Estadística espacial para datos punto referenciados

Ariadna Beltrán Rodríguez

Trabajo de fin de grado de Matemáticas
Universidad de Zaragoza

Directores del trabajo: Ana Carmen Cebrián Guajardo
y Jorge Castillo-Mateo

4 de septiembre de 2023

Prólogo

La metodología espacial y espacio-temporal ha sido utilizada de forma creciente (especialmente durante los últimos 50 años) para resolver problemas en muchos campos. Los orígenes de la estadística espacial se remontan a principios del siglo XX. En esa época, los científicos comenzaron a reconocer que muchos fenómenos naturales y sociales exhibían patrones espaciales y que el análisis tradicional de datos estadísticos no era adecuado para abordar estos aspectos espaciales. En la década de 1920, el estadístico británico Karl Pearson introdujo el concepto de 'autocorrelación' para medir la dependencia espacial entre las observaciones. Sin embargo, el desarrollo de la estadística espacial como disciplina independiente tuvo que esperar hasta la década de 1950 y 1960. En la década de 1970, los avances en la computación y la tecnología permitieron el desarrollo de métodos más sofisticados para el análisis espacial. Se introdujeron técnicas como el kriging, que es un método de interpolación espacial utilizado para estimar valores en lugares no muestreados, y los procesos puntuales, que se utilizan para modelar la distribución de eventos puntuales en el espacio.

En los últimos años, la estadística espacial ha experimentado avances significativos. El aumento en la disponibilidad de datos geoespaciales a gran escala (Big Geospatial Data) ha impulsado la necesidad de desarrollar técnicas avanzadas para analizar volúmenes masivos de información de ubicación. Esto ha llevado a un enfoque en el análisis en tiempo real y al surgimiento de la inteligencia artificial geoespacial (GeoAI), donde la inteligencia artificial se combina con análisis espacial para detectar patrones y generar modelos predictivos más precisos. Además, la existencia de bibliotecas de programación y sistemas de información geográfica (SIG) ha permitido a profesionales de diferentes campos llevar a cabo análisis espaciales más avanzados sin requerir un profundo conocimiento técnico.

El principal objetivo del presente trabajo es desarrollar la teoría de la estadística espacial en el ámbito de los datos punto referenciados. En el Capítulo 1, se comienza introduciendo brevemente los procesos espaciales y sus distintas aplicaciones. En el Capítulo 2, se presentarán los diferentes conceptos básicos de los modelos para puntos referenciados, y, tras introducir la definición del variograma, se exponen distintas formas para llevar a cabo su estimación. A continuación, en el Capítulo 3, se desarrolla la teoría y aplicación práctica del kriging ordinario y universal, metodologías fundamentales en la estimación de valores en ubicaciones no muestreadas. Finalmente, en el Capítulo 4, se ilustrarán algunos de los procedimientos vistos mediante dos conjuntos de datos de temperatura mensual.

Antes de dar paso al trabajo, agradecer a mis directores de TFG, Ana Carmen Cebrián y Jorge Castillo-Mateo por su implicación a lo largo del trabajo. También me gustaría nombrar en estas lineas a mi familia y amigos que me han acompañado y ayudado a lo largo de estos años y sin los que esta etapa de mi vida no hubiera sido lo mismo.

Summary

Spatial statistics focuses on the analysis of geographically located data, aiming to identify spatial patterns and relationships. It helps to understand how data are distributed in space and is applied in fields such as geography, ecology, and epidemiology. It employs statistical techniques to model and make decisions based on spatial information.

Let us summarize the contents of this project. It begins with a brief introduction in Chapter 1 about spatial statistics. Three different types of spatial processes are introduced: point-referenced data, areal data, and models for point processes. These are distinguished from each other depending on the assumptions made on their domain D, which represents the study region. Additionally, some of the many applications of spatial statistics in current times are briefly discussed, including environmental sciences, epidemiology, and public health.

The second chapter, based on the work made by Michael Sherman [5] and Cressie [4] in their respective books, is dedicated to models for point-referenced data. The purpose of this section is to review the main and most important concepts of this type of models. We begin by introducing the concept of stationarity, which is essential to facilitate the interpretation and prediction of data at different spatial locations. Next, we define the variogram and the covariance function, which are the functions commonly used for modeling spatial dependence and are considered as tools of special interest in the process. In practice, it is usually used the variogram, not only because it is more general, but also due to its advantages in estimation. Furthermore, we will demonstrate some essential properties of the variogram and explain a way to obtain the variogram from the covariance function. Later, we will focus on variograms for isotropic processes and will describe the different types of models used in practice, such as the exponential or the sperical models, and the most well-known, the Matern model. The different parameters of the variogram will also be introduced.

In the second half of Chapter 2, the explanation of variogram estimation, which is crucial to understanding the spatial structure of the data, is carried out. Firstly, we will briefly discuss some of the non-parametric estimators of the variogram, including the most common one, the method of moments estimator. Next, we will introduce parametric estimation, which involves finding a valid parametric model that adequately describes the spatial dependence present in the data. Several methods of adjustment are employed for this purpose. In our work, we will develop the least squares methods and maximum likelihood estimation. Lastly, we will introduce the estimation with a non-constant mean function.

In Chapter 3, the concepts related to spatial prediction or kriging are developed, which involve the process of estimating values of a spatial process at unsampled locations within a geographical area using information from previously sampled locations. This chapter is again based on the book written by Michael Sherman [5].

In the first part, the model for optimal prediction and the objective of kriging, which is to obtain the best prediction that minimizes the mean squared error, are explained in detail. Next, we will distinguis three types of kriging, among the many that exist. Firstly, simple kriging assumes that the mean is known. Then, ordinary kriging assumes that the mean is constant and unknown, and also that the variogram exists and is known. Lastly, universal kriging assumes that the mean is unknown and not constant, but rather a linear combination of $(p+1)$ explanatory variables. The kriging variance will also be computed.

At the end of the chapter, cross-validation is explained, which is a technique commonly used in spatial models to diagnose whether a model captures the spatial variability of the data in an adequate way. This involves removing some data points and using the remaining data to predict the removed observation. Specifically, the simplest version of cross-validation, known as "leave-one-outçross-validation, will be explained. As examples of methods to measure the accuracy of predictions, the root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) will be briefly explained.

Finally, in chapter 4, we will apply some of the concepts developed throughout the study. For this purpose, we will use two sets of average temperatures for the months of January and August obtained from [11]. First, we will conduct a brief analysis of these two data sets, and will study the spatial correlation present in both of them.

Next, a spatial model will be proposed in order to compare the temperatures of both months, and will introduce the covariates of elevation, distance to the coast, and an interaction of elevation and distance to the coast. With all of this, sample variograms will be computed, and subsequently, we will attempt to find a parametric variogram model that fits the data better. To achieve this, we will use the OLS and WLS fitting methods, and finally, we will perform cross-validation.

Finally, ordinary and universal kriging will be computed for each data set using the best models obtained previously. We will also carry out several interpretationes of the obtained results throughout the chapter.

