left(|\mathbf{Ah}|\right)$$

Esto significa que se pueden transformar las coordenadas del espacio anisotrópico al espacio isotrópico mediante la matriz $\mathbf{A}$, que reescala y rota las direcciones principales de anisotropía. Con estos conceptos claros, ya se pueden entender los casos de anisotropía en dos y tres dimensiones.

Anisotropía geométrica en $\mathbb{R}^2$

Si se tiene una situación en $\mathbb{R}^2$ en la que la isotropía se puede conseguir mediante una rotación y un reescalado, la estructura de las isolíneas presenta forma elíptica, la matriz de transformación correspondiente es de la siguiente forma:

$$\mathbf{A} = \begin{bmatrix} b_1 & 0 \\ 0 & b_2 \end{bmatrix}\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

La primera matriz es la encargada de llevar a cabo el reescalado de las coordenadas según los factores b_1 y b_2, mientras que la segunda matriz rota el sistema en un ángulo θ.

Como ya se vio en la sección anterior, los valores $b_i (i = 1, 2)$ representan lo rápido que crece el variograma en las direcciones principales. Por esta razón, los rangos a_i, que indican hasta qué distancia los puntos siguen correlacionados, se definen como su inverso: $a_i = \frac{1}{b_i}$.

Por lo tanto, se tendrá que el rango en la dirección θ será $a_1 = \frac{1}{b_1}$ y en la dirección perpendicular $\theta + \frac{\pi}{2}$ será $a_2 = \frac{1}{b_2}$. Estos valores son importantes porque son los rangos a_1 y a_2 los que normalmente cuantifican la intensidad de la anisotropía mediante la relación $\frac{a_1}{a_2}$ que se conoce como **radio de anisotropía**.

La Figura 2.15 ilustra cómo el sistema de coordenadas original $h = (h_1, h_2)$ se transforma en un nuevo sistema h' alineado con los ejes principales de las elipses concéntricas. Esta rotación y el posterior reescalado permiten analizar la anisotropía geométrica de manera más clara.

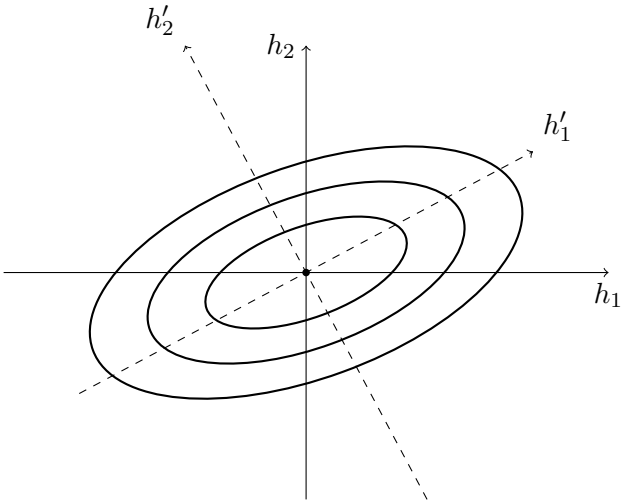


Figura 2.15: Transformación del sistema de coordenadas $h = (h_1, h_2)$ al sistema $h' = (h_1', h_2')$, alineado con los ejes principales de las elipses concéntricas(Wackernagel, 2003, p.63) [26].

Anisotropía geométrica en $\mathbb{R}^3$

El caso de la anisotropía en $\mathbb{R}^3$ se soluciona mediante una composición de rotaciones elementales. En este caso, las isolíneas conforman una superficie elipsoidal que puede transformarse mediante rotaciones y reescalado para convertirse en una esfera isotrópica, donde la dependencia espacial solo dependa de la distancia.

$$\mathbf{A} = \begin{bmatrix} b_1 & 0 & 0 \\ 0 & b_2 & 0 \\ 0 & 0 & b_3 \end{bmatrix} \begin{bmatrix} \cos\theta_3 & \sin\theta_3 & 0 \\ -\sin\theta_3 & \cos\theta_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_2 & \sin\theta_2 \\ 0 & -\sin\theta_2 & \cos\theta_2 \end{bmatrix} \begin{bmatrix} \cos\theta_1 & \sin\theta_1 & 0 \\ -\sin\theta_1 & \cos\theta_1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

El ángulo θ_1 define una rotación del plano h_1h_2 alrededor de h_3, de modo que h_1 se lleva al plano $h_1'h_2'$. Con θ_2, se realiza una rotación alrededor de la intersección de los planos h_1h_2 y $h_1'h_2'$, llevando h_3 a la posición de h_3'. La tercera rotación, con un ángulo θ_3, rota todo alrededor de h_3' hasta su posición final.

El reescalado se lleva a cabo al igual que en el caso de anisotropía geométrica en $\mathbb{R}^2$ mediante los factores b_1, b_2 y b_3.

2.5.2. Anisotropía Zonal

Este tipo de anisotropía se caracteriza porque, aunque la dependencia espacial varía según la dirección, dicha variación no implica un reescalado o deformación de las distancias. La existencia de anisotropía solo afecta al nivel de variación, la varianza, del fenómeno dependiendo de la

dirección.

De forma más intuitiva, en la anisotropía zonal el rango, la distancia hasta la que existe correlación entre puntos, puede ser igual en todas las direcciones, pero la intensidad de la variación (umbral), el valor del variograma a la distancia del rango, cambia según la dirección.

En estas situaciones la forma más habitual de lidiar con la anisotropía es "dividir" el variograma en dos o más componentes dependiendo del valor que alcanza el umbral en las diferentes direcciones. El ejemplo en dos dimensiones es bastante ilustrativo, imagínese que se tiene un umbral mucho mayor en x_2 que en x_1. En este caso se puede generar un primer variograma experimental, $\gamma_1(h)$ en la dirección x_1 que será isotrópico. Y para modelar la diferencia de umbral se toma el variograma experimental en la dirección x_2 como $\gamma_2(h)$ que se define de tal forma que no tenga ningún efecto en la coordenada x_1. De esta forma se construye el modelo de variograma como:

$$\gamma(h) = \gamma_1(h) + \gamma_2(h) \tag{2.14}$$

En la práctica, raramente se encuentra una anisotropía zonal pura; es más común encontrar una mezcla de anisotropías zonal y geométrica juntas. En este sentido y al igual que en el caso del estudio de modelos de variogramas, existen más modelos de anisotropía que los presentados en este trabajo, como la estratificada, la periódica, etc ... Así como combinaciones de los mismos, la geoestadística es una rama de las matemáticas que se puede complicar al intentar modelar una realidad muy compleja.

2.6. Estimación del Variograma: Caso práctico

Para ilustrar el proceso descrito anteriormente, se estimará el variograma experimental y se ajustará un modelo teórico para el caso de estudio presentado en este trabajo utilizando datos reales de contaminación ambiental en Galicia . Este cálculo forma parte del caso práctico que se presenta en el Capítulo 4. En dicho capítulo se ofrece una descripción completa de los datos utilizados con sus características y contexto, además se incluye su exploración inicial y se presentan los resultados de su interpolación.

Sin embargo, esta sección se limitará solo al cálculo práctico del variograma a partir de dichos datos, para ello se hará uso del libro de Bivand et al. (2008) [1], así como del software estadístico R y de su libreria gstat.

2.6.1. Análisis Exploratorio del Variograma

Como se expone en la presentación de los datos, en este caso práctico se tomarán tres contaminantes y se estudiará la variación de su distribución espacial tomando la media por hora de concentración de contaminante por estación. Como eso proporciona una gran cantidad de datos, el análisis aquí presentado se centrará en el comportamiento de los tres contaminantes principales en una única hora, las 07 : 00.

Después de un análisis inicial, se decidió utilizar la transformación logarítmica de los datos porque mejora la correlación al reducir el impacto de valores atípicos. Además se opto por una distancia, *lag*, de 5 km. Aunque en un primer momento se tomó una distancia mayor de 10 km los datos mostraron una correlación espacial más débil. Finalmente se presentan los diagramas de dispersión resultantes que servirán de base para el cálculo del variograma experimental posterior:

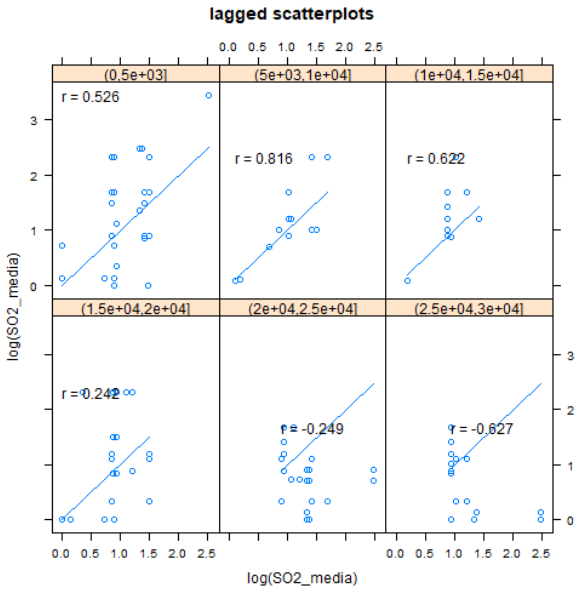


Figura 2.16: Diagrama de dispersión para SO_2 (log).

Los gráficos de dispersión para los contaminantes NO_X, PM_{10} y SO_2, a una distancia, o *lag*, de 5 km, muestran diferencias importantes en su comportamiento espacial. En el caso de SO_2 presenta una alta correlación en distancias cortas mientras que su influencia disminuye significativamente con la distancia, hasta convertirse en negativa.

Para PM_{10}, la correlación es más débil y menos consistente. Por último en el NO_X, se observa una correlación prácticamente nula entre los 0 y los 5 km pero que aumenta entre los 5 y los 10

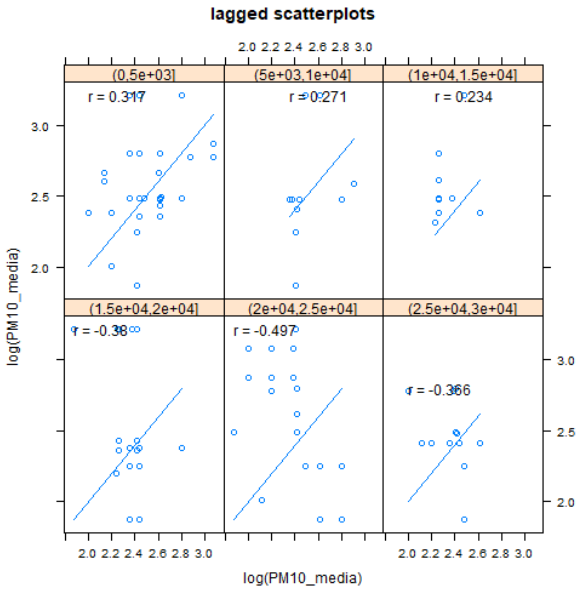


Figura 2.17: Diagrama de dispersión para PM_{10} (log).

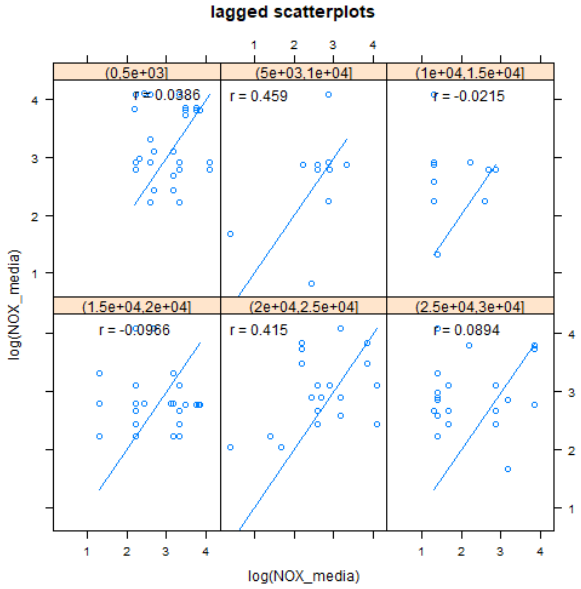


Figura 2.18: Diagrama de dispersión para NO_X (log).

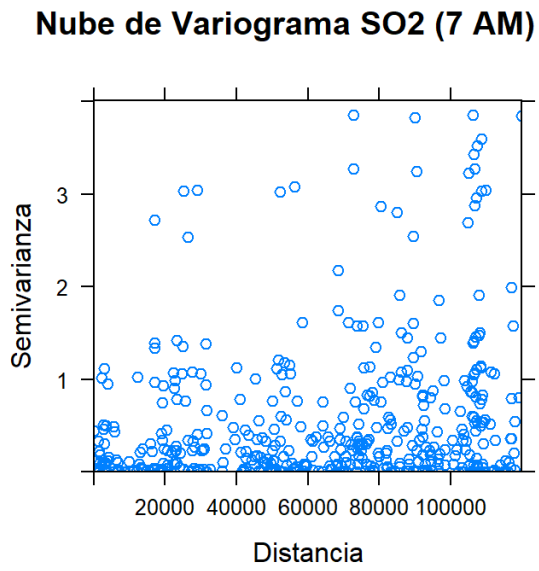
km para volver a debilitarse y volverse negativa después.

Estos gráficos indican una dependencia espacial que varía según el contaminante. En el caso de SO_2, la fuerte dependencia espacial a distancias cortas lleva a pensar en un variograma experimental con un rango corto. Para PM_{10}, la dependencia espacial es más débil, lo que podría provocar un variograma menos definido. Por último, para NO_X, las oscilaciones en la

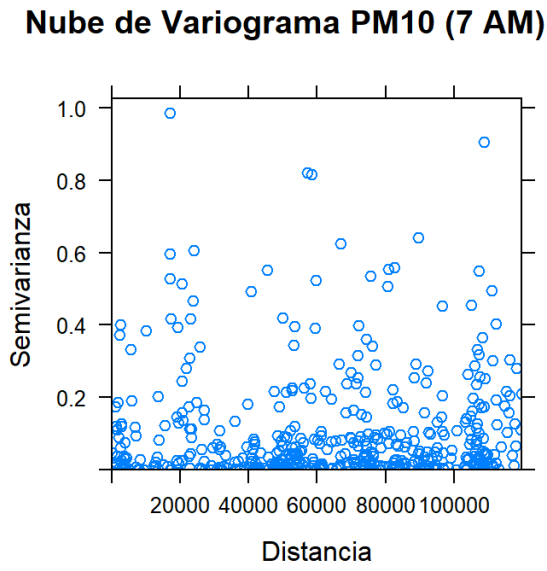
correlación podrían indicar la presencia de anisotropía o tendencias que podrían requerir ajustes en el modelo teórico del variograma. Para comenzar con el estudio del variograma se presentará la nube de variograma de cada uno de los contaminantes, para ello se utilizarán las siguientes funciones de la libreria gstat en R.