Índice general

Capítulo 1

Introducción

1.1. Estadísica espacial y procesos espaciales

De manera general, los *datos espaciales* son aquellos que tienen asociada una localización en el espacio. Consecuentemente, y ampliando la idea anterior, un dato espacio-temporal es simplemente la observación de una variable en cierta localización espacial considerando al tiempo como coordenada adicional. Las variables espacio-temporales de interés pueden ser continuas o discretas, y se presupone que existe correlación entre dos variables que tengan asociada distinta referencia geográfica. La idea fundamental con este tipo de datos es que las observaciones más cercanas son más parecidas entre sí, y conforme éstas se distancian, la correlación entre las variables tiende a disminuir, anulándose en algún momento.

La estadística espacial es el conjunto de modelos y métodos que tienen por objetivo el análisis de datos espacialmente referenciados. De manera más formal se puede decir que la estadística espacial trata con el análisis de realizaciones de un proceso estocástico $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$, en el que $\mathbf{s} \in \mathbb{R}^d$ representa una ubicación en el espacio euclídeo d-dimensional y varía sobre un conjunto $D \subset \mathbb{R}^d$. Además, $Z(\mathbf{s})$ es una variable aleatoria en la ubicación $\mathbf{s}$.

Los modelos para datos espaciales tienen una estructura simple pero lo suficientemente flexible para manejar una clase extremadamente grande de problemas. Los datos pueden ser continuos o discretos, pueden ser agregaciones espaciales u observaciones en puntos del espacio, sus ubicaciones espaciales pueden ser regulares o irregulares, y esas ubicaciones pueden ser de un conjunto espacial continuo o discreto.

1.2. Aplicaciones de los procesos espaciales

Uno de los objetivos principales de la estadística espacial es la predicción. Los modelos espaciales se utilizan para realizar predicciones en ubicaciones no muestreadas. Estos modelos capturan la autocorrelación espacial y utilizan la información de las observaciones cercanas para estimar valores en lugares no muestreados. Algunas de sus aplicaciones más importantes son las siguientes.

- **Ciencias medioambientales:** Se emplean para evaluar el impacto de actividades humanas en el medio ambiente. Permiten analizar la dispersión de contaminantes, la fragmentación del hábitat, el cambio de uso del suelo y la evaluación de la calidad ambiental. Esto ayuda en la toma de decisiones para la conservación de recursos naturales y la planificación sostenible. También se usan en la predicción de variables ambientales, como la calidad del aire, la concentración de contaminantes o la distribución de especies.

- **Epidemiología y salud pública:** Los modelos espaciales se utilizan en la epidemiología para analizar la propagación de enfermedades, identificar áreas de alta incidencia de enfermedades, estudiar factores de riesgo espaciales y evaluar la efectividad de intervenciones de salud pública.

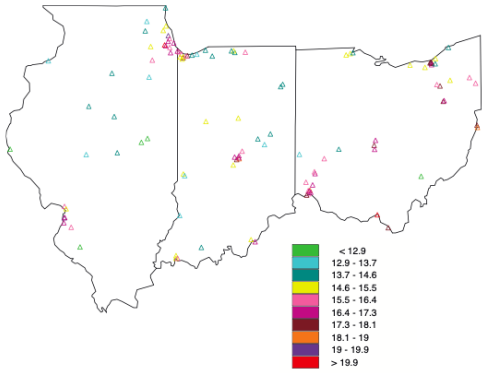

Figura 1.1: Mapa de sitios de muestreo de PM2.5 en tres estados del medio oeste de los Estados Unidos; el carácter de trazado indica el rango del nivel promedio de PM2.5 monitoreado durante el año 2001. Fuente: [1, cap. 1].

También existen aplicaciones en el análisis de mercado, donde estos modelos ayudan en la toma de decisiones de marketing y estrategias comerciales. Se emplean también en la planificación urbana para la ubicación óptima de infraestructuras, la evaluación de accesibilidad y la distribución de servicios públicos. Estas son solo algunas de las aplicaciones de los modelos espaciales, y su utilidad se extiende a muchos otros campos. Se pueden consultar más detalladamente sus diversas aplicaciones en los libros [2] y [3].

1.3. Tipos de procesos espaciales

En el campo de la estadística espacial pueden aparecer tres tipos de datos, que requieren tipos de análisis y herramientas diferentes. Las distintas clases de datos son descritas a continuación.

Para el desarrollo de esta sección se ha seguido [1, cap. 1].

Datos punto-referenciados

En este tipo de datos, $Z(\mathbf{s})$ es un vector aleatorio en una ubicación $\mathbf{s} \in \mathbb{R}^d$, donde $\mathbf{s}$ varía continuamente sobre D, un subconjunto fijo de $\mathbb{R}^d$ que contiene un rectángulo d-dimensional de volumen positivo. Este caso es generalmente conocido como datos *geoestadísticos*.

Podemos observar un ejemplo de este caso en la Figura 1.1, en la cual se muestran las ubicaciones de 114 sitios de monitorización de la contaminación del aire en tres estados del medio oeste de Estados Unidos (Illinois, Indiana y Ohio). El carácter del trazado indica el nivel promedio anual de PM2.5 en 2001 (medido en ppb) en cada sitio. PM2.5 representa las partículas en suspensión de menos de 2.5 micrones de diámetro y es una medida de la densidad de partículas muy pequeñas que pueden viajar a través de la nariz y la tráquea hasta los pulmones, potencialmente dañando la salud de una persona. Aquí podríamos estar interesados en un modelo de distribución geográfica de estos niveles que tenga en cuenta la correlación espacial y posiblemente covariables subyacentes (industrialización regional, densidad del tráfico, entre otros). El uso de colores facilita la lectura en cierta medida, ya que el color permite ordenar las categorías de manera más natural y ayuda a resaltar el contraste entre las áreas urbanas y rurales.

Datos de área

En este tipo de puntos, D es nuevamente un subconjunto fijo (de forma regular o irregular), pero ahora particionado en un número finito de unidades de área con límites bien definidos. Este caso a menudo se denomina datos de rejilla (lattice), un término que encontramos engañoso ya que sugiere observaciones correspondientes a «esquinas» de una cuadrícula similar a un tablero de ajedrez. Por supuesto, hay

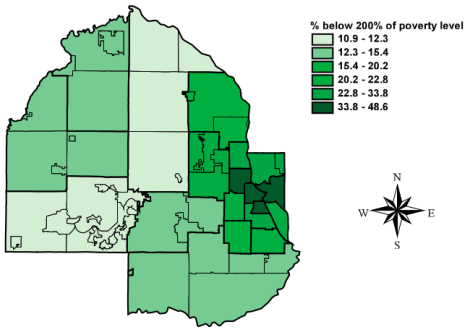

Figura 1.2: Mapa que muestra el porcentaje de la población encuestada con ingresos familiares por debajo del 200 % del límite de pobreza federal, en unidades de encuesta regionales en el condado de Hennepin, Minnesota. Fuente: [1, cap. 1].

conjuntos de datos de este tipo; por ejemplo, los que surgen de ensayos agrícolas de campo (donde las parcelas cultivadas forman una rejilla regular).