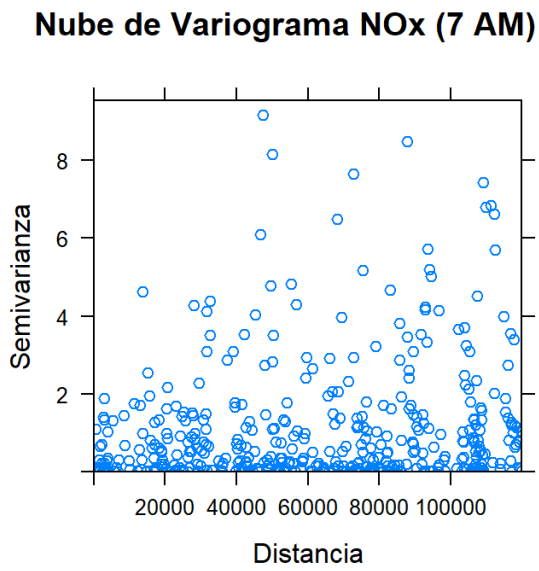
```
#Ejemplificamos para el caso del SO2 la Nube de Variograma
variogram(log(SO2_media) ~ 1, data_7am_so2, cloud = TRUE,cutoff = 120000)
```

Es importante recalcar el porque del *cutoff* tomado. La variable *cutoff* fija la distancia máxima a la que se calculan los pares de puntos para estudiar la dependencia espacial. La comunidad autónoma gallega tiene una superficie de entorno a $30000km^2$ [4], si se supone que su área es cuadrada, aunque no es una aproximación del todo exacta, se tiene como resultado un cuadrado de lado $173km$. Pero, como se vio en esta sección, no se debe tomar una distancia máxima del tamaño de la superficie por lo tanto se decide tomar dos tercios de la misma, $120km$. De esta forma se obtienen las siguientes nubes de variograma para cada contaminante a las $07:00$.



Como ya se puede observar en la Figura 2.6.1 y se podrá ver en los gráficos de la nube de variograma siguientes, aunque con la distancia tenemos puntos con una semivarianza mayor, hay otros puntos que mantienen una semivarianza baja a pesar de la distancia. Esto se puede deber a una homegeneidad en los datos, debido, por ejemplo, a que las fuentes que generan estos contaminantes están distribuidas de manera más o menos uniforme. También puede deberse a





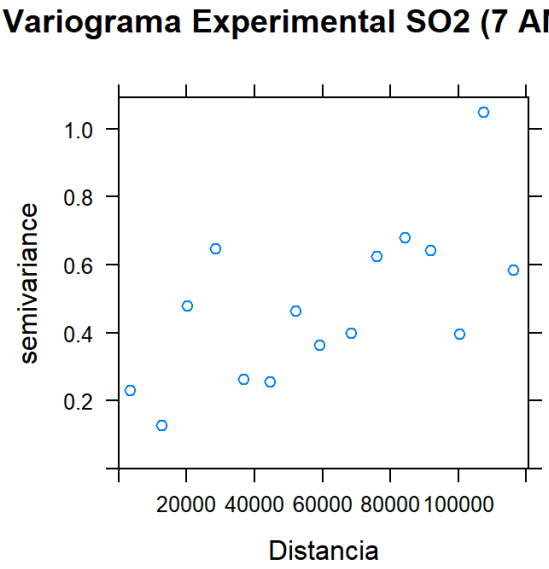
que el transporte a través del aire de los contaminantes provoca que su concentración tienda a estabilizarse en toda la región.

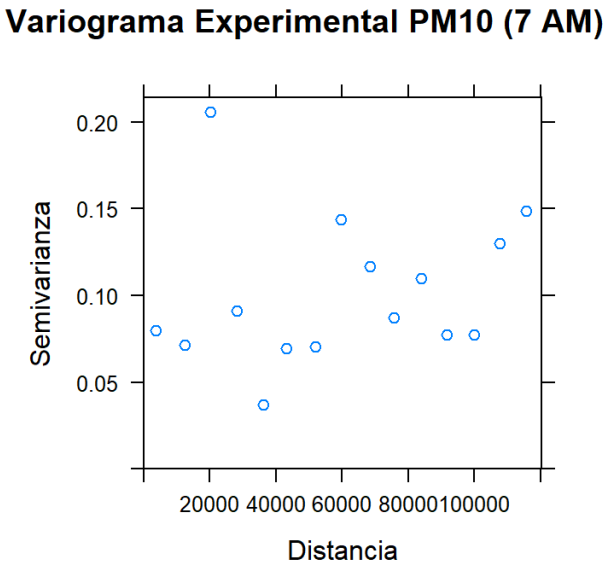
Véase ahora el variograma experimental, para su calculo se utilizará el siguiente código:

```
#Ejemplificamos para el caso del SO2 el Variograma experimental
vario_sample_so2 <- variogram(log(SO2_media) ~ 1, data_7am_so2,cutoff = 120000)
```

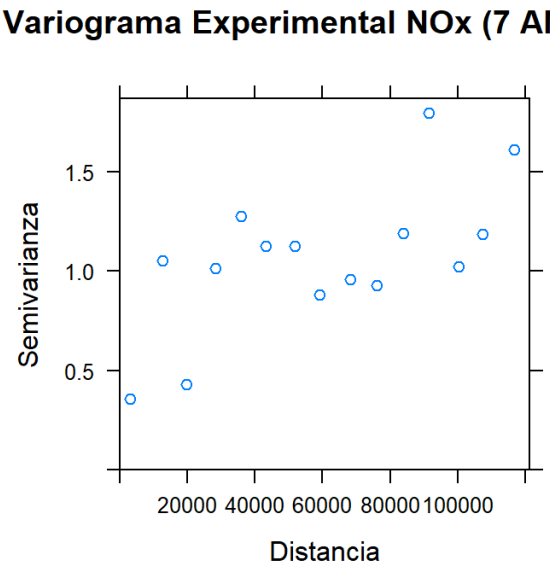
Como en el caso de la nube de variograma en este caso es importante mencionar que la función variogram toma un *width* o separación determinada, que se corresponde con dividir la

distancia máxima, el *cutoff*, entre 15 [19].





Es importante destacar que en estos casos además de las decisiones en cuanto a la distancia máxima, *cutoff*, y la separación, *width*, la función variogram toma otras decisiones, como ignorar la dirección.



2.6.2. Modelado del Variograma Teórico

Antes de presentar los modelos de variograma por los que se ha optado para modelizar cada uno de los contaminantes es importante entender como funciona la expresión *fit variogram* de la librería gstat de R.

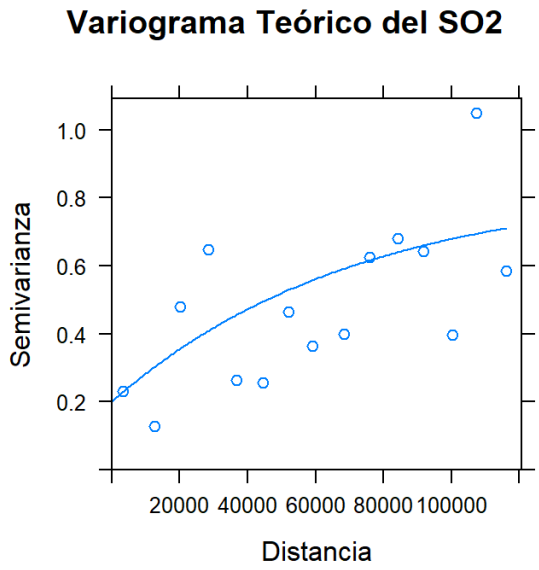
Pero antes es más importante hablar, como se mencionó en la Subsección 2.3, sobre la elección del modelo del variograma. Esta decisión ha de ser intencionada y basarse en las características de los datos del estudio, más que en el aspecto del variograma experimental. En este sentido la bibliografía menciona los tres modelos de variogramas presentes en este trabajo, Sección 2.4. El modelo esférico se menciona en Van Zoest et al. (2019) [25] para el modelado del contaminante aéreo NO_2 en áreas urbanas, el modelo exponencial se presenta en Gräler et al. (2013) [10] para modelar el contaminante PM_{10} en el aire de Europa, y, por último, el modelo gaussiano se menciona en Rivera-González et al. (2015) [20] donde se estudia la distribución de diferentes contaminantes, entre ellos el SO_2, el NO_2 y el PM_{10}; utilizando los tres modelos, el esférico, el exponencial y el gaussiano.

De esta forma y con el apoyo de la bibliografía podemos constatar que cualquiera de estos tres modelos podría utilizarse para el calculo del variograma teórico, en este momento es cuando entra en juego el funcionamiento de la expresión *fit variogram*. Esta función lo que hace es ajustar un modelo teórico del variograma de forma que se ajuste lo máximo posible al variograma experimental. Para ello es necesario presentarle una hipótesis inicial sobre la forma del variograma, que se puede extraer visualmente a partir de la bibliografía mencionada, que proporciona una base sobre los modelos disponibles, y los gráficos del variograma experimental. De esta forma,

con este código:

```
#Ajustamos el variograma teórico al variograma muestral para el SO2
modelo_variograma_so2 <- fit.variogram(vario_sample_so2,
              model = vgm(psill = 1,           #Valor del sill
              model = "Exp",          #Modelo Exponencial
              range = 50000,          #Rango
              nugget = 0.2)           #Valor del nugget
)
```

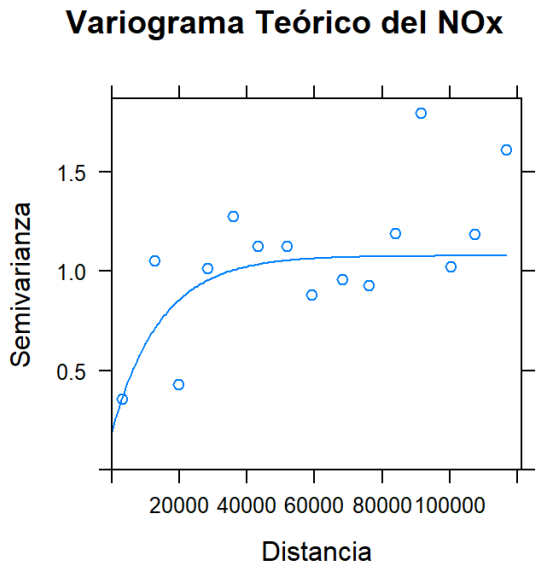
Se llega a los siguientes gráficos:



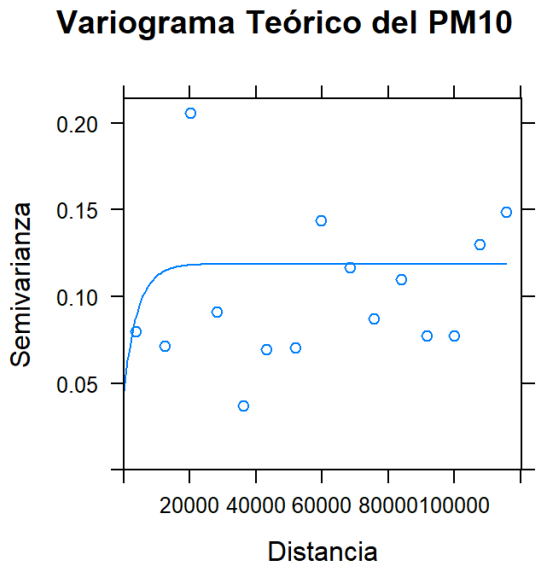
Observación 2.15. Es importante remarcar que en el caso del NO_x se tomó también el modelo exponencial, pero se omite su código en este trabajo.

El problema surge en este caso en el variograma experimental del PM_{10}, que no se ajusta correctamente a ninguno de los tres modelos presentados en la bibliografía. La propia función de R nos informa del error mediante el mensaje *Warning message: No convergence after 200 iterations: try different initial values?*

En esta situación se pueden considerar dos opciones, la primera es la planteada por Bivand et al. (2008, p.210) [1], que sugiere el "argumento para defender el ajuste visual antes que el numérico puede ser que la persona que realiza el ajuste tenga conocimiento más allá de la información en los datos", es decir, realizar un ajuste de forma visual basado en la experiencia del analista. La



segunda opción es utilizar el modelo exponencial debido a su uso en estudios similares. En la literatura consultada, específicamente en estudios sobre distribución espacial de contaminantes, el modelo exponencial se usa para modelar el contaminante PM_{10}, Gräler et al. (2013) [10]. Está segunda opción es la que se considerará en este trabajo.



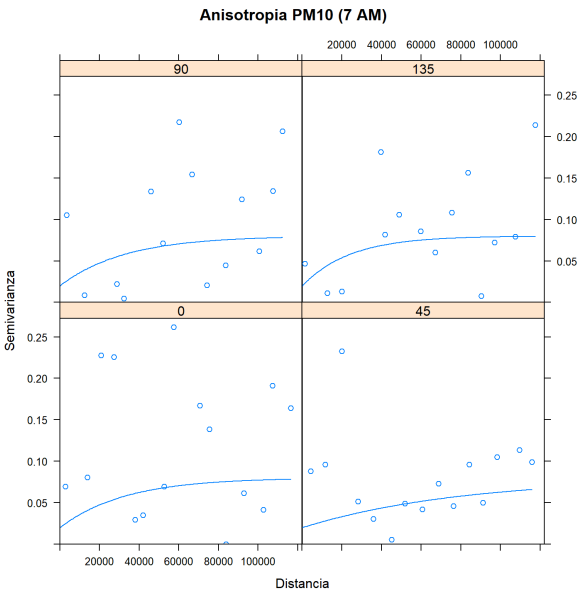
2.6.3. Anisotropía

El comportamiento poco consistente del variograma teórico del PM_{10} lleva a plantearse si las hipótesis sobre las que se fundamente el análisis de dicho contaminante son firmes. Las dos hipótesis principales en cualquier estudio geoestadístico son la dependencia espacial y la isotropía. Debido a su estructura se van a realizar un par de test para comprobar si el PM_{10} satisface dichas hipótesis, para ello se usará la librería sm de R.

Antes de usar esta librería es de interés ver si el comportamiento varía según la dirección, para ello con el siguiente código:

```
model_anis_pm10<-vgm(psill=0.06, model = "Exp", range = 80000, nugget = 0.02, anis=c(45,0.3))
```

Se obtiene el siguiente gráfico:



En él se puede observar cierta diferencia en el comportamiento del variograma, sobre todo en la dirección de 45 grados. Pero, para confirmarlo, se realiza el siguiente test:

```
library(sm)
indep_test <- sm.variogram(coords_matrix, data_7am_pm10$log_PM10,
    model = "independent", display = "none", se = FALSE)#Probamos la independencia espacial
#Probamos la isotropia
iso_test <- sm.variogram(coords_matrix, data_7am_pm10$log_PM10, model = "isotropic")
```

Estos test devuelven los siguientes p-valores. Para la prueba de independencia espacial, p-valor=0.18 y para la de isotropía, p-valor=0.747. Por lo tanto, no se tienen pruebas significativas para rechazar la hipótesis de isotropía, aunque su estructura dista de ser círculos concéntricos como los mencionados en la Subsección 2.5, como se puede ver en la Figura 2.19. Sin embargo el test es bastante concluyente a este respecto.

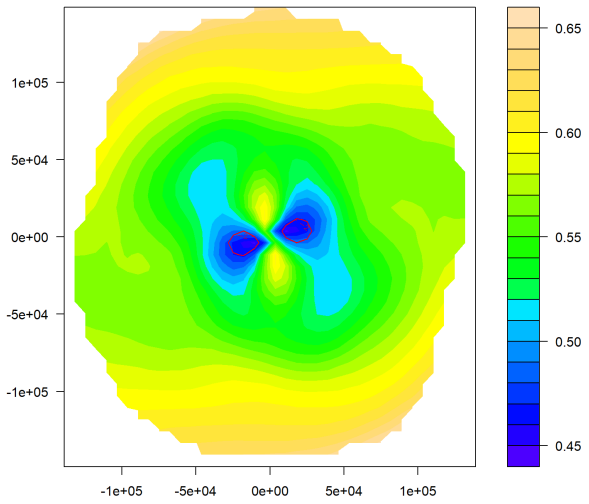


Figura 2.19: Estructura de la isolíneas caso del PM_{10}

Una situación más compleja es el caso de la dependencia espacial, el test no proporciona pruebas a los niveles de significación habituales para que se pueda aceptar la existencia de dependencia espacial. Si se realiza el test de independencia en el caso de los dos otros contaminantes presentados en este trabajo, cuyos variogramas teóricos si que parecen capturar correctamente la tendencia de los datos, se obtienen los p-valores 0.148 en el caso del SO_2 y 0.152 en el caso del NO_x; es cierto que son menores que en el caso del PM_{10}, pero siguen sin ser significativos a los niveles habituales.

En este momento es importante recordar el caso de estudio, se está intentando interpolar la contaminación en una region de $30000km^2$ a partir de 35 estaciones, la muestra es relativamente pequeña y su densidad baja. En este sentido, y aunque el test no proporcione pruebas claras de dependencia espacial, se puede considerar que tanto el variograma como las razones físicas detrás del fenómeno sugieren su existencia y permiten asumir que la falta de significancia del test se debe a la distribución o a la cantidad de estaciones.

Capítulo 3

Teoría Geoestadística y Método de Kriging

Como ya se ha visto durante todo este trabajo, la geoestadística surge de la necesidad, no solo existente en el análisis de fenómenos espaciales, de obtener predicciones precisas a partir de los datos. Esta necesidad no se limita a las matemáticas, sino que incluye a casi todas las ramas científicas, lo que convierte a la geoestadística en una herramienta muy útil para la ciencia en general.

Para lidiar con este problema de interpolación espacial existen diferentes técnicas matemáticas, como puede ser la aproximación mediante funciones paramétricas (por ejemplo a través de polinomios) o métodos no paramétricos (mediante splines). Esta sección, sin embargo, se centrará en otro enfoque, el enfoque desarrollado por G. Matheron en 1963, conocido como **Kriging**.

El Kriging puede adoptar distintas variantes según las hipótesis sobre la media de los datos o la cantidad de variables disponibles. En este trabajo se presentarán tres variantes principales: el **Kriging ordinario**, que asume una media constante pero desconocida; el **Kriging universal**, que permite modelar tendencias más complejas mediante una media variable; y el **Kriging multivariado**, diseñado para trabajar con múltiples variables correlacionadas.

Tipo de Kriging	**Media**	**Requisito mínimo**	**Nombre del modelo**
Kriging simple (KS)	Constante, conocida	Covarianza	Estacionario
Kriging ordinario (KO)	Constante, desconocida	Variograma	Intrínseco
Kriging universal (KU)	Variable, desconocida	Variograma	Modelo KU

Cuadro 3.1: Esquema replicado de los tipos de Kriging y sus características principales(Chiles & Delfiner, 1999, p. 151) [2].

En el Cuadro 3.1 se menciona el Kriging simple, se podría decir que esta es la forma más básica de este conjunto de métodos de interpolación espacial. Asume que la media del proceso es constante y conocida, lo que simplifica los cálculos, pero lo convierte en un método tremendamente limitado para su aplicación práctica donde las condiciones suelen ser mucho más complejas.

Aunque, al igual que en el caso de las series temporales, se podría tomar la media muestral y restársela al conjunto de datos para reducir el problema a uno de media cero. El problema de este enfoque es que "es muy difícil analizar las propiedades teóricas de una combinación ad-hoc de dos procedimientos de estimación. Siempre queda la duda de cuánto las variaciones de 'alta frecuencia' están corrompiendo la estimación de 'baja frecuencia' de la media"(Chiles & Delfiner, 1999, p. 164, citando a Duda, 1982) [2]. A causa de estas restricciones, no se abordará este método en profundidad en este trabajo, dándole prioridad a otras técnicas como el Kriging ordinario y universal, que son más relevantes en aplicaciones reales.

3.1. Kriging Ordinario

En esta subsección se expondrá el Kriging ordinario, "un método que a menudo se asocia con el acrónimo B.L.U.E."(Isaaks & Srivastava, 1989, p.278) [11]. Dicho acrónimo representa las siglas de "mejor estimador lineal insesgado" (en inglés: Best Linear Unbiased Estimator, BLUE).

Se dice que es "lineal" porque genera estimaciones utilizando combinaciones lineales ponderadas de los datos disponibles. Es "insesgado", ya que busca garantizar que la media de los residuos o errores (m_R) sea igual a 0. Además, es el "mejor" método porque su objetivo principal es reducir al mínimo la varianza de los errores (σ_E^2).

Estos "objetivos del Kriging ordinario son ambiciosos y, en sentido práctico, inalcanzables dado que m_R y σ_E^2 son siempre desconocidos"(Isaaks & Srivastava, 1989, p.278) [11]. Por lo tanto, como conocer el valor real de la media de los errores o reducir al mínimo su varianza resulta imposible, el problema se enfrentará planteando un modelo de los datos y trabajando con el error medio y la varianza del error del modelo creado.

La idea detrás del Kriging ordinario es bastante intuitiva. Imagínese que se tiene una situación como la de la figura 3.1, donde los puntos negros son puntos conocidos de la variable de interés, z_α, y x_0 es el punto del cual queremos tener una estimación de su valor. A partir del concepto de dependencia espacial visto en los anteriores capítulos, es lógico pensar que el valor de la variable aleatoria en x_0 se puede expresar como una combinación lineal del valor de dicha variable en los puntos cercanos, asignándole un peso a cada uno de esos puntos. De dicha forma, se puede

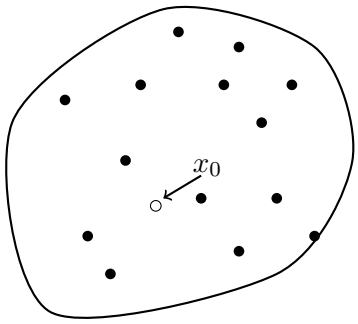


Figura 3.1: Ilustración de los puntos conocidos y el punto de interés x_0 en un dominio espacial.

definir el valor de la variable aleatoria en x_0 como:

$$Z^*(x_0) = \sum_{\alpha=1}^{n} w_\alpha Z(x_\alpha) \tag{3.1}$$

Se está asumiendo que los datos son la realización de funciones aleatorias $Z(x)$.

Ahora es importante recordar la suposición en la que se sustenta el Kriging ordinario; que la media, aunque desconocida, es constante. Por lo tanto, se debe cumplir que:

$$\mathbb{E}[Z^*(x_0) - Z(x_0)] = \mathbb{E}\left[\sum_{\alpha=1}^{n} w_\alpha Z(x_\alpha) - Z(x_0)\right] = \sum_{\alpha=1}^{n} w_\alpha \mathbb{E}[Z(x_\alpha)] - \mathbb{E}[Z(x_0)] = 0 \tag{3.2}$$

De forma intuitiva se observa que, para que se conserve la insesgadez, el sumatorio de los pesos ha de ser igual a 1, $\sum_{\alpha=1}^{n} w_\alpha = 1$. Además, es la existencia de una media de los errores insesgada la que permite el uso del variograma. No se debe olvidar que el objetivo del método es minimizar la varianza de los errores, la estimación de dicha varianza, $\sigma_E^2 = \mathrm{var}(Z^*(x_0) - Z(x_0))$, se expresa como:

$$\sigma_E^2 = \mathbb{E}\left[(Z^*(x_0) - Z(x_0))^2\right] = -\gamma(x_0 - x_0) - \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} w_\alpha w_\beta \gamma(x_\alpha - x_\beta) + 2\sum_{\alpha=1}^{n} w_\alpha \gamma(x_\alpha - x_0) \tag{3.3}$$

Ahora bien, para minimizarla se puede usar el método de multiplicadores de Lagrange que introducirá un nuevo parámetro, μ, y que dará como resultado un sistema conocido como **sistema del Kriging ordinario(KO)**, que es de la forma:

$$\begin{pmatrix} \gamma(x_1 - x_1) & \cdots & \gamma(x_1 - x_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(x_n - x_1) & \cdots & \gamma(x_n - x_n) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_n \\ \mu \end{pmatrix} = \begin{pmatrix} \gamma(x_1 - x_0) \\ \vdots \\ \gamma(x_n - x_0) \\ 1 \end{pmatrix} \tag{3.4}$$

A partir de dicho sistema, y llevando a cabo las multiplicaciones necesarias, se puede obtener

el sistema reescrito de la siguiente forma:

$$\begin{cases} \sum_{\beta=1}^{n} w_\beta \gamma(x_\alpha - x_\beta) + \mu = \gamma(x_\alpha - x_0), \quad \text{para } \alpha = 1, \ldots, n, \\ \sum_{\beta=1}^{n} w_\beta = 1. \end{cases} \tag{3.5}$$

Con lo que la estimación de la varianza del Kriging ordinario será:

$$\sigma^2_{KO} = \mu - \gamma(x_0 - x_0) + \sum_{\alpha=1}^{n} w_\alpha \gamma(x_\alpha - x_0) \tag{3.6}$$

Observación 3.1. Resulta obvio que este método necesita que el variograma esté previamente calculado, ya que el sistema de ecuaciones se construye a partir del mismo; por esa razón, la elección del modelo de variograma, o de función de covarianza, es un prerrequisito para poder utilizar el Kriging ordinario. En situaciones anisotrópicas, como las descritas en la Subsección 2.5, donde es posible transformar el problema en isotrópico mediante una transformación lineal de las coordenadas, el variograma ajustado puede utilizarse directamente. Para resolver el sistema, basta emplear un método numérico, como la factorización de matrices o métodos iterativos estándar.

Lo cierto es que lo mejor sería ejemplificar el método. Para ello se presentará un caso muy simplificado dado que la bibliografía en este sentido da ejemplos muy extensos, que encajarían más en un libro que en un trabajo de fin de grado. Resulta interesante mencionar el presentado por Noel Cressie en su libro Statistics for Spatial Data (1991, p. 212) [3]. En este caso, se describe un estudio llevado a cabo en el acuífero de Wolfcamp ubicado entre Texas y Nuevo México en Estados Unidos. En él se utilizó el Kriging ordinario para estimar si era posible que un potencial vertedero nuclear que se tenía planeado instalar en Texas pudiese contaminar el acuífero. El ejemplo es muy ilustrativo, tanto de la utilidad que los métodos de interpolación espacial tienen, como por sus características, dado que la situación que se encuentra el equipo encargado del estudio es una anisotropía geométrica, tratada en la Subsección 2.5.1.

El ejemplo aquí expuesto será "de juguete" para facilitar la comprensión del método sin complicar los cálculos.

Ejemplo 3.2. Imagínese un estudio en el que se toman muestras de la concentración de un elemento en línea recta, avanzando únicamente en la dirección de la coordenada x. Y que los valores obtenidos son los siguientes:

N. Muestra	**Coordenada** x	$Z(x)$
1	0	5.3
2	1	6.8
3	2	7.15

En este caso, se asume que la media de la variable aleatoria es constante pero desconocida, como es característico en el Kriging ordinario. Y se busca estimar el valor de $Z(x_0)$ en $x_0 = 1.5$. Se supone que el ajuste del variograma se asocia con un modelo de variograma exponencial sin efecto-nugget con rango $a = 1$ y meseta $c = 1$.

$$\gamma(h) = c \cdot \left(1 - e^{\frac{-|h|}{a}}\right)$$

Como se conoce la forma del variograma, se calcula su valor para las diferentes distancias:

- Entre $x = 0$ y $x = 1$ la distancia es $h = 1$ por lo tanto:

$$\gamma(1) = 1 \cdot \left(1 - e^{-1}\right) = 1 - e^{-1} \approx 0.6321$$

- Entre $x = 0$ y $x = 2$ la distancia es $h = 2$ por lo tanto:

$$\gamma(2) = 1 \cdot \left(1 - e^{-2}\right) = 1 - e^{-2} \approx 0.8647$$

- Entre $x = 1$ y $x = 2$ la distancia es $h = 1$ por lo tanto:

$$\gamma(1) = 1 - e^{-1} \approx 0.6321$$

Por el Sistema 3.4, se tendrá que $\Gamma \cdot w = \gamma$, donde:

$$\Gamma = \begin{bmatrix} \gamma(0) & \gamma(1) & \gamma(2) & 1 \\ \gamma(1) & \gamma(0) & \gamma(1) & 1 \\ \gamma(2) & \gamma(1) & \gamma(0) & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0.6321 & 0.8647 & 1 \\ 0.6321 & 0 & 0.6321 & 1 \\ 0.8647 & 0.6321 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

El vector γ contiene los variogramas entre los puntos conocidos y $x_0 = 1.5$:

$$\gamma(1.5 - 0) = \gamma(1.5) = 1 \cdot \left(1 - e^{-1.5}\right) \approx 0.7769,$$

$$\gamma(1.5 - 1) = \gamma(0.5) = 1 \cdot \left(1 - e^{-0.5}\right) \approx 0.3935,$$

$$\gamma(1.5 - 2) = \gamma(0.5) \approx 0.3935.$$

De donde, $\gamma^T = [0.7769, 0.3935, 0.3935, 1]$. Por último se resuelve el sistema $\lambda = \Gamma^{-1} \cdot \gamma$, los cálculos se omiten en este trabajo, y se obtienen como resultado los siguientes pesos $w_1 = 0.0430, \quad w_2 = 0.4706, \quad w_3 = 0.4864$.