Sin embargo, en la práctica, la mayoría de los datos de área son resúmenes sobre una rejilla irregular, como una colección de límites de condados u otras regiones, como se muestra en la Figura 1.2. Dicho ejemplo proporciona información sobre el porcentaje de una población encuestada con ingresos familiares por debajo del 200 % del límite de pobreza federal para una colección de regiones que comprenden el condado de Hennepin, MN. Es importante destacar que no tenemos información sobre ningún hogar individual en el área de estudio, solo resúmenes regionales para cada región. La Figura 1.2 es un ejemplo de un mapa coroplético, lo que significa que utiliza tonos de color (o escala de grises) para clasificar los valores en unas pocas categorías amplias (seis en este caso), similar a un histograma para datos no espaciales.

En este caso, las ubicaciones $\mathbf{s} \in D$ son en realidad las regiones (o bloques) en sí mismos, por lo tanto no se designarán como $\mathbf{s}_i$, sino como B_i con $i = 1, \ldots, n$, para evitar confusiones entre los puntos $\mathbf{s}_i$ y los bloques B_i.

Modelos para procesos puntuales

Ahora, el conjunto D en sí mismo es aleatorio; su conjunto de índices proporciona las ubicaciones de sucesos aleatorios que constituyen el patrón espacial de puntos. El proceso $Z(\mathbf{s})$ en sí mismo puede ser simplemente igual a 1 para todos los $\mathbf{s} \in D$ (indicando la ocurrencia del evento), o posiblemente proporcionar información adicional de covariable (produciendo un proceso de patrón de puntos marcado).

Ejemplos de este caso pueden ser las residencias de personas que sufren de una enfermedad particular, o la ubicación de una especie de árbol en un bosque. Aquí, la respuesta Z suele ser fija (ocurrencia del evento), y solo las ubicaciones se consideran aleatorias. En algunos casos, esta información podría complementarse con la edad u otra información de covariable, dando lugar a un patrón de puntos marcados. Este tipo de datos suelen ser de interés en estudios de agrupación de sucesos, donde el objetivo es determinar si un patrón de puntos espaciales observado es un ejemplo de un proceso agrupado o simplemente el resultado de un proceso de eventos aleatorios que operan de manera independiente y homogénea en el espacio.

En este trabajo nos centraremos en el estudio de modelos espaciales para datos punto-referenciados. Estos modelos son generalmente conocidos con el término *geoestadística*. Los orígenes de esta rama de las matemáticas aplicadas se remontan a principios de la década de 1950 con un intento por mejorar los métodos para calcular las reservas en las minas de oro en Sudáfrica y que posteriormente llamaron la atención de los ingenieros de minas franceses, en especial de Georges Matheron, quien formalizó los conceptos.

Capítulo 2

Modelos para puntos georreferenciados

Los datos georreferenciados son una fuente valiosa de información en muchos campos, desde la geología hasta la epidemiología. Estos datos representan la ubicación espacial de puntos de interés y pueden proporcionar información valiosa sobre la distribución y variabilidad de fenómenos naturales o sucesos sociales. Sin embargo, analizar y comprender la estructura espacial de estos datos puede resultar desafiante debido a la naturaleza compleja de su distribución. En este capítulo, nos centraremos en el análisis de puntos georreferenciados utilizando modelos espaciales. Estos modelos nos permiten estudiar la dependencia espacial y la variabilidad de los datos en función de la distancia y la dirección.
En primer lugar, se presentarán los conceptos básicos de la geoestadística, incluyendo la definición de variograma, la noción de estacionariedad e isotropía, y la importancia de la correlación espacial. A continuación, nos centraremos en los variogramas para procesos isotrópicos. Se explicarán los fundamentos teóricos del variograma y su interpretación, así como los diferentes tipos de modelos utilizados en la práctica. Estos modelos nos permiten caracterizar la estructura de dependencia espacial de los datos y proporcionan una base sólida para la predicción y la interpolación espacial. Por último, abordaremos la estimación del variograma, la cual se realiza a partir de los datos observados. Se presentarán diferentes métodos de estimación y se discutirán las ventajas y desventajas de cada uno, así como las consideraciones prácticas para su aplicación.

La bibliografía consultada para realizar este capítulo ha sido [1, cap. 2], [4, cap. 2] y [5, cap. 3].

2.1. Conceptos básicos

Asumimos que el proceso espacial $Z(\mathbf{s})$ tiene una media asociada, $\mu(\mathbf{s}) = E(Z(\mathbf{s}))$ y que su varianza existe para todo $\mathbf{s} \in D$.

Un proceso espacial $Z(\mathbf{s})$ se dice estacionario si mantiene sus propiedades estadísticas constantes en todas las ubicaciones dentro del área de estudio. La estacionariedad es una hipótesis que suele ser razonable y muy empleada en el caso de datos espaciales. A continuación presentamos tres tipos.

Definición 1. El proceso espacial $Z(\mathbf{s})$ es *estrictamente estacionario* (*estacionariedad fuerte*) si, para cada $n \geq 1$, cada conjunto de n puntos $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ en D y cada $\mathbf{h} \in \mathbb{R}^d$, la distribución de $(Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))$ es la misma que la de $(Z(\mathbf{s}_1 + \mathbf{h}), \ldots, Z(\mathbf{s}_n + \mathbf{h}))$.

Definición 2. Un proceso espacial $Z(\mathbf{s})$ se denomina *débilmente estacionario* (o *estacionariedad de segundo orden*) si $\mu(\mathbf{s}) \equiv \mu$ (es decir, tiene media constante) y $Cov(Z(\mathbf{s}), Z(\mathbf{s} + \mathbf{h})) = C(\mathbf{h})$ para todo $\mathbf{h} \in \mathbb{R}^d$.

En otras palabras, la estacionariedad fuerte implica que la función de distribución de probabilidad es invariante a cualquier traslación respecto a un vector $\mathbf{h}$. Por otro lado, la estacionariedad débil implica que la covarianza entre dos puntos distintos depende únicamente del vector de desplazamiento $\mathbf{h}$ y no de dichos puntos. Además, esta covarianza puede definirse mediante una función $C : D \to \mathbb{R}$ denominada *función de covarianza* que depende únicamente de $\mathbf{h}$. La existencia de la covarianza implica que la varianza existe y es finita, es decir, $Var(Z(\mathbf{s})) = C(0) = \sigma^2$.

La estacionariedad fuerte implica que toda la distribución es estacionaria, por lo tanto, en particular, los momentos de primer y segundo orden, siempre que existan, también van a ser estacionarios, obteniendo así la estacionariedad débil. El recíproco no es cierto generalmente, pero sí se cumple para procesos gaussianos. Recordemos que un proceso $Z(\mathbf{s})$ se dice que es gaussiano si, para cada $n \geq 1$ y para cada conjunto de puntos $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))^T$ tiene una distribución normal multivariante.

Existe un tercer tipo de estacionariedad llamada *estacionariedad intrínseca*, la cual impone condiciones sobre la media y la varianza de los incrementos $[Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})]$.

Definición 3. El proceso espacial $Z(\mathbf{s})$ se dice que es *estacionariamente intrínseco* si cumple que:

$$E\left[Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})\right] = 0, \tag{2.1}$$

$$Var\left(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})\right) = 2\gamma(\mathbf{h}). \tag{2.2}$$

La definición de estacionariedad intrínseca impone que la esperanza de la diferencia sea 0 y, en consecuencia, esto implica que la esperanza de $Z(\mathbf{s})$, si existe, es constante. Por otro lado, nótese que para cualquier vector $\mathbf{h}$, la varianza del incremento está definida y es una función que depende únicamente de la distancia y no de $\mathbf{s}$. Además, por (2.1) tenemos que $Var\left(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})\right) = E\left[Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})\right]^2$.

La definición de estacionariedad intrínseca lleva a introducir la definición de $\gamma(\mathbf{h})$.

Definición 4. Se denomina *variograma* a la funcíón $2\gamma(\mathbf{h})$ y *semivariograma* a la función $\gamma(\mathbf{h})$, definidas en (2.2). Nótese que es una funcion $\gamma : D \to \mathbb{R}$.

Tanto el variograma como la función de covarianza se emplean para la determinación de la dependencia espacial entre los datos observados. Veamos de forma intuitiva, cómo es el comportamiento del variograma. Para valores de $\|\mathbf{h}\|$ pequeños (distancias pequeñas), se espera que $Z(\mathbf{s}+\mathbf{h})$ y $Z(\mathbf{s})$ sean muy similares, es decir, que sean más dependientes y por lo tanto, la diferencia $(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s}))^2$ sea pequeña. Por otro lado, cuando $\|\mathbf{h}\|$ crece, se espera una menor similitud entre $Z(\mathbf{s}+\mathbf{h})$ y $Z(\mathbf{s})$, y por lo tanto $(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s}))^2$ será más grande. Luego el valor de $\gamma(\mathbf{h})$ se espera que crezca con $\|\mathbf{h}\|$, proporcionando así una idea sobre la dependencia espacial.

Estudiemos ahora algunas propiedades del variograma.

Teorema 2.1. *Sea la función $\gamma : D \to \mathbb{R}$, $\boldsymbol{h} \in \mathbb{R}^d$ y $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ un conjunto de n puntos en D. Se verifican las siguientes propiedades:*

a) $\gamma(\boldsymbol{h}) = \gamma(-\boldsymbol{h})$, $\gamma(\boldsymbol{h}) \geq 0$ *y* $\gamma(0) = 0$.

b) El variograma tiene la propiedad de ser definido negativo, es decir, para cada conjunto de puntos $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ y cada conjunto de constantes $a_1, \ldots, a_n$ tales que $\sum_i a_i = 0$, si $\gamma(\boldsymbol{h})$ es válido, entonces:

$$\sum_i \sum_j a_i a_j \gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) \leq 0. \tag{2.3}$$

Demostración. a) En primer lugar vamos a comprobar que $\gamma(\mathbf{h}) = \gamma(-\mathbf{h})$. Sabemos, por definición, que el variograma es $2\gamma(\mathbf{h}) = Var(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s}))$. Luego:

$$2\gamma(\mathbf{h}) = Var(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})) = Var(Z(\mathbf{s}+\mathbf{h})) + Var(Z(\mathbf{s})) - 2Cov(Z(\mathbf{s}+\mathbf{h}), Z(\mathbf{s})).$$

Como la varianza es constante, $Var(Z(\mathbf{s}+\mathbf{h})) = Var(Z(\mathbf{s}-\mathbf{h}))$, y dada la simetría de la covarianza obtenemos:

$$2\gamma(\mathbf{h}) = Var(Z(\mathbf{s}+\mathbf{h})) + Var(Z(\mathbf{s})) - 2C(\mathbf{h}) = Var(Z(\mathbf{s}-\mathbf{h})) + Var(Z(\mathbf{s})) - 2C(-\mathbf{h}) = 2\gamma(-\mathbf{h}),$$

y esto ocurre sí y solo sí $\gamma(\mathbf{h}) = \gamma(-\mathbf{h})$.

Por otro lado, $\gamma(0) = 0$, ya que $2\gamma(0) = Var(Z(s) - Z(s)) = Var(0)$, luego $\gamma(0) = 0$.

Finalmente, dado que el semivariograma es una varianza sabemos que $\gamma(\mathbf{h}) \geq 0$.

b) Sea un conjunto de puntos $\mathbf{s}_1,\dots,\mathbf{s}_n$ y un conjunto de constantes $a_1,\dots,a_n$ tales que $\sum_i a_i = 0$, si $\gamma(\mathbf{h})$ es válido, entonces:

$$\sum_i \sum_j a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j) = \frac{1}{2} E \sum_i \sum_j a_i a_j (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2$$

$$= -E \sum_i \sum_j a_i a_j Z(\mathbf{s}_i) Z(\mathbf{s}_j) = -E \left[\sum_i a_i Z(\mathbf{s}_i) \right]^2 \leq 0.$$

□

Vamos a estudiar ahora la relación entre el variograma y la función de covarianza. En primer lugar veremos cómo se puede calcular γ a partir de C.

Teorema 2.2. *Sea $Z(\boldsymbol{s})$ un proceso estacionario con función de covarianza C y $\boldsymbol{h} \in \mathbb{R}^d$, entonces se cumple:*

$$\gamma(\boldsymbol{h}) = C(0) - C(\boldsymbol{h}). \tag{2.4}$$

Demostración. Por la definición del variograma obtenemos:

$$2\gamma(\mathbf{h}) = Var(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})) = C(0) + C(0) - 2C(\mathbf{h}) = 2\left[C(0) - C(\mathbf{h})\right].$$

Luego $\gamma(\mathbf{h}) = C(0) - C(\mathbf{h})$.

□

En general, la covarianza no siempre puede expresarse en función del variograma. Un caso particular en el cual se verifica, es cuando asumimos que el proceso espacial es *ergódico*, es decir, $C(\mathbf{h}) \to 0$ cuando $\|\mathbf{h}\| \to \infty$, donde $\|\mathbf{h}\|$ denota la longitud del vector $\mathbf{h}$. Esta condición indica que la covarianza entre los valores en dos puntos tiende a cero a medida que los puntos se separan más en el espacio. Veamos entonces cómo podemos expresar la covarianza en función del variograma. En primer lugar, tenemos $\gamma(\mathbf{u}) = C(0) - C(\mathbf{u})$ y tomamos el límite. Debido a que el proceso espacial es ergódico obtenemos:

$$\lim_{\|u\|\to\infty} \gamma(\mathbf{u}) = C(0) - \lim_{\|u\|\to\infty} C(\mathbf{u}) = C(0). \tag{2.5}$$

Por otro lado, $C(\mathbf{h}) = C(0) - \gamma(\mathbf{h})$ y sustituyendo por la expresión anterior (2.5) obtenemos:

$$C(\mathbf{h}) = \lim_{\|u\|\to\infty} \gamma(\mathbf{u}) - \gamma(\mathbf{h}). \tag{2.6}$$

Logramos así una forma de determinar la función de covarianza C a partir del semivariograma γ.