Estos pesos son los que permiten calcular el valor en x_0 como combinación lineal a partir de la Ecuación 3.1. El valor estimado para $Z(x_0 = 1.5)$ usando Kriging ordinario es:

$$Z(x_0) = 6.9057$$

3.2. Kriging Universal

Al igual que en el caso del Kriging ordinario, el Kriging universal parte de premisas similares en cuanto al uso de la dependencia espacial y del variograma. Como se puede ver en la tabla del principio de este capítulo, la diferencia que genera la existencia de dos modelos claramente diferenciados es la hipótesis de media constante. El Kriging ordinario supone que la media es constante, pero desconocida, mientras que en el caso del Kriging universal se asume que la media es desconocida, pero no constante.

Esta idea de que la media de la variable aleatoria no es constante es útil en el caso de que los datos presenten tendencias. En el segundo capítulo, en la Subsección 2.2.1, ya se mencionaron razones por las cuales el variograma podía no capturar correctamente las características de los datos, una de estas razones era la presencia de tendencias. Este es el problema que el Kriging universal intenta solventar dividiendo la función aleatoria en una combinación lineal de dos componentes. El primero de ellos será una función determinista cuyo valor es conocido en todos los puntos de la región, el otro es el componente aleatorio, una función estacionaria de segundo orden. De esta forma, la función aleatoria $Z(x)$ se podrá expresar como:

$$Z(x) = m(x) + Y(x) \tag{3.7}$$

Con $m(x)$ como componente determinista, de forma que $\mathbb{E}[Z(x)] = m(x)$, e $Y(x)$ como componente aleatorio con media cero y un variograma (o función de covarianza) asociado que captura la dependencia espacial del fenómeno.

De esta forma se asume que la media de la función aleatoria no es constante, sino una combinación lineal de funciones conocidas $\{f_0(x), \ldots, f_l(x)\}$ y coeficientes $\{a_0, \ldots, a_l\}$ desconocidos, por lo que:

$$\mathbb{E}[Z(x)] = m(x) = \sum_{l=0}^{L} a_l f_l(x) \tag{3.8}$$

De esta forma $Z(x)$ se puede expresar como:

$$Z(x) = \sum_{l=0}^{L} a_l f_l(x) + Y(x) \tag{3.9}$$

Y, de forma equivalente al caso del Kriging ordinario, la estimación en un punto no muestreado se calculará como combinación lineal de los puntos conocidos:

$$Z^*(x_0) = \sum_{\alpha=1}^{n} w_\alpha Z(x_\alpha) \tag{3.10}$$

Ahora bien, se quiere que la estimación sea insesgada, para ello:

$$\mathbb{E}[Z(x_0) - Z^*(x_0)] = m(x_0) - \sum_{\alpha=1}^{n} w_\alpha Z(x_\alpha) = \sum_{l=0}^{L} a_l (f_l(x_0) - \sum_{\alpha=1}^{n} w_\alpha f_l(x_\alpha)) = 0 \tag{3.11}$$

Es trivial que para que se cumpla esta condición, dado que los a_l son distintos de cero, se debe respetar que:

$$\sum_{\alpha=1}^{n} w_\alpha f_l(x_\alpha) = f_l(x_0), \quad \text{para} \quad l = 0, \ldots, L \tag{3.12}$$

A estas condiciones se las conoce como **condiciones de universalidad**. Este adjetivo, "universal", lo acuñó Matheron para referirse a un método que predice un fenómeno que tiene una tendencia que es una combinación lineal desconocida de funciones conocidas. De esta definición proviene el nombre del Kriging universal.

Observación 3.3. Es fácil ver que si $L = 1$ y $f_0(x) \equiv 1$, se obtiene el modelo del Kriging ordinario. En este caso, el componente determinista $m(x)$ es constante, ya que se reduce a $m(x) = a_0$ y, por tanto, no hay tendencia o variación de la media de la variable aleatoria. En esta situación cualquier variación en los datos se modela suponiendo un comportamiento únicamente aleatorio.

Al igual que en el caso de Kriging ordinario, el objetivo de este método es minimizar la varianza de los errores. Dicha varianza se define como $\sigma = var(Z^*(x_0) - Z(x_0))$. Sustituyendo $Z^*(x_0) = \sum_{\alpha=1}^{n} w_\alpha Z(x_\alpha)$ y $Z(x_0) = m(x_0) + Y(x_0)$, se tiene que:

$$Z^*(x_0) - Z(x_0) = \sum_{\alpha=1}^{n} w_\alpha Z(x_\alpha) - \left(\sum_{l=0}^{L} a_l f_l(x_0) + Y(x_0) \right)$$

Separando términos:

$$Z^*(x_0) - Z(x_0) = \sum_{\alpha=1}^{n} w_\alpha Y(x_\alpha) - Y(x_0) + \underbrace{\left(\sum_{\alpha=1}^{n} w_\alpha \sum_{l=0}^{L} a_l f_l(x_\alpha) - \sum_{l=0}^{L} a_l f_l(x_0) \right)}_{\text{Restricciones de tendencia}}$$

Por la condición de universalidad $\sum_{\alpha=1}^{n} w_\alpha f_l(x_\alpha) = f_l(x_0)$, el término relacionado con la tendencia desaparece, y por tanto:

$$Z^*(x_0) - Z(x_0) = \sum_{\alpha=1}^{n} w_\alpha Y(x_\alpha) - Y(x_0) + 0$$

Por último, la estimación de la varianza del error en términos de los pesos w_α y del variograma $\gamma(x_\alpha - x_\beta)$ será:

$$\sigma_E^2 = \gamma(x_0 - x_0) + \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} w_\alpha w_\beta \gamma(x_\alpha - x_\beta) - 2 \sum_{\alpha=1}^{n} w_\alpha \gamma(x_\alpha - x_0) \tag{3.13}$$

Esta formulación es igual que la del método del Kriging ordinario, Ecuación 3.1. Sin embargo, las condiciones que determinan los pesos w_α son diferentes en el universal y por ello la varianza del Kriging universal será diferente.

Al igual que en el caso del Kriging ordinario, la forma clásica de resolver este problema de minimización es usar el método de multiplicadores de Lagrange. Se considera la función:

$$\mathcal{L} = \underbrace{\mathrm{Var}(Z^*(x_0) - Z(x_0))}_{\text{Varianza del error}} + 2\sum_{l=0}^{L} \mu_l \underbrace{\sum_{\alpha=1}^{n} w_\alpha f_l(x_\alpha) - f_l(x_0)}_{\text{Condición de universalidad (penalización)}} \tag{3.14}$$

La función Lagrangiana se compone de dos términos, el primero representa la varianza del error, que es lo que se desea minimizar, mientras que el segundo, $2\sum \mu_l \left[\sum w_\alpha f_l(x_\alpha) - f_l(x_0)\right]$, introduce la penalización que fuerza el cumplimiento de las condiciones de universalidad. Este segundo término incluye un tipo de variable desconocida adicional, μ_l para $l = 0, \ldots, L$, los multiplicadores de Lagrange.

Si las restricciones no se cumpliesen, el valor de $\mathcal{L}$ aumentaría, pero esto se evita durante la minimización, la cual se determina igualando las ecuaciones en derivadas parciales de $\mathcal{L}$ a cero.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial w_\alpha} &= 2\sum_{\beta=1}^{n} w_\beta C(\mathbf{x}_\alpha - \mathbf{x}_\beta) - 2C(\mathbf{x}_\alpha - \mathbf{x}_0) + 2\sum_{l=0}^{L} \mu_l f_l(\mathbf{x}_\alpha) = 0, \quad \alpha = 1, \ldots, n, \\ \frac{\partial \mathcal{L}}{\partial \mu_l} &= 2\left[\sum_{\alpha=1}^{n} w_\alpha f_l(\mathbf{x}_\alpha) - f_l(\mathbf{x}_0)\right] = 0, \quad l = 0, 1, \ldots, L. \end{aligned} \tag{3.15}$$

De donde, simplificando, se obtiene el **sistema del Kriging universal (KU)**:

$$\begin{cases} \sum_{\beta=1}^{n} w_\beta C(\mathbf{x}_\alpha - \mathbf{x}_\beta) - \sum_{l=0}^{L} \mu_l f_l(\mathbf{x}_\alpha) = C(\mathbf{x}_\alpha - \mathbf{x}_0), & \text{para } \alpha = 1, \ldots, n, \\ \sum_{\beta=1}^{n} w_\beta f_l(\mathbf{x}_\beta) = f_l(\mathbf{x}_0), & \text{para } l = 0, \ldots, L. \end{cases} \tag{3.16}$$

En forma matricial:

$$\underbrace{\begin{bmatrix} \Sigma & F \\ F' & 0 \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} w \\ -\mu \end{bmatrix}}_{\mathbf{X}} = \underbrace{\begin{bmatrix} c \\ f \end{bmatrix}}_{\mathbf{B}}$$

Siendo Σ la matriz de covarianza, F la matriz de diseño, w los pesos del modelo y μ los multiplicadores de Lagrange. Para que dicho sistema lineal tenga solución:

- **Existencia**: Si y solo si $\mathbf{A}$ es no singular ($det(\mathbf{A}) \neq 0$), lo cual requiere que F tenga rango completo igual a $L + 1$. Es decir, los vectores columna f_l deben ser linealmente independiente. Por esta razón "las funciones $f_l(x)$ tienen que ser escogidas con cariño" (Wackernagel, 2003, p.301) [26].

- **Unicidad**: Si además Σ es estrictamente definida positiva ($\det(\Sigma) > 0$).

Ahora bien, la varianza del método será:

$$\sigma^2_{KU} = \sum_{\alpha=1}^{n} w_\alpha \gamma(x_\alpha - x_0) + \sum_{l=0}^{L} \mu_l f_l(x_0) \tag{3.17}$$

Es importante resaltar que además de la formulación de la Ecuación 3.16, el Kriging universal también se puede expresar en función del variograma en los siguientes términos:

$$\begin{cases} \sum_{\beta=1}^{n} w_\beta \gamma(\mathbf{x}_\alpha - \mathbf{x}_\beta) - \sum_{l=0}^{L} \mu_l f_l(\mathbf{x}_\alpha) = \gamma(\mathbf{x}_\alpha - \mathbf{x}_0), & \text{para } \alpha = 1, \ldots, n, \\ \sum_{\beta=1}^{n} w_\beta f_l(\mathbf{x}_\beta) = f_l(\mathbf{x}_0), & \text{para } l = 0, \ldots, L. \end{cases} \tag{3.18}$$

Como se puede ver, es prácticamente equivalente a la Ecuación 3.16, de todas formas eso no la hace menos importante. Su utilidad reside en que, como se ha visto en la Subsección 2.3.1, en el caso de que el variograma no esté acotado la función de covarianza no estará bien definida, aunque el variograma exista. En este contexto, esta formulación basada en el variograma permite realizar la interpolación de puntos no muestreados incluso si el variograma no está acotado.

Observación 3.4. Exponer un ejemplo ficticio de la construcción del Kriging universal no aporta mucho valor, su cálculo es similar al de Kriging ordinario con la diferencia principal de la existencia de las funciones $\{f_0(x), \ldots, f_l(x)\}$. Dichas funciones son conocidas y dependen del fenómeno analizado, en un ejemplo ficticio carece de mucho sentido tomarlas de forma arbitraria dada su importancia para el modelo. Al igual que en el caso del Kriging ordinario, los ejemplos de la bibliografía suelen ser muy extensos, pero merece la pena mencionar otra vez a Noel Cressie y su libro Statistics for Spatial Data (1991, p. 224) [3]. En él se presenta un caso de uso real donde se utilizó el Kriging universal para estudiar la dependencia espacial de la tensión del agua en el suelo bajo diferentes formas de labranza durante una temporada de cultivo en Iowa.

3.3. Kriging Multivariado (Cokriging)

Los métodos de Kriging presentados hasta el momento se centraban en la aproximación de la variable objetivo a partir únicamente de la dependencia espacial entre los puntos del dominio de dicha variable objetivo. En este sentido, el Cokriging es la extensión natural del Kriging cuando se tiene acceso a datos multivariantes. Este método no utiliza solo la información de la variable de interés, sino que se apoya en variables auxiliares para mejorar la estimación del método.

La mayoría de estudios reales utilizan este enfoque usando casi siempre más de una variable. Los ejemplos en ciencias naturales son numerosos, pero se puede tomar el que menciona (Chiles & Delfiner, 1999, p. 292) [2]. La industria petrolera, por ejemplo, utiliza herramientas a la hora de realizar prospecciones que no miden únicamente la profundidad, sino que son capaces de captar

otros parámetros como la porosidad, permeabilidad y saturación de fluidos, todos ellos están, en mayor o menor medida, relacionados con la profundidad.

De estas situaciones surge el interés por una generalización multivariante del Kriging, la cual se presentará en este capítulo y recibe el nombre de **Cokriging**.

Observación 3.5. Ya se ha mencionado a lo largo de este trabajo la gran complejidad a la que puede llegar la geoestadística al intentar modelar fenómenos físicos. En este sentido, aunque la geoestadística multivariante puede involucrar conceptos como isotopía y heterotopía, como se puede ver en el Figura 3.2, en relación con la disposición de los datos multivariantes no se profundizará en estos aspectos debido a su complejidad y a que no son relevantes para el caso práctico analizado.

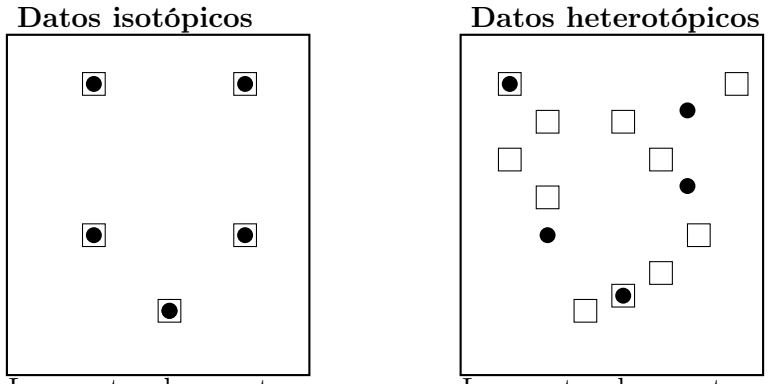


Figura 3.2: Datos isotópicos y heterotópicos en el caso multivariante. Los círculos negros representan los datos primarios, mientras que los cuadrados blancos representan los datos secundarios, replicado de (Wackernagel, 2003, p.159) [26].

Observación 3.6. Se supondrá que el muestreo es igual para todos los componentes, por esa razón se seguirá la siguiente notación presentada por Wackernagel (2003) [26], en la cual la función aleatoria vectorial se define como:

$$\mathbf{Z}(\mathbf{x}) = \begin{bmatrix} Z_1(\mathbf{x}) \\ Z_2(\mathbf{x}) \\ \vdots \\ Z_N(\mathbf{x}) \end{bmatrix}$$

Donde $\mathbf{x}$ denota una ubicación específica en el espacio y cada componente corresponde a una variable regionalizada en esa ubicación.

Antes de presentar el modelo, se debe hacer una introducción a los conceptos de covarianza cruzada y variograma cruzado, que son las herramientas encargadas de analizar la relación

espacial entre las dos o más variables regionalizadas que están conectadas entre sí.

3.3.1. Función de Covarianza cruzada y Variograma cruzado

La función de covarianza se puede calcular no solo en determinadas localizaciones x, sino que también se puede calcular en función del vector distancia h. Bajo la hipótesis de estacionariedad de segundo orden se expresa como:

$$\begin{cases} \mathbb{E}[Z_i(\mathbf{x})] = m_i, & \text{para todo } \mathbf{x} \in \mathcal{D};\ i = 1, \dots, N, \\ \mathbb{E}\big[(Z_i(\mathbf{x}) - m_i) \cdot (Z_j(\mathbf{x}+\mathbf{h}) - m_j)\big] = C_{ij}(\mathbf{h}), & \text{para todo } \mathbf{x}, \mathbf{x}+\mathbf{h} \in \mathcal{D};\ i, j = 1, \dots, N. \end{cases} \tag{3.19}$$

Al igual que la función de covarianza en el caso univariante, la función de covarianza cruzada es invariante por traslaciones al depender únicamente de la distancia definida por el vector $\mathbf{h}$, pero, a diferencia de la función de covarianza, en este caso no es una función necesariamente par o impar. Generalmente, un cambio en el orden de las variables o en el signo del vector distancia cambia el valor de la covarianza cruzada:

$$C_{ij}(\mathbf{h}) \neq C_{ji}(\mathbf{h}) \quad \text{y} \quad C_{ij}(-\mathbf{h}) \neq C_{ij}(\mathbf{h}) \tag{3.20}$$

En este sentido, y al igual que la noción de variograma se podía generalizar a partir de la función de covarianza, la existencia de una función de covarianza cruzada propicia la aparición del concepto de variograma cruzado:

$$\gamma_{ij}(\mathbf{h}) = \frac{1}{2}\mathbb{E}\left[\big(Z_i(\mathbf{x}+\mathbf{h}) - Z_i(\mathbf{x})\big) \cdot \big(Z_j(\mathbf{x}+\mathbf{h}) - Z_j(\mathbf{x})\big)\right]$$

Que no es más que el producto del incremento esperado de las dos variables, a diferencia de la covarianza cruzada, sí que será par $\gamma_{ij}(h) = \gamma_{ij}(-h)$ y satisfará la siguiente desigualdad:

$$\gamma_{ii}(\mathbf{h})\gamma_{jj}(\mathbf{h}) \geq |\gamma_{ij}(\mathbf{h})| \tag{3.21}$$

Por último, la relación entre ambos conceptos, es decir, la forma en que el variograma cruzado puede derivarse de la función de covarianza cruzada, se expresa de la siguiente forma:

$$\gamma_{ij}(\mathbf{h}) = C_{ij}(0) - \frac{1}{2}(C_{ij}(-\mathbf{h}) + C_{ij}(+\mathbf{h})) \tag{3.22}$$

Es fácil observar que el variograma cruzado se puede entender como una combinación de la covarianza cruzada para los valores $-\mathbf{h}$ y $\mathbf{h}$.

Lo cierto es que las diferencias entre el Kriging y el Cokriging no son grandes, como mencionan Chiles y Delfiner (1999,p. 292) [2],"a pesar de una mayor complejidad [del Cokriging] principalmente debido a las notaciones, hay poca novedad desde un punto de vista teórico" .

Al igual que en el caso del Kriging univariante, podemos diferenciar dos formas de Cokriging; una donde se asume que la media de la variable principal es desconocida pero constante, Cokriging ordinario, y otra en la que se asume que su media es desconocida y variable.

En el Cokriging, las variables secundarias se usan para complementar la estimación de la variable principal, por lo que las suposiciones de media constante o variable se aplican solo a dicha variable.

3.3.2. Cokriging Ordinario

Al igual que en el caso del Kriging univariante, el Cokriging se basa en la estimación del valor de un punto no muestreado mediante una combinación lineal de los puntos cercanos, asignando un peso a cada variable en función de la dependencia espacial, definida por el variograma. En este sentido, la formulación del Cokriging ordinario es:

$$Z^*_{i_0}(x_0) = \sum_{i=1}^{N} \sum_{\alpha=1}^{n_i} w^i_\alpha Z_i(x_\alpha) \tag{3.23}$$

Donde, $Z^*_{i_0}(x_0)$ es la estimación de la variable principal en el punto x_0, N es el número total de variables involucradas en el Cokriging, lo que incluye tanto la principal i_0 como las secundarias, n_i es el número de muestras de la variable i esto permite que el tamaño muestral sea diferente para cada variable, haciendo posible que el modelo se ajuste a casos heterotópicos, en los cuales no todas las variables están muestreadas en todos los puntos.

Por último, w^i_α son los pesos para una determinada posición α y una determinada variable i y $Z_i(x_\alpha)$ es el valor de la variable Z_i en x_α.

Al igual que en el caso univariante, el uso del variograma se sustenta en la hipótesis intrínseca, la cual establece que los incrementos del proceso aleatorio son estacionarios. Esta hipótesis se satisfará si:

$$\sum_{\alpha=1}^{n_i} w^i_\alpha = \delta_{ii_0} = \begin{cases} 1 & \text{si } i = i_0, \\ 0 & \text{en caso contrario.} \end{cases} \tag{3.24}$$

Observación 3.7. Es decir, los pesos de las variables secundarias no contribuyen directamente al valor de la estimación, dado que su suma es cero, pero modulan su valor reduciendo su varianza y dando una mejor estimación.

De esta forma se llega a que la expresión para la estimación del error medio es:

$$\mathbb{E}\left[Z^*_{i_0}(x_0) - Z_{i_0}(x_0)\right] = \mathbb{E}\left[\sum_{i=1}^{N}\sum_{\alpha=1}^{n_i} w^i_\alpha Z_i(x_\alpha) - \underbrace{\sum_{\alpha=1}^{n_{i_0}} w^{i_0}_\alpha}_{1} Z_{i_0}(x_0) - \sum_{i=0, i\neq i_0}^{N} \underbrace{\sum_{\alpha=1}^{n_i} w^i_\alpha}_{0} Z_i(x_0)\right] \tag{3.25}$$

$$= \sum_{i=1}^{N} \sum_{\alpha=1}^{n_i} w_\alpha^i \underbrace{\mathbb{E}\left[Z_i(x_\alpha) - Z_i(x_0)\right]}_{0} = 0.$$

El objetivo del método es minimizar la varianza de los errores, $\sigma_E^2 = \mathbb{E}\left[\left(\sum_{i=1}^{N}\sum_{\alpha=1}^{n_i} w_\alpha^i Z_i(\mathbf{x}_\alpha) - Z_{i_0}(\mathbf{x}_0)\right)^2\right]$, tomando la variable aleatoria ficticia $Z_i(0)$ y sustituyendo, se tiene que:

$$\sigma_E^2 = \mathbb{E}\left[\left(\sum_{i=1}^{N}\left(\sum_{\alpha=0}^{n_i} w_\alpha^i Z_i(\mathbf{x}_\alpha) - Z_i(0)\sum_{\alpha=0}^{n_i} w_\alpha^i\right)\right)^2\right] = \tag{3.26}$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^{N}\sum_{\alpha=0}^{n_i} w_\alpha^i \underbrace{\left(Z_i(\mathbf{x}_\alpha) - Z_i(0)\right)}_{\text{incrementos}}\right)^2\right] = \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{\alpha=0}^{n_i}\sum_{\beta=0}^{n_j} w_\alpha^i w_\beta^j C_{ij}^I(\mathbf{x}_\alpha, \mathbf{x}_\beta).$$

Si se quiere expresar en función del variograma:

$$\sigma_E^2 = 2\sum_{i=1}^{N}\sum_{\alpha=1}^{n_i} w_\alpha^i \gamma_{ii_0}(\mathbf{x}_\alpha - \mathbf{x}_0) - \gamma_{i_0 i_0}(\mathbf{x}_0 - \mathbf{x}_0) - \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{\alpha=1}^{n_i}\sum_{\beta=1}^{n_j} w_\alpha^i w_\beta^j \gamma_{ij}(\mathbf{x}_\alpha - \mathbf{x}_\beta). \tag{3.27}$$

Y, tras la minimización mediante multiplicadores de Lagrange, se llega a que el sistema del Cokriging ordinario es:

$$\begin{cases} \displaystyle\sum_{j=1}^{N}\sum_{\beta=1}^{n_j} w_\beta^j \gamma_{ij}(\mathbf{x}_\alpha - \mathbf{x}_\beta) + \mu_i = \gamma_{ii_0}(\mathbf{x}_\alpha - \mathbf{x}_0), & \text{for } i = 1, \ldots, N;\ \alpha = 1, \ldots, n_i, \\ \displaystyle\sum_{\beta=1}^{n_i} w_\beta^i = \delta_{ii_0}, & \text{for } i = 1, \ldots, N, \end{cases} \tag{3.28}$$

Y la varianza del método es:

$$\sigma_{CK}^2 = \sum_{i=1}^{N}\sum_{\alpha=1}^{n_i} w_\alpha^i \gamma_{ii_0}(\mathbf{x}_\alpha - \mathbf{x}_0) + \mu_{i_0} - \gamma_{i_0 i_0}(\mathbf{x}_0 - \mathbf{x}_0).$$

Este método aprovecha la ayuda de variables secundarias para proporcionar una mejor estimación de la variable principal en los puntos no muestreados. Al igual que en el caso univariante, los ejemplos de la bibliografía suelen ser bastante extensos y se omitirán en este trabajo.

Cabe mencionar dos variaciones del método importantes, el Cokriging universal y el Cokriging colocalizado.

El Cokriging universal es una variante avanzada del Cokriging que permite modelar tendencias en los datos. Lo que hace a este método realmente complejo es su capacidad de modelar no solo la tendencia de las variables de forma independiente, sino también la relación que surge entre las tendencias de las diferentes variables. Se pueden clasificar estas tendencias en tres casos: algebraicamente independientes, linealmente dependientes y mixtas. Aunque esta complejidad

lo convierte en un método muy potente, también complica su implementación y hace que quede fuera del alcance de este trabajo.

Por otro lado, el Cokriging colocalizado se enfoca en situaciones donde una variable está densamente muestreada, prácticamente en todos los puntos del dominio, y las demás no, lo que lo hace especialmente relevante en estudios heterotópicos. Al igual que el Cokriging universal, su análisis detallado no se incluirá en este trabajo debido a su complejidad y extensión.

Capítulo 4

Caso Práctico: Aplicación en Datos Reales

En este caso práctico se presentará el estudio de la distribución de la contaminación en Galicia. El objetivo principal es utilizar las técnicas de interpolación espacial presentadas en este trabajo para generar un mapa que estime la concentración de los distintos contaminantes presentes en el aire. Este mapa permitirá obtener una estimación de la concentración de los diferentes contaminantes aéreos en cada punto del territorio gallego.

4.1. Descripción de los Datos

El acceso a los datos de la concentración de contaminantes es público a través de la página de Meteogalicia, pero su descarga solo está disponible pasado un tiempo determinado, durante el cual se realiza la validación de los datos. Por esta razón, y dado que algunos datos pueden tener la etiqueta de temporal (T) meses después de su recogida, se tomarán los datos de Enero de 2024 para el desarrollo del mapa.

Los contaminantes atmosféricos son muy variados e incluyen tanto sustancias químicas como partículas. Las estaciones de la red de Meteogalicia recogen información sobre algunos de ellos como el dióxido de azufre (SO_2), el monóxido de carbono (CO), los óxidos de nitrógeno (NO, NO_2, y NO_x), el ozono (O_3), las partículas en suspensión ($PM_{2.5}$ y PM_{10}), el sulfuro de hidrógeno (SH_2), el fluoruro de hidrógeno (FH) y el benceno (BEN). Todos estos contaminantes provienen, principalmente, de actividades humanas [7], como la quema de combustibles fósiles, procesos industriales, las emisiones de vehículos de combustión o actividades agrícolas, aunque en casos puntuales se deben a fenómenos naturales como incendios forestales.

Su recopilación y el estudio de su distribución son de interés público debido a su relación con problemas de salud, como enfermedades respiratorias y cardiovasculares [5], así como por su impacto ambiental, contribuyendo al cambio climático.

En cuanto al funcionamiento de las estaciones, es importante mencionar que estas no registran todas las variables de contaminación, es decir, no todas las estaciones recogen los mismos contaminantes. Por ejemplo, la estación de la Torre de Hércules, en A Coruña, monitoriza ocho de los contaminantes mencionados: SO_2, CO, NO, NO_2, NO_x, O_3, $PM_{2.5}$ y PM_{10}; mientras que la de Areeiro, en Pontevedra, solo recoge tres: SO_2, SH_2 y PM_{10}.