Veamos por último que la estacionariedad débil implica estacionariedad intrínseca. Para ello hay que probar lo siguiente:

i) $E(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})) = 0$.

   En efecto, si el proceso es débilmente estacionario, $E(Z(\mathbf{s})) = \mu$, y por lo tanto $E(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})) = \mu - \mu = 0$.

ii) $Var(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s}))$ es una función que depende únicamente de $\mathbf{h}$, no del punto $\mathbf{s}$.

   Sabemos que,

$$Var(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})) = Var(Z(\mathbf{s}+\mathbf{h})) + Var(Z(\mathbf{s})) - 2Cov(Z(\mathbf{s}+\mathbf{h}), Z(\mathbf{s})).$$

   Si el proceso es débilmente estacionario $Var(Z(\mathbf{s}+\mathbf{h})) = Var(Z(\mathbf{s})) = \sigma^2$, no depende de $\mathbf{s}$, ni de $\mathbf{h}$ y $Cov(Z(\mathbf{s}), Z(\mathbf{s}+\mathbf{h})) = C(\mathbf{h})$ es una función que solo depende de $\mathbf{h}$.

El recíproco, sin embargo, no es cierto.

Procesos isotrópicos

A continucación vamos a introducir un concepto de gran importancia en los procesos espaciales que es la isotropía. Dicha propiedad es valiosa debido a su capacidad para simplificar el análisis y mejorar las predicciones.

Definición 5. Un proceso espacial $Z(\mathbf{s})$ se dice *isotrópico* si la correlación entre los datos no depende de la dirección en la que esta se calcule, sino únicamente de la distancia. En caso contrario se hablará de *anisotropía* o de proceso espacial *anisotrópico*. En otras palabras, $Z(\mathbf{s})$ es isotrópico si su función semivariograma $\gamma(\mathbf{h})$, depende del vector sólo a través de su longitud $\|\mathbf{h}\|$.

Denotaremos a partir de ahora $\|\mathbf{h}\|$ por h para simplificar la notación.

2.2. Tipos de variograma para procesos isotrópicos

Más adelante veremos cómo la formulación de modelos paramétricos para el variograma resulta una herramienta de gran utilidad durante el proceso de estimación. En general dichos modelos pueden dividirse en no acotados (lineal) y acotados (esférico, exponencial, gaussiano). Los del segundo grupo garantizan que la covarianza de los incrementos es finita, por lo cual son ampliamente usados cuando hay evidencia de que presentan buen ajuste. Todos estos modelos tienen tres parámetros comunes que son descritos a continuación y su interpretación gráfica puede observarse en la Figura 2.1.

- *Pepita (nugget)*

  Este parámetro, denotado generalmente por τ^2, representa una discontinuidad puntual del semivariograma en el origen, la cual puede ser causada por errores de medición en la variable o debido a que existe una variabilidad aleatoria en escalas muy pequeñas.

- *Rango (range)*

  El rango representa la distancia a la cual el variograma alcanza su valor máximo. En términos prácticos corresponde a la distancia a partir de la cual dos observaciones son independientes. Cuanto más pequeño sea el rango, más cerca se está del modelo de independencia espacial. Sin embargo, el rango no siempre aparece de manera explícita en la fórmula del semivariograma.

- *Meseta (sill)*

  La meseta representa el valor máximo del variograma a medida que la distancia entre los puntos aumenta y se aleja del punto de origen. Se calcula a partir de $\tau^2 + \sigma^2$, donde τ^2 es la pepita y σ^2 la meseta parcial. También puede definirse como el límite del semivariograma cuando la distancia h tiende a infinito. Dicho límite puede ser o no finito.

A continuación, consideramos algunos de los modelos de variogramas más importantes.

- **Lineal**

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2 h, & \text{si } h > 0, \quad \tau^2 > 0, \quad \sigma^2 > 0 \\ 0, & \text{en otro caso.} \end{cases} \tag{2.7}$$

  Notar que $\gamma(h) \to \infty$ cuando $h \to \infty$, luego este variograma no se corresponde con un proceso estacionario débil, pero sí que es intrínsecamente estacionario. En este modelo la pepita es τ^2 pero la meseta y el rango son infinitos.

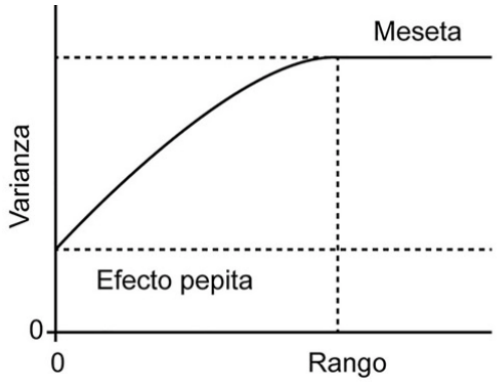

Figura 2.1: Parámetros del semivariograma. Fuente [6].

- **Esférico**

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2, & \text{si } h \geq 1/\phi, \\ \tau^2 + \sigma^2 \left\{ \frac{3\phi h}{2} - \frac{1}{2}(\phi h)^3 \right\}, & \text{si } 0 < h < 1/\phi, \\ 0, & \text{en otro caso.} \end{cases} \tag{2.8}$$

Este variograma ofrece una ilustración muy clara de los parámetros definidos anteriormente. Mientras que $\gamma(0) = 0$ por definición, $\gamma(0^+) \equiv \lim_{d \to 0^+} \gamma(h) = \tau^2$ es la pepita. Por otro lado, $\lim_{h \to \infty} \gamma(h) = \tau^2 + \sigma^2$ es la meseta. Finalmente, el valor $h = 1/\phi$ en el cual $\gamma(h)$ alcanza por primera vez su valor máximo es el rango, donde ϕ se denomina parámetro de decaimiento.

- **Exponencial**

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi h)), & \text{si } h > 0, \\ 0, & \text{en otro caso.} \end{cases} \tag{2.9}$$

El variograma exponencial tiene la ventaja sobre el esférico de que es más simple en forma funcional sin dejar de ser un variograma válido en todas las dimensiones. Sin embargo, notar que la meseta solo se alcanza asintóticamente; estrictamente hablando, el rango $R \equiv 1/\phi$ es infinito. En estos casos se suele emplear la noción de *rango efectivo*, definido como la distancia a partir de la cual la correlación puede considerarse despreciable. Generalmente se considera una correlación menor a 0.05, la cual para el caso exponencial se alcanza para una distancia h tal que $\exp(-\phi h) = 0{,}05$, dando un rango efectivo de $h = -\log(0{,}05)/\phi \approx 3/\phi$.

- **Gaussiano**

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi^2 h^2)) & \text{si } h > 0 \\ 0 & \text{en otro caso.} \end{cases} \tag{2.10}$$

Al igual que en el modelo exponencial, la dependencia espacial se anula solo en una distancia que tiende a infinito. El principal distintivo de este modelo es su forma parabólica cerca del origen.