Por último, mencionar que la creación de la base de datos se tuvo que hacer manualmente debido a que la descarga no se puede hacer directamente desde la web de Metogalicia, la web solo permite la descarga estación por estación. Después de esta descarga, se indexaron todas en la misma base de datos. La posición exacta de las estaciones (longitud y latitud) tampoco se proporciona, lo que hizo necesaria su búsqueda a mano una por una en Google Maps para obtener la base de datos final con la que se generará el mapa, un resumen del código de construcción de la base de datos se encuentra en el Anexo I. La posición de las estaciones se puede observar en la siguiente Figura 4.1

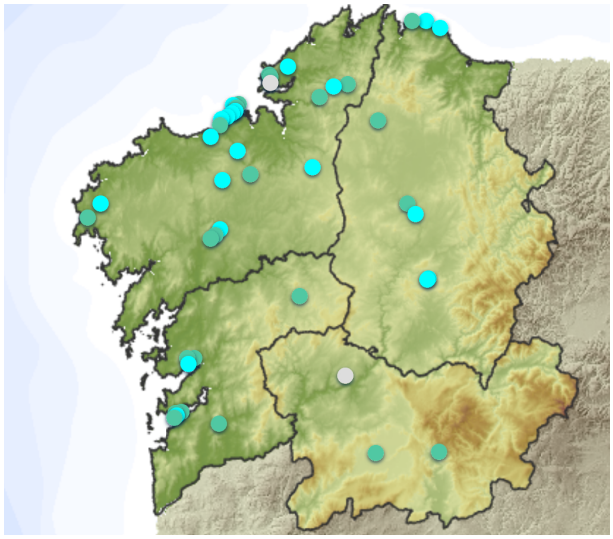


Figura 4.1: Mapa que muestra la ubicación de las estaciones de monitoreo de contaminación ambiental en la Comunidad Autónoma de Galicia [15].

El dataset final obtenido tras integrar los datos y localizar las estaciones consta de cerca de 32000 registros de las 43 estaciones de monitoreo de Meteogalicia. Además, incluye 19 columnas que representan tanto variables geográficas como las variables de contaminación. Las columnas incluyen:

- **Localización** y **Provincia**: nombre de la estación y la provincia donde se encuentra
- **Latitud** y **Longitud**: coordenadas geográficas de cada estación.
- **Variables de contaminación**:SO_2, PM_{10}, O_3, NO_x, NO_2, NO, $PM_{2.5}$, CO, BEN, SH_2, FH.

4.2. Análisis Exploratorio de los Datos

La intención ahora será entender cómo se comportan los datos. En este sentido, el primer paso será visualizar el funcionamiento de las diferentes estaciones de monitoreo de contaminación.

Como se ha visto en la sección anterior, se tiene constancia de que no todas las estaciones recogen exactamente los mismos datos, por eso lo natural es preguntarse cuáles son los contaminantes más muestreados y en qué proporción lo son. Para ello basta obtener el número de valores NA por columna, con lo que se obtiene la siguiente Tabla 4.1.

Variable	**Porcentaje NA (%)**
CO ng/m^3	97.67
BEN µg/m^3	97.67
BEN µg/m^{3}N	97.67
SH_2 µg/m^3	93.02
FH µg/m^3	93.02
CO mg/m^3	65.12

Cuadro 4.1: Porcentaje de valores NA por variable en orden descendente

En dicha tabla aparecen representadas únicamente las variables con un porcentaje de valores NA superior al 50 %. A partir de esta tabla ya se puede intuir que la interpolación espacial de variables con tan pocos datos muestrales va a ser complicada, especialmente en el caso de variables como FH, BEN o SH_2. En este momento es importante mencionar la diferencia entre $CO\ ng/m^3$ y $CO\ mg/m^3$, lo cierto es que la primera de dichas cantidades solo se registra en una estación, la de Fraga Redonda pero, aunque se hiciese una conversión de cifras, el porcentaje de valores NA del CO se mantendría alrededor del 60 %, el mismo problema ocurre con la diferencia entre los bencenos, BEN. Por esta razón no se tratarán estos compuestos en el estudio.

Ahora que se conoce qué contaminantes se recogen en pocas estaciones, la pregunta es si el hecho de que una determinada estación recoja estos datos repercute en la recogida de otros contaminantes, es decir, si medir contaminantes poco habituales afecta a la recopilación de contaminantes más comunes.

Se querrá conocer para qué contaminantes son relevantes todas las estaciones o si existen estaciones concretas que recogen solo un determinado contaminante poco común y no los demás, lo que permitiría descartar esa estación del estudio. Por esa razón, se presenta el porcentaje de datos NA de las variables principales en las estaciones donde las variables problemáticas registran algún valor. Las tablas asociadas al CO y al BEN y el código desarrollado en R carecen de interés en este caso, pero se presenta en la Sección I.2 del Anexo I. Las tablas que sí

$[FH$ **µg/m³**$]$

Variable	**Porcentaje NA**
SO_2 µg/m³	0.00
PM_{10} µg/m³	0.00
O^3 µg/m³	66.67
NO_X µg/m³	66.67
NO_2 µg/m³	66.67
NO µg/m³	66.67
$PM_{2.5}$ µg/m³	0.00

$[SH_2$ **µg/m³**$]$

Variable	**Porcentaje NA**
SO_2 µg/m³	0.00
PM_{10} µg/m³	0.00
O^3 µg/m³	66.67
NO_X µg/m³	33.33
NO_2 µg/m³	33.33
NO µg/m³	33.33
$PM_{2.5}$ µg/m³	66.67

Cuadro 4.2: Porcentaje de valores `NA` de las variables de interés en las estaciones donde las variables FH y $SH2$ recopilan datos.

son de interés son las que representan las estaciones que registran sulfuro de hidrógeno (SH_2) y fluoruro de hidrógeno(FH), se puede observar que dichas estaciones tienden a no registrar la presencia de óxidos de nitrógeno, que son marcadores claros de la contaminación humana.

Haciendo un análisis del comportamiento de las estaciones se llega a que en ocho estaciones no registran datos sobre este tipo de contaminantes, exactamente en Xubia, A Grela, Campo Futbol, Sabon Embalse, Cuiña, Escola Música, Rio Cobo y Areeiro. Además, los tres tipos de óxidos de nitrógeno se registran en las mismas estaciones, así que podemos tomar cualquiera de ellos para nuestro estudio, por ejemplo el NO_x. A parte de los óxidos de nitrógeno, se tomarán el SO_2 y el PM_{10} para mapearlos en el territorio gallego. Se comprueba de forma equivalente que estos dos componentes no se registran en siete estaciones, coincidiendo la falta de registro de estos dos contaminantes en una única estación, Penedo. Es importante tener en cuenta que estaciones están disponibles para el estudio de cada uno de los contaminantes.

4.2.1. Enfoque del Análisis de los Contaminantes SO_2, NO_x y PM_{10}

Ahora, la estructura de los datos es tanto geográfica como temporal, pues se tienen las mediciones de los contaminantes por hora en cada una de las estaciones; intentar ver la dependencia temporal excede el ámbito de este trabajo, por esa razón se planteará el enfoque siguiente.

El objetivo del análisis será evaluar si existen patrones en los niveles de contaminación a lo largo del día, es decir, estudiar la contaminación de fondo presente en todo el territorio en función de la hora. Para ello, se considerarán únicamente los días laborables, excluyendo tanto los fines de semana como los festivos. Además se eliminará la primera semana del mes porque los datos se corresponden con Enero y sus numerosos días festivos podrían estar influyendo en ellos, haciéndolos menos representativos.

Los datos descargados están divididos en horas, así que se calculará la media de contaminación para cada hora, la creación de dicho dataset es simple y se refleja en la Sección I.3 del Anexo I.

4.2.2. Análisis Exploratorio de los Contaminantes SO_2, NO_x y PM_{10}

El análisis exploratorio de los contaminantes SO_2, NO_x y PM_{10} que se realizará a continuación se basa en la propuesta de análisis planteada en Bivand et al. (2008, p.192) [1]. En este sentido, y como este estudio incluye un componente temporal y otro espacial, se presentará un gráfico del comportamiento medio de los contaminantes a lo largo de las horas del día y las gráficas de contaminación media por contaminante y por provincia. De esta forma se obtienen las siguientes gráficas:

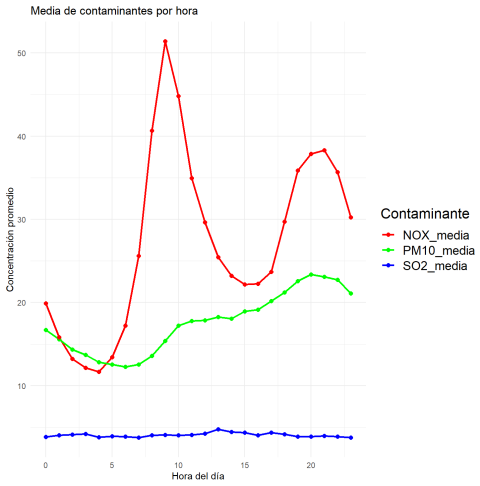


Figura 4.2: Media horaria de contaminantes (SO_2, PM_{10} y NO_X) en todas las estaciones.

Se puede observar en el gráfico que el contaminante que presenta más variación es el NO_x, un contaminante perteneciente a la clase de los óxidos de nitrogeno estrechamente relacionado con los procesos de combustión [16], lo que podría explicar los picos que presenta a las 8:00 y 19:00. Por otra parte el PM_{10} aumenta gradualmente, alcanzando su máximo por la tarde (20:00). Sin embargo, el dióxido de azufre, SO_2 se mantiene bajo y estable. Se presentan ahora los mapas

acerca de la media por provincia de cada contaminante.

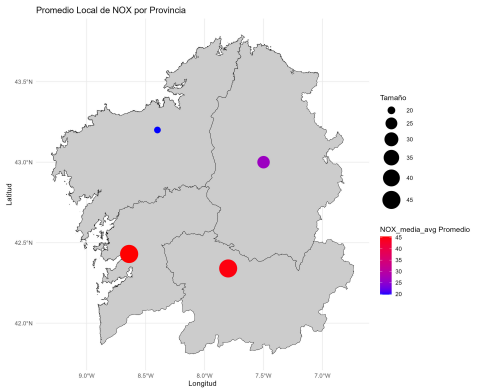


Figura 4.3: Media de NO_x por provincia

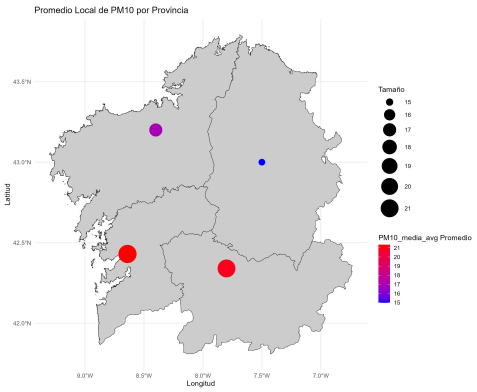


Figura 4.4: Media de PM_{10} por provincia

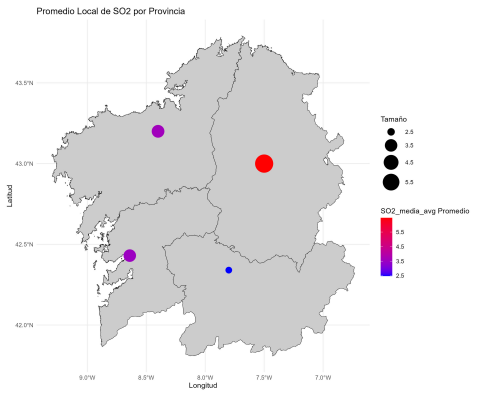


Figura 4.5: Media de SO_2 por provincia

4.3. Aplicación de Kriging

El desarrollo del variograma experimental se omite en este capítulo porque ya ha sido expuesto en la Subsección 2.6 pero se usará como base para la interpolación espacial mediante Kriging de esta sección. Para ello, al igual que en el caso del estudio del variograma, se hará uso de la librería gstat de R. Al igual que en el caso del variograma se presentarán los resultados para el caso de los tres contaminantes a una hora determinada, las 7 : 00. Se comenzará presentado la función de mayor interés en este caso:

```
#Ejemplificamos para el caso del SO2 a las 7:00
puntos_datos_sp_so2 <- as(data_7am_so2, "Spatial")
resultado_kriging_so2_7 <- krige(
  formula = log(SO2_media) ~ 1,
  locations = puntos_datos_sp_so2,
  newdata = grid_galicia_sf,
  model = modelo_variograma_so2, #Variograma teórico hallado en la Sección 2.6
  nmin = 1,  #Número mínimo de vecinos
  nmax = 5  #Número máximo de vecinos )
```

Es importante hablar del proceso de creación del grid de Galicia. Para ello primero se descargó un archivo shapefile de Galicia, que no es más que un archivo que almacena datos geoespaciales de una región. Se puede descargar directamente de diferentes webs publicas, en este caso fue descargado de `https://www.sergas.es/Saude-publica/GIS-Limites-administrativos`. Después es necesario tomar el mismo sistema de coordenadas tanto para el mapa de Galicia como para la posición de las estaciones, es un proceso un poco tedioso, pero nada complejo, que se presenta en la Sección I.4 del Anexo I.

En cuanto al desarrollo del método de interpolación eso es todo. En cuanto a las gráficas, existen diferentes opciones, la primera es representar los datos directamente con un grid mediante la librería ggplot2, lo que devuelve un gráfico similar a este, Figura 4.6. El caso de las figuras de este estilo para el PM_{10} y el NO_x se presenta en la Subsección II.1.

También se puede presentar una gráfica utilizando la librería de R raster, que permite graficar el resultado de la interpolación como una superficie continua representada mediante un degradado de colores, ese es el caso de la Figura 4.7. Se pueden ver las figuras de este estilo para el PM_{10} y el NO_x s en la Subsección II.2.

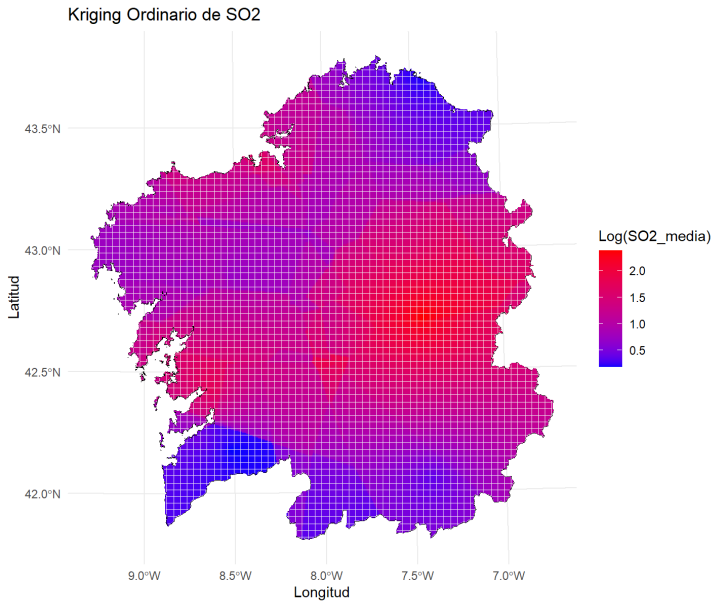


Figura 4.6: Estimación espacial de los niveles de SO_2 representada en una malla.