- **Matérn**

El modelo Matérn es conocido por su flexibilidad para describir diferentes patrones de variabilidad espacial en datos georreferenciados. La forma general de la función de variograma se define como

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2 \left[1 - \frac{(2\sqrt{\nu} h \phi)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(2\sqrt{\nu} h \phi) \right] & \text{si } h > 0 \\ 0 & \text{si } h = 0 \end{cases} \tag{2.11}$$

donde $\Gamma(\cdot)$ es la función Gamma usual y K_ν es la función Bessel modificada de orden ν. Además, este modelo incorpora un parámetro adicional, ν, que proporciona mayor flexibilidad. El valor de

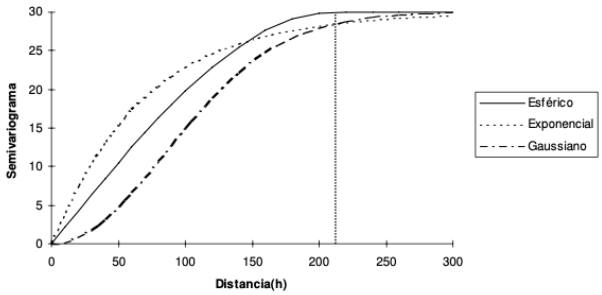

Figura 2.2: Comparación de los modelos exponencial, esférico y gaussiano. La línea vertical representa el rango en el caso del modelo esférico y el rango efectivo en el de los modelos exponencial y gaussiano. Fuente: [1, cap. 2].

Modelo	Covarianza, $C(h)$	Variograma
Lineal	$C(h)$ no existe	$\gamma(h) = \begin{cases} \tau^2 + \sigma^2 h, & \text{si } h > 0, \quad \tau^2 > 0, \quad \sigma^2 > 0 \\ 0, & \text{en otro caso.} \end{cases}$
Esferico	$C(h) = \begin{cases} 0 & \text{si } h \geq 1/\phi \\ \sigma^2\left[1 - \frac{3}{2}\phi h + \frac{1}{2}(\phi h)^3\right] & \text{si } 0 < h \leq 1/\phi \\ \tau^2 + \sigma^2 & \text{si } h = 0 \end{cases}$	$\gamma(h) = \begin{cases} \tau^2 + \sigma^2, & \text{si } h \geq 1/\phi, \\ \tau^2 + \sigma^2\left\{\frac{3\phi h}{2} - \frac{1}{2}(\phi h)^3\right\}, & \text{si } 0 < h < 1/\phi, \\ 0, & \text{en otro caso.} \end{cases}$
Exponencial	$C(h) = \begin{cases} \sigma^2 \exp(-\phi h) & \text{si } h > 0 \\ \tau^2 + \sigma^2 & \text{si } h = 0 \end{cases}$	$\gamma(h) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi h)), & \text{si } h > 0, \\ 0, & \text{en otro caso.} \end{cases}$
Gaussiano	$C(h) = \begin{cases} \sigma^2 \exp(-\phi^2 h^2) & \text{si } h > 0 \\ \tau^2 + \sigma^2 & \text{si } h = 0 \end{cases}$	$\gamma(h) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi^2 h^2)) & \text{si } h > 0 \\ 0 & \text{en otro caso.} \end{cases}$
Matern	$C(h) = \begin{cases} \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\phi h)^\nu K_\nu(\phi h) & \text{si } h > 0 \\ \tau^2 + \sigma^2 & \text{si } h = 0 \end{cases}$	$\gamma(h) = \begin{cases} \tau^2 + \sigma^2\left[1 - \frac{(2\sqrt{\nu}h\phi)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(2\sqrt{\nu}h\phi)\right] & \text{si } h > 0 \\ 0 & \text{si } h = 0 \end{cases}$

Cuadro 2.1: Funciones de covarianzas asociadas a cada modelo de variograma.

ν determina cómo se comporta la función de variograma Matérn a medida que aumenta la distancia espacial. Concretamente, si consideramos el caso particular $\nu = 0{,}5$ obtenemos

$$K_{0,5}(x) = \sqrt{\frac{\pi}{2x}}\exp(-x),$$

y por tanto $\gamma(h) = \tau^2 + \sigma^2(1 - \exp(-\phi h))$, que es exactamente la función variograma del modelo exponencial. Por otro lado, cuando ν tiende a infinito, la función variograma se aproxima a un modelo gaussiano.

Las distintas formas de variogramas y las covarianzas a las que corresponden dichos modelos, se resumen en la Tabla 2.1.

2.3. Estimación del variograma

Sea $Z(\mathbf{s})$ un proceso espacial y sean $Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)$ sus n observaciones. Nuestro objetivo es estimar la correlación espacial entre las observaciones del proceso $Z(\mathbf{s})$. Generalmente, en el caso de datos georreferenciados, dicha correlación es modelada por la función semivariograma $2\gamma(\mathbf{h})$ definida por (2.2).

El objetivo es estimar la función variograma a partir de la muestra observada. Además, en toda la sección se va asumir que el proceso espacial $Z(\mathbf{s})$, del que proviene la muestra, es intrínsicamente estacionario e isotrópico. De nuevo emplearemos la notación h para referirnos a la longitud $\|\mathbf{h}\|$.

2.3.1. Estimación no paramétrica

Bajo las suposiciones anteriores y empleando el método de los momentos, se obtiene el denominado *estimador del método de momentos* (o clásico) del semivariograma

$$\hat{\gamma}(h) = \frac{1}{2\,|N(h)|} \sum_{|N(h)|} [Z(\mathbf{s}_i) - Z(\mathbf{s}_j)]^2, \tag{2.12}$$

donde $N(h) = \left\{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i - \mathbf{s}_j = h\right\}$ es el conjunto de todos los pares de localizaciones separadas por la distancia h y $|N(h)|$ es el cardinal de dicho conjunto.

De forma análoga, en el caso de procesos con estacionariedad de segundo orden, se obtiene el estimador clásico para la función de covarianza

$$\hat{C}(h) = \frac{1}{|N(h)|} \sum_{|N(h)|} [Z(\mathbf{s}_i) - \bar{Z}_n]\,[Z(\mathbf{s}_j) - \bar{Z}_n], \tag{2.13}$$

siendo $\hat{\mu} = \bar{Z}_n$ la media muestral. Nótese que, la necesidad de estimar la media μ produce que el estimador sea sesgado.

Estimador robusto

El estimador del método de momentos del semivariograma es un promedio de diferencias al cuadrado y, por lo tanto, puede verse muy influenciado por un pequeño número de valores atípicos. Por esta razón, para disminuir la importancia de cualquier diferencia grande al cuadrado, a veces es preferible considerar un estimador robusto. Por ejemplo, Cressie y Hawkins (1980) propusieron el siguiente estimador robusto

$$\bar{\gamma}(h) = \frac{1}{2\,|N(h)|} \frac{\sum_{|N(h)|} \left\{[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)]^{1/2}\right\}^4}{0{,}457 + 0{,}494/\,|N(h)|}, \tag{2.14}$$

donde se añade un factor de corrección de forma que el estimador sea insesgado cuando el proceso espacial es normal.

La robustez en este contexto se refiere a la capacidad del estimador de resistir la influencia de valores atípicos o datos erróneos, lo que puede distorsionar los resultados y afectar negativamente a la interpretación del análisis. Esto se puede ver en el caso del estimador (2.14) con el uso de diferencias de raíces cuadradas, en lugar de diferencias cuadradas, como en el estimador (2.12).