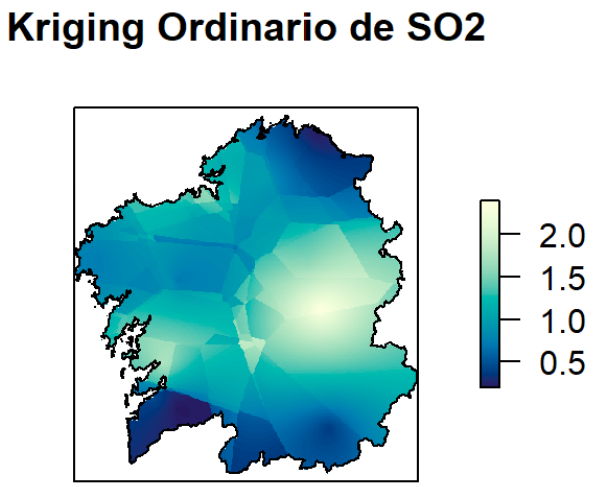


Figura 4.7: Visualización de las concentraciones interpoladas de SO_2 como una superficie continua.

Ahora se van a presentar los mapas de la evolución media de los contaminantes por hora a lo largo del día. El código es equivalente al realizado para el caso del SO_2 para las 07 : 00 simplemente que se implementa un bucle for para iterar según lar horas, se presenta en la Sección I.5 del Anexo I.

Los gráficos resultantes se presentan en vídeo, por limitaciones técnicas debidas a la discontinuación de la librería media9 por parte de Adobe los vídeos no pueden presentarse en este PDF [18], por esa razón se adjunta un enlace a un repositorio donde serán de acceso libre `https://github.com/CarlosLLM1/Kriging-Calidad-Aire-Gallego`. Estos videos son importantes para la la interpretación de los resultados, presentada en la subsección siguiente, que se basará en su análisis.

4.4. Interpretación de Resultados y Conclusión

El primer contaminante del que se va a analizar su comportamiento es el SO_2. Como ya se puede observar en el mapa de la Figura 4.5 la región con mayor presencia de este contaminante en su aire es Lugo. Este contaminante es liberado por procesos de combustión además de ser usado de forma extensiva por la industria química aunque lo más probable es que su presencia en Lugo se deba a su relación con la agricultura y la ganadería, usado principalmente en la fumigación [28] pero puede también ser el producto de la reacción de fertilizantes que contienen azufre con el suelo o bacterias presentes en él. Esta relación con el medio natural puede ser la causante de su estabilidad independientemente de la hora del día, lo que puede indicar que es un contaminante que no está muy relacionado con la actividad humana diaria. Su comportamiento en el resto de la comunidad así lo muestra, pues, aunque se aprecia un muy leve aumento de su concentración conforme avanza el día, la diferencia es mínima.

Un contaminante que está más estrechamente relacionado con la actividad humana es el NO_x, el mapa de este contaminante si que proporciona información de más interés debido, primero, a su distribución geográfica. Las zonas con mayor concentración del mismo son siempre los núcleos urbanos de la comunidad, principalmente Coruña y Vigo. Pero lo que es de más interés no es como se puede detectar claramente la presencia humana y movimiento en las grandes ciudades sino que refleja claramente lo vaciadas que están algunas zonas de la comunidad principalmente el sur de Orense y el norte de Lugo. Está tendencia también se aprecia a lo largo del día aunque de madrugada las diferencias entre diferentes puntos de la comunidad son menores. A partir de las siete de la mañana, hora aproximada a la que empieza el movimiento de personas, las principales ciudades cambian rápidamente de color hasta alcanzar un pico a mediodía para comenzar a decrecer hasta las seis de la tarde cuando vuelve a aumentar la concentración de NO_x. Esto se puede deber a la relación de este contaminante con la combustión [16], que lo convierte en una clara señal de la actividad humana y su desplazamiento basado en vehículos de combustión, lo que hace probable que los picos a la primera hora y última del día estén relacionados con el movimiento de personas.

Por último, el PM_{10} se comporta de forma similar al NO_x aumentando por la mañana, en

este caso a partir de las ocho, la diferencia sustancial es que no decrece a la misma velocidad que el NO_x, lo que probablemente sea debido a la naturalidad del PM_{10}. Este contaminante representa las partículas microscópicas suspendidas en el aire que pueden permanecer en él, dependiendo de distintos fenómenos atmosféricos, hasta varias horas [24]. Estas partículas provienen principalmente de actividades humanas, aunque no exclusivamente, como construcción, procesos industriales, el tránsito de coches por carretas tanto pavimentadas como no y de procesos de combustión en general [24].

Para finalizar el trabajo, se presentan también en el repositorio `https://github.com/CarlosLLM1/Kriging-Calidad-Aire-Gallego`, los vídeos de interpolación de un día aleatorio, se tomó el 12 de Enero. El objetivo es comprobar si los datos se asemejan a los de la media, dado que no se ha hecho estudio de datos atípicos en este trabajo podría ser que los datos estuviesen mal representados por sus medias. Sin embargo, como se puede comprobar en dichos vídeos, parece que los datos si se comportan según lo esperado, sobre todo el NO_x y el PM_{10} aunque es cierto que las diferencias regionales son mayores al no estar calculadas sobre la media, sino sobre el valor "real".

El comportamiento que se sale más de la norma es el del SO_2 aunque se mantiene bastante estable y en valores bajos en la mayor parte de la comunidad, lo cierto es que en Lugo su valor no es tan estable. Presenta dos picos, uno entre las 14-15 horas y otro entre las 21-22 horas, este comportamiento es extraño y no aparece tan claramente en el caso de la media por día laborable de Enero, lo que sugiere una fuente de emisión puntual difícil de explicar.

Bibliografía

[1] Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R*. Springer.

[2] Chiles, J., & Delfiner, P. (1999). *Geostatistics: Modeling spatial uncertainty*. Wiley-Interscience.

[3] Cressie, N. A. C. (1993). *Statistics for spatial data*. John Wiley & Sons, Inc.

[4] Datosmacro.com. (n.d.). *Galicia*. Expansion.com. Recuperado el 25 de enero de 2025, de `https://datosmacro.expansion.com/ccaa/galicia#:~:text=Galicia%20con%20una%20superficie%20de,resto%20de%20las%20Comunidades%20Aut%C3%B3nomas.`

[5] Díez, F. B., Tenías, J. M., & Pérez-Hoyos, S. (1999). Efectos de la contaminación atmosférica sobre la salud: una introducción. Revista Española de Salud Pública, 73(2), 109-121.

[6] Edward Isaaks. (28 de febrero de 2013). *What the Heck is a Variogram?*.[Archivo de Video]. Disponible en: https://www.youtube.com/watch?v=SJLDlasDLEU

[7] European Environment Agency. (2022). *Air quality in Europe 2022: Sources and emissions of air pollutants in Europe*. Recuperado de `https://www.eea.europa.eu/publications/air-quality-in-europe-2022/sources-and-emissions-of-air`.

[8] Gneiting, T., Sasvári, Z., & Schlather, M. (2000). Analogies and correspondences between variograms and covariance functions. *NRCSE Technical Report Series*, NRCSE-TRS No. 056. National Research Center for Statistics and the Environment.

[9] Gorai, A. K., Jain, K. G., Shaw, N., Tuluri, F., & Tchounwou, P. B. (2015). Kriging analysis for spatio-temporal variations of ground level ozone concentration. *Asian Journal of Atmospheric Environment, 9*, 247–258.

[10] Gräler, B., Rehr, M., Gerharz, L., & Pebesma, E. (2013). *Spatio-temporal analysis and interpolation of PM10 measurements in Europe for 2009*. ETC/ACM Technical Paper 2012/8 (revised version). European Topic Centre on Air Pollution and Climate Change Mitigation (ETC/ACM).

[11] Isaaks, E. H. & Srivastava, R. M. (1989). *An Introduction to Applied Geostatistics*. Oxford University Press.

[12] Kadane, J. B., Matheron, G., & Hasofer, A. M. (1990). Estimating and Choosing: An Essay on Probability in Practice. *Journal Of The American Statistical Association*, 85(412), 1167.

[13] Loots, C., Vaz, S., Planque, B., & Koubbi, P. (2010). Spawning distribution of North Sea plaice and whiting from 1980 to 2007. *Journal Of Oceanography, Research And Data, 3*, 77-95.

[14] Matheron, G. (1971). *The theory of regionalized variables and its applications*. Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau, No. 5.

[15] MeteoGalicia. (2025). *Calidad del aire en Galicia*. Recuperado el 7 de enero de 2025, de `https://www.meteogalicia.gal/web/ica/portada`

[16] Ministerio para la Transición Ecológica y el Reto Demográfico, Óxidos de Nitrógeno, disponible en: `https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/glosario-de-terminos/glosario-contaminantes/oxidos-nitrogeno.html`, [Recuperado el 24 de enero de 2025].

[17] Oliver, M. A., & Webster, R. (2015). Basic Steps in Geostatistics: The Variogram and Kriging. En *SpringerBriefs in agriculture*

[18] Overleaf. (2025). *How can I embed a video in my PDF using LaTeX?* Overleaf Documentation. Recuperado el 28 de enero de 2025, de `https://www.overleaf.com/learn/latex/Questions/How_can_I_embed_a_video_in_my_PDF_using_LaTeX%3F`

[19] Pebesma, E., & Graeler, B. (2024). *gstat: Spatial and Spatio-Temporal Geostatistical Modelling, Prediction and Simulation (Version 2.1-2)* [Paquete R]. Recuperado de `https://cran.r-project.org/web/packages/gstat/gstat.pdf`

[20] Rivera-González, L. O., Zhang, Z., Sánchez, B. N., Zhang, K., Brown, D. G., Rojas-Bracho, L., Osornio-Vargas, A., Vadillo-Ortega, F., & O'Neill, M. S. (2015). An assessment of air pollutant exposure methods in Mexico City, Mexico. *Journal of the Air & Waste Management Association*, *65*(5), 581–591.

[21] Serres, M. (1968). *Le système de Leibniz et ses modèles mathématiques*. Paris: Presses Universitaires de France.

[22] Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. Economic Geography, 46, 234.

[23] Tolosana-Delgado, R. (2005). Geostatistics for constrained variables: positive data, compositions and probabilities. Applications to environmental hazard monitoring. *Geography.*

[24] U.S. Environmental Protection Agency. (2023, Junio 29). *Overview of Particulate Matter (PM) Air Quality in the United States.* U.S. EPA. Recuperado de `https://www.epa.gov/system/files/documents/2023-06/PM_2022.pdf`

[25] Van Zoest, V., Osei, F. B., Hoek, G., & Stein, A. (2019). Spatio-temporal regression kriging for modelling urban NO_2 concentrations. *Spatial Statistics, 32*, 851-865.

[26] Wackernagel, H. (2003). *Multivariate geostatistics: an introduction with applications.* Springer.

[27] Wang, D., Ding, W., Lo, H., Morabito, M., Chen, P., Salazar, J., & Stepinski, T. (2013). Understanding the spatial distribution of crime based on its related variables using geospatial discriminative patterns. *Computers, Environment and Urban Systems, 39*, 93–106.

[28] Yong-Biao Liu. *Sulfur Dioxide Fumigation for Postharvest Control of Mealybugs on Harvested Table Grapes.* Journal of Economic Entomology, vol. 112, no. 2, pp. 597–602, April 2019.

Anexo I

Código Fuente

I.1. Construcción de la Base de Datos

```
#A modo de ejemplo se realizarán todos los casos con las estaciones de Lugo.
matriz <- matrix(nrow = 43, ncol = 4)
lugo <- rep("Lugo", 8)
matriz[c(26:33), 2] <- lugo # Asignamos las provincias a la matriz.

#Ahora escribimos el nombre de las localizaciones a mano.
nombres_lugo <- c("Cuiña", "Escola_Música", "Est_Ou", "Fingoi", "Mourence",
                  "Rio_Cobo", "Sur", "Xove")

#Asignamos las localizaciones en función de la provincia en la que se encuentran.
matriz[c(26:33), 1] <- nombres_lugo

lugo_coordenadas <- matrix(nrow = 8, ncol = 2)#Coordenadas de Lugo.
lugo_coordenadas[1, ] <- c(42.9968594, -7.900446)
# ...
lugo_coordenadas[8, ] <- c(43.69375, -7.50398)

matriz[c(26:33), 3] <- lugo_coordenadas[, 1]#Añadimos las coordenadas a la matriz.
matriz[c(26:33), 4] <- lugo_coordenadas[, 2]

#Ahora, en una carpeta se almacenan los dataset individuales de cada estación,
#leeremos cada uno de los archivos y asociamos la información que contienen con su ubicación.
```

```
lista_datos <- list()
nombres_archivos <- matriz[, 1]

#Bucle que lee y procesa cada dataset.
for (nombre_archivo in nombres_archivos) {
  archivo <- paste0(ruta_csv, nombre_archivo, ".csv")
  if (file.exists(archivo)) {
    datos <- read.csv(archivo, header = FALSE, sep = ";", stringsAsFactors = FALSE) #Leemos.
    datos <- datos[-1, , drop = FALSE] #Eliminamos la primera fila del dataset.
    colnames(datos) <- datos[1, ]
    datos <- datos[-1, , drop = FALSE]

    if ("Data" %in% colnames(datos)) { #Renombramos la columna "Data" como "Hora".
      colnames(datos)[colnames(datos) == "Data"] <- "Hora" }

    #Buscamos en la matriz generada la información sobre la ubicación de la estación.
    loc <- matriz[matriz[, 1] == nombre_archivo, 1]
    #...
    lat <- matriz[matriz[, 1] == nombre_archivo, 4]

    #Añadimos las columnas Localización, Provincia, Longitud y Latitud a los datos.
    datos <- datos %>% mutate(
      Localizacion = loc,
      #...,
      Latitud = lat   )
    lista_datos[[nombre_archivo]] <- datos
  } else {
    warning(paste("El archivo", archivo, "no existe.")) }}
matriz_combinada <- bind_rows(lista_datos) #Combinamos en un solo dataframe.
```