Estimador Kernel

El objetivo principal del *estimador kernel* es obtener un estimador que permita estimar la función variograma en cualquier valor h y de una forma suave. Dicho estimador emplea una función kernel que suaviza los pares de diferencias para posteriormente, calcular el semivariograma a partir de esas diferencias suavizadas.

En efecto, definimos una función de densidad bidimensional simétrica y no negativa, $w(\mathbf{u})$, centrada en 0, cuya integral sobre todo su dominio debe sumar uno, es decir, $\int w(\mathbf{u})d\mathbf{u} = 1$. Por tanto, la estimación kernel del semivariograma quedaría

$$\hat{\gamma}_\delta(h) = \sum_{(i,j)} \frac{\left\{[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)]^2\right\} w_\delta(h - h_{ij})}{\sum_{i,j} w_\delta(h - h_{ij})}, \tag{2.15}$$

donde $h_{ij} = \mathbf{s}_i - \mathbf{s}_j$ es la distancia espacial observada entre i y j.

Además, $w_\delta(\mathbf{u}) = \frac{1}{\delta} w(\frac{\mathbf{u}}{\delta})$ donde δ es el parámetro denominado *parámetro de suavizado* y determina el peso que se asigna a cada observación en la estimación en cada h. Un valor de δ más pequeño resultará en una estimación con más variabilidad, mientras que un valor más grande dará como resultado una estimación más suave. En resumen, la elección del parámetro es crítica en la suavización de datos utilizando funciones kernel. En la práctica, encontrar el δ óptimo a menudo implica experimentar con diferentes valores y evaluar cómo afectan a la suavización y la interpretación de los resultados.

2.3.2. Estimación paramétrica

El estimador del método de momentos del semivariograma es intuitivo e imparcial. Aún así, este estimador presenta dos insuficiencias principales. La primera surge del hecho de que las varianzas son positivas. Específicamente, esto requiere que para cualquier constante b_i, $i = 1, \ldots, k$ y cualquier conjunto de ubicaciones espaciales, $\mathbf{s}_i$, $i = 1, \ldots, k$, se cumple que

$$\text{Var}\left[\sum_{i=1}^{k} b_i Z(\mathbf{s}_i)\right] = \sum_{i=1}^{k}\sum_{j=1}^{k} b_i b_j \,\text{Cov}\left[Z(\mathbf{s}_i), Z(\mathbf{s}_j)\right] \geq 0.$$

De manera análoga, como hemos comentado la sección 2.1, se requiere que $\sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0$, para todo a_i tal que $\sum_{i=1}^{k} a_i = 0$. El estimador del método de los momentos dado en (2.12) no garantiza dichas propiedades.

La segunda razón por la que el estimador (2.12) no es adecuado, es debido a que proporciona estimaciones de correlación sólo en las distancias espaciales h observadas. Sin embargo, por lo general, en la práctica se requiere un estimador de la correlación espacial en cualquier h, incluidas las distancias espaciales que pueden no haberse observado.

Una solución común a estas dos dificultades es ajustar modelos paramétricos válidos, como los desarrollados en la Sección 2.2, para los cuales la función semivariograma es definida no positiva, lo que luego permite el cálculo del semivariograma con cualquier distancia espacial. A continuación presentamos un proceso general para la estimación paramétrica del semivariograma.

Sea $\mathbf{\Lambda} = \{h_1, \ldots, h_m\}$ un conjunto de distancias. Consideramos el vector $\hat{\boldsymbol{\gamma}} = \{\hat{\gamma}(h_1), \ldots, \hat{\gamma}(h_m)\}$ de longitud m, que denota las estimaciones del semivariograma en las m distancias espaciales, calculadas mediante cualquiera de los estimadores no paramétricos mencionados anteriormente. Por ejemplo, $\hat{\boldsymbol{\gamma}}$ puede ser el estimador clásico, el estimador robusto o el estimador kernel. Definimos $\hat{\gamma}(h_i)$ como la i-ésima componente del vector, $i = 1, \ldots, m$.

Buscamos entonces una variograma paramétrico que sea el más cercano en algún sentido a este estimador no paramétrico. Por lo general, esto significa minimizar una distancia cuadrada entre los dos. Concretamente, supongamos que la verdadera función semivariograma se encuentra dentro de una familia paramétrica particular, denotada como $[\gamma_Z(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbf{\Theta}]$. Definimos entonces, para cada $\boldsymbol{\theta}$, $\boldsymbol{\gamma}_Z(\boldsymbol{\theta})$ como el vector de longitud m de los valores del semivariograma paramétrico en las distancias espaciales $\mathbf{\Lambda}$. Además, $\gamma_Z(h_i, \theta)$ denotará la i-ésima componente del vector con $i = 1, \ldots, m$.

A continuación se realiza la estimación de los parámetros, cuyo objetivo es encontrar los valores de los parámetros del modelo paramétrico que se ajusten lo mejor posible a las estimaciones $\hat{\boldsymbol{\gamma}}$ en las distancias $\mathbf{\Lambda}$. Para ello existen diversos criterios de ajuste. Entre ellos hay que destacar los basados en mínimos cuadrados y en máxima verosimilitud, descritos a continuación.

Estimación por mínimos cuadrados

- *Mínimos cuadrados ordinarios (OLS)*

  En el método de mínimos cuadrados ordinarios, se busca elegir $\boldsymbol{\theta}$ de tal manera que se minimice la expresión

$$R(\boldsymbol{\theta}) = \sum_{i=1}^{m} \left[\hat{\gamma}(h_i) - \gamma_Z(h_i, \boldsymbol{\theta})\right]^2, \tag{2.16}$$

donde $R(\boldsymbol{\theta})$ es la función del error que queremos minimizar.

Aunque esta minimización parece natural y computacionalmente sencilla, no es el estimador más eficaz. Esto se debe a que el método de mínimos cuadrados ordinarios no tiene en cuenta el número de diferencias al cuadrado promediadas en cada distancia, $N(h_i)$, y asigna el mismo peso a cada sumando. Por lo tanto, el método no considera la variabilidad asociada a cada diferencia al cuadrado, ni la posible correlación entre elementos de $\hat{\boldsymbol{\gamma}}$.

- *Mínimos cuadrados generalizados (GLS)*

  Este criterio tiene en cuenta los aspectos anteriores y no asume necesariamente que los errores sean independientes entre sí y tengan varianza constante. Si definimos la matriz $m \times m$ de varianza-covarianza como $\mathbf{V}(\boldsymbol{\theta}) := Var[\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_Z(\boldsymbol{\theta})]$, entonces el GLM busca minimizar la función

  $$R(\boldsymbol{\theta}) = (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_Z(\boldsymbol{\theta}))^\top \mathbf{V}(\boldsymbol{\theta})^{-1} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_Z(\boldsymbol{\theta})). \tag{2.17}$$

  A pesar de ser un estimador mejor, existen algunas dificultades prácticas. La primera es que la matriz $\mathbf{V}(\boldsymbol{\theta})$ es difícil de calcular en la práctica. Se pueden encontrar algunas aproximaciones de muestras grandes para procesos gaussianos, pero en general no es una tarea sencilla. En segundo lugar, minimizar este criterio puede resultar complicado, ya que $R(\boldsymbol{\theta})$ es una función no lineal.