I.2. Análisis Exploratorio de la Base de Datos

```
#Porcentaje de valores NA por columna.
columnas_a_eliminar <- c("Hora", [...], "Latitud")#variables de ubicacion
matriz_combinada_filtrada <- matriz_combinada[, !(names(matriz_combinada) %in%
columnas_a_eliminar)]

porcentaje_na_columnas_filtradas <- sapply(matriz_combinada_filtrada, function(x)
sum(is.na(x)) / length(x) * 100)
umbral <- 50#Definimos el 50% como el umbral de porcentaje elevado.

columnas_altos_na <- data.frame(
  Variable = names(porcentaje_na_columnas_filtradas)[porcentaje_
  na_columnas_filtradas > umbral],
  Porcentaje_NA = porcentaje_na_columnas_filtradas[porcentaje_
  na_columnas_filtradas > umbral])

#Como podemos ver, la mayoría de valores aun en presencia de valores en las columnas
#problemáticas no son NA para las demás variables, pero atendamos a los NO.
ubicaciones_totales <- unique(matriz_combinada$Localizacion)
#Filtramos las estaciones donde NOx o ambos NO y NO2 no se registran.
ubicaciones_con_nox <- unique(matriz_combinada[!is.na(matriz_combinada$`NOX Âµg/mÂ³`) |
                (!is.na(matriz_combinada$`NO Âµg/mÂ³`) &
                !is.na(matriz_combinada$`NO2 Âµg/mÂ³`)), ]$Localizacion)
ubicaciones_sin_nox <- setdiff(ubicaciones_totales, ubicaciones_con_nox)
print(ubicaciones_sin_nox)
#Lo hacemos de forma equivalente para SO2 y PM10.
#Comprobamos si las estaciones que no muestrean estas variables coinciden.
coinciden_nox_so2 <- intersect(ubicaciones_sin_nox, ubicaciones_sin_so2)
#...
coinciden_so2_pm10 <- intersect(ubicaciones_sin_so2, ubicaciones_sin_pm10)
```

I.3. Construcción Final de la Base de Datos con Media por Horas del Día

```
library(lubridate)

#El primer problema es el formato de la hora que nos da la fecha y hora en la misma celda
#pero nos interesa convertirlo a formato datetime para ello utilizamos la librería lubridate.
matriz_combinada$Hora <- dmy_hm(matriz_combinada$Hora)
matriz_combinada$Dia <- day(matriz_combinada$Hora)
matriz_combinada$Hora_Del_Dia <- hour(matriz_combinada$Hora)

#Tomamos solo las variables que hemos considerado de interés en el análisis anterior.
datos_final <- matriz_combinada[, c("Hora_Del_Dia", "SO2 Âµg/mÂ³", "PM10 Âµg/mÂ³", "Localizac
                                    "Provincia", "Longitud", "Latitud", "NOX Âµg/mÂ³", "Dia")
dias_festivos <- c(1,2,3,4,5,6,7,13,14,20,21,27,28)#Tomamos una lista de los días festivos
datos_laborales <- datos_final[!datos_final$Dia %in% dias_festivos, ]

#Ahora lo que queremos es calcular la media por hora y por estación.
datos_laborales$`SO2 Âµg/mÂ³` <- as.numeric(datos_laborales$`SO2 Âµg/mÂ³`)
#...
#Tomamos todas las posibles combinaciones de horas y estaciones.
combinaciones <- unique(datos_laborales[, c("Hora_Del_Dia", "Localizacion")])
resultado <- data.frame(#Creamos un data frame para almacenar los resultados
  Hora_Del_Dia = integer(),
  ...
  NOX_media = numeric(),
  stringsAsFactors = FALSE )
for (i in 1:nrow(combinaciones)) {#Iteramos respecto a las posibles combinaciones.
  hora <- combinaciones$Hora_Del_Dia[i] #Valor de la Hora.
  localizacion <- combinaciones$Localizacion[i]#Valor de la localización.
  #Y tomamos los datos para las características de la combinación actual.
  subset_datos <- datos_laborales[datos_laborales$Hora_Del_Dia == hora &
                                  datos_laborales$Localizacion == localizacion, ]
  #Calculamos las medias ignorando los valores Nan.
  SO2_media <- mean(subset_datos$`SO2 Âµg/mÂ³`, na.rm = TRUE)
  PM10_media <- mean(subset_datos$`PM10 Âµg/mÂ³`, na.rm = TRUE)
  NOX_media <- mean(subset_datos$`NOX Âµg/mÂ³`, na.rm = TRUE)
```

```
  #Y se agregan al dataframe con las correspondientes medias.
  resultado <- rbind(resultado, data.frame(
    ...
    NOX_media = NOX_media,
    stringsAsFactors = FALSE))}
```

I.4. Construcción del Grid de Galicia a partir de un Shapefile

```
#Primero al igual que para los mapas de la media del punto anterior cargamos los mapas.
mapa_galicia <- st_read("C:/[...]/TFG/Provincias/Provincias_IGN.shp")
mapa_galicia_union <- st_union(mapa_galicia)#Combinamos todas las provincias,
#porque sino a la hora de crear el grid nos da error.
cell_size <- 500  #Creamos el grid con el tamaño por celda de 500 metros.
grid_galicia <- st_make_grid(mapa_galicia_union, cellsize = cell_size, what = "centers")
grid_galicia_sf <- st_sf(geometry = grid_galicia) #Y lo convertimos en un objeto sf.
grid_galicia_sf <- grid_galicia_sf[st_intersects(grid_galicia_sf, mapa_galicia_union, sparse =
#Visualizamos.
plot(st_geometry(mapa_galicia), col = "gray95", border = "black")
plot(st_geometry(grid_galicia_sf), add = TRUE, col = "blue", pch = 20, cex = 0.5)

#Tenemos problemas con la estructura de las coordenadas, por esa razón transformamos
#el grid de Galicia al CRS de los datos, ETRS89 / UTM zone 29N (en metros).
grid_galicia_sp <- st_transform(grid_galicia, crs = st_crs(puntos_datos_sp_nox))
```

I.5. Interpolación Espaciotemporal mediante Kriging

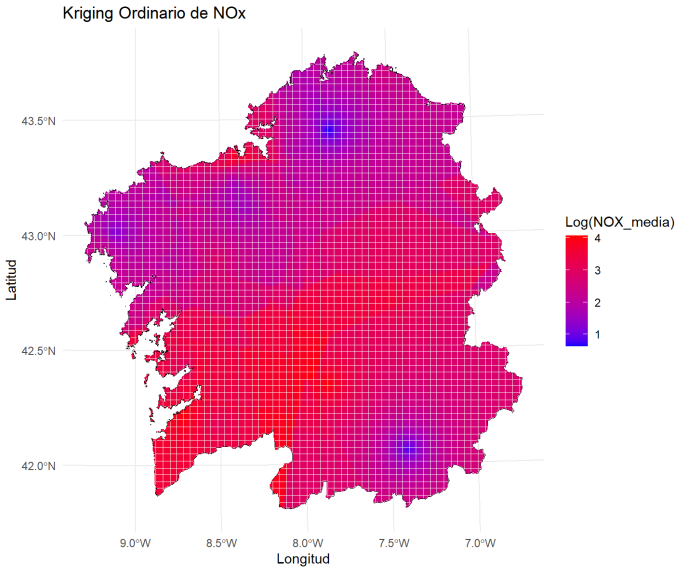
```
#Creamos un bucle para iterar sobre todas las horas del día para el SO2.
for (hora in 0:23) {
  #Filtramos los datos para la hora específica.
  data_hora <- data_interpo %>% filter(Hora_Del_Dia == hora & !is.na(SO2_media))

  #Ajustamos el variograma para la hora especifica.
  vario_cloud_so2 <- variogram(log(SO2_media) ~ 1, data_hora, cloud = TRUE, cutoff = 120000)
  vario_sample_so2 <- variogram(log(SO2_media) ~ 1, data_hora, cutoff = 120000)
```

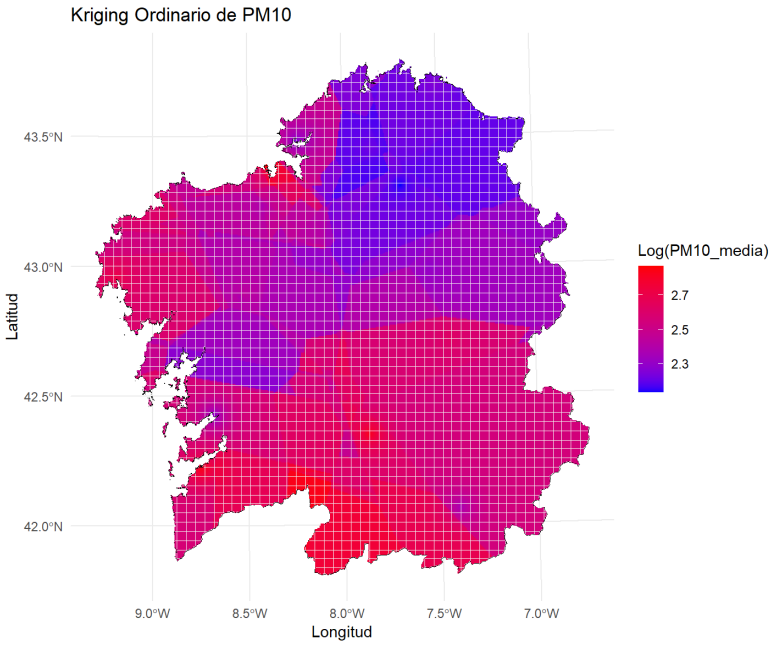
```
  modelo_variograma_so2 <- fit.variogram(
    vario_sample_so2,
    model = vgm(psill = 0.7, model = "Exp", range = 50000, nugget = 0.2) )

  puntos_datos_sp_so2 <- as(data_hora, "Spatial")
  puntos_datos_sp_so2 <- puntos_datos_sp_so2[!duplicated(coordinates(puntos_datos_sp_so2)), ]
  #Hacemos la interpolación Kriging.
  resultado_kriging_so2 <- krige( formula = log(SO2_media) ~ 1,
    locations = puntos_datos_sp_so2, newdata = grid_galicia_sf,
    model = modelo_variograma_so2,  nmin = 1, nmax = 5 )

  resultado_kriging_so2_sf <- st_as_sf(resultado_kriging_so2)
  #Generamos el gráfico.
  plot <- ggplot() +
    geom_sf(data = mapa_galicia_union, fill = "gray95", color = "black") +
    geom_sf(data = resultado_kriging_so2_sf, aes(color = var1.pred), size = 0.1) +
    scale_color_gradient_fixed("Log(SO2_media)", global_min_so2, global_max_so2) +
    labs(title = paste("Kriging Ordinario de SO2 - Hora:", hora),
      x = "Longitud",
      y = "Latitud") +
    theme_minimal() +
    theme(plot.background = element_rect(fill = "white", color = NA),
      panel.background = element_rect(fill = "white", color = NA),
      plot.title = element_text(hjust = 0.5, face = "bold", size = 14))

  #Guardamos el archivo.
  ggsave(filename = paste0("kriging_so2_hora_", hora, ".png"), plot = plot, width = 8,
    height = 6)
}
```

Anexo II

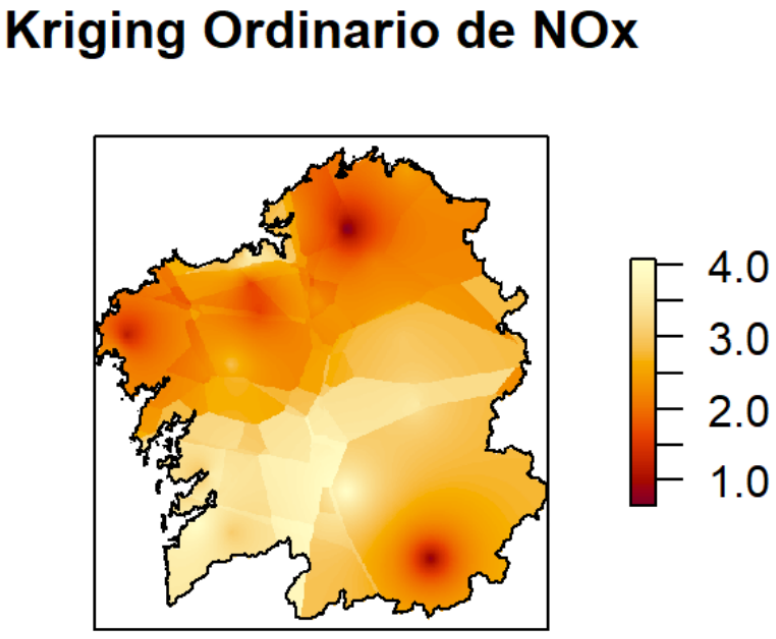
Mapas

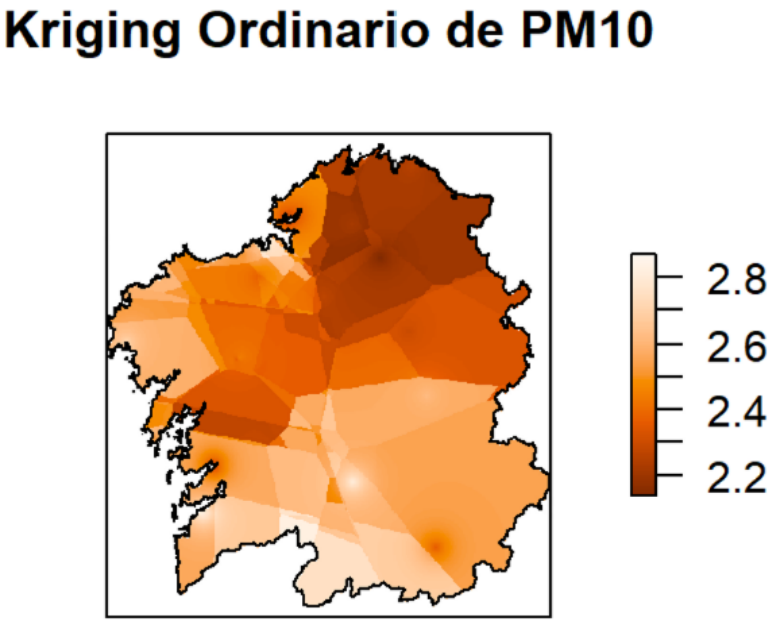
II.1. Gráficas con Grid del Kriging Ordinario





II.2. Gráficas Continuas del Kriging Ordinario




Kriging Ordinario de PM10
2.8
2.6
2.4
2.2