- *Mínimos cuadrados ponderados (WLS)*:

  Debido a las dificultades comentadas con el GLS, el método de mínimos cuadrados ponderados busca dar mayor importancia a las estimaciones más fiables. Para ello, el WLS asigna pesos diferentes a cada término, para así tener en cuenta la variabilidad de los estimadores. Este criterio busca encontrar el $\boldsymbol{\theta}$ que minimice

  $$R(\boldsymbol{\theta}) = (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_Z(\boldsymbol{\theta}))^\top \mathbf{W}(\boldsymbol{\theta}) (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_Z(\boldsymbol{\theta})) = \sum_{i=1}^{m} w_i^2 [\hat{\gamma}(h_i) - \gamma_Z(h_i, \theta)]^2, \tag{2.18}$$

  donde $\mathbf{W}(\boldsymbol{\theta})$ denota una matriz diagonal definida por los pesos w_i^2.

  Para los pesos, nótese que las observaciones gaussianas, $Z(\mathbf{s}+h) - Z(h)$ tienen una distribución $N(0, 2\gamma(h))$, por la definición del variograma. Notar que $[Z(\mathbf{s}+h) - Z(h)] / \sqrt{2\gamma(h)}$ tiene una distribución $N(0,1)$. Por lo tanto, $[Z(\mathbf{s}+h) - Z(h)]^2 / 2$ tiene una distribución $\gamma(h)\chi_1^2$, donde χ_1^2 denota la distribución chi-cuadrado con un grado de libertad. Recordemos que el estimador del método de momentos tiene la expresión (2.12). Si, para cada i, asumimos que las $|N(h_i)|$ diferencias cuadradas son independientes, entonces $\text{Var}[\widehat{\gamma}(h_i)] = 2\gamma^2(h_i) / |N(h_i)|$. Esto sugiere dar los valores $w_i^2 = |N(h_i)| / 2\gamma^2(h_i)$. En otras palabras, buscamos minimizar la expresión

  $$\sum_{i=1}^{m} N(h_i) \left[\frac{\hat{\gamma}(h_i)}{\gamma_Z(h_i, \theta)} - 1 \right]^2. \tag{2.19}$$

  Se podría decir que el método WLS se encuentra entre el estimador OLS, que resultaba fácil de computar, y el estimador GLS, que aunque eficaz, era dificil de computar. En general, el método de mínimos cuadrados ponderados suele ser el criterio de ajuste por defecto.

Una vez que se han ajustado los parámetros, la estimación paramétrica del variograma puede utilizarse para predecir la dependencia espacial en ubicaciones no muestreadas y proporcionar información sobre cómo las observaciones se correlacionan a diferentes distancias.

Estimador máximo verosímil

La estimación por máxima verosimilitud (*maximum likelihood*, ML) es un método muy conocido en inferencia estadística paramétrica, aunque su uso en estadística espacial ha sido relativamente reciente.

Si suponemos que la distribución de los datos es normal, $\mathbf{Z} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$, donde $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ es la matriz de covarianza de una distribución normal multivariante. Se puede deducir fácilmente la expresión de la función de verosimilitud y obtener las estimaciones de los parámetros buscando los valores que la maximizan.

En este caso, la expresión del logaritmo negativo de la función de verosimilitud (*negative log likelihood*, NLL) es:

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = (n/2)\log(2\pi) + (1/2)\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + (1/2)(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}) \tag{2.20}$$

Los detalles sobre el cálculo de $\hat{\boldsymbol{\beta}}$ y $\hat{\boldsymbol{\theta}}$ se pueden consultar en [7].

Una ventaja del método de máxima verosimilitud es que permite estimar de forma conjunta $\boldsymbol{\beta}$ y $\boldsymbol{\theta}$ directamente de los datos y no es necesario calcular estimaciones no paramétricas del variograma. Sin embargo, uno de los principales problemas de la estimación ML es que los estimadores pueden tener un sesgo considerable, especialmente cuando la media no es constante. Este problema se puede resolver, por lo menos en parte, utilizando una variante de este método.

El método de máxima verosimilitud restringida (*restricted maximum likelihood*, REML) se basa en la idea de filtrar los datos de forma que la distribución conjunta no dependa de $\boldsymbol{\beta}$. En este caso, no vamos a trabajar con la función de verosimilitud original de los datos, sino con la función de verosimilitud de las diferencias de los datos, es decir, $\mathbf{W} = (Z(1) - Z(2), Z(2) - Z(3), \ldots, Z(n-1) - Z(n))^\top$. Por lo tanto, se busca minimizar:

$$L_W(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{(n-1)}{2}\log(2\pi) + (1/2)\log\left|A^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})A\right| + \frac{1}{2}\left(\mathbf{W} - A^\top \mathbf{X}\boldsymbol{\beta}\right)^\top \left(A^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})A\right)^{-1}\left(\mathbf{W} - A^\top \mathbf{X}\boldsymbol{\beta}\right), \tag{2.21}$$

donde $A = (a_{ij})$ es una $(n-1) \times n$ matriz cuyos elementos son:

$$a_{ij} = \begin{cases} 1, \text{ para } i = j, j = 1, \ldots, n-1, \\ -1, \text{ para } i = j+1, j = 1, \ldots, n-1, \\ 0, \text{ en otro caso.} \end{cases} \tag{2.22}$$

Asumiendo que el proceso $Z(\mathbf{s})$ tiene media constante μ, entonces $A^\top \mathbf{X}\boldsymbol{\beta} = 0$ ya que $\mathbf{X}\boldsymbol{\beta} = (1, \ldots, 1)^\top \mu$, y entonces L_W no depende de $\boldsymbol{\beta}$. El objetivo es que, sacrificando una observación (porque al trabajar con las diferencias tendremos $(n-1)$ observaciones en lugar de n), podamos obtener un estimador de $\boldsymbol{\theta}$ basado en L_W que tenga mejores propiedades de insesgamiento.

En general, la estimación REML mejora, a veces significativamente, los resultados obtenidos con la estimación ML.

2.3.3. Estimación con función media no constante

La suposición de media constante suele ser apropiada cuando la función media verdadera es constante o presenta muy poca variación. Sin embargo, hay muchos casos en los que la función media no es constante de una manera significativa y sistemática. En este tipo de situaciones resulta interesante considerar la función media variable.

Supongamos entonces que la media del proceso espacial $Z(\mathbf{s})$ depende de la ubicación. Es decir, el modelo que siguen las observaciones será

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s}), \quad s \in D,$$

donde $\delta(\cdot)$ es un proceso estocástico estacionariamente intrínseco de media cero.

Una forma habitual de modelizar medias espacialmente heterogéneas es $\mu(\mathbf{s}) = \mathbf{X}\boldsymbol{\beta}$, donde $\mathbf{X}$ es un vector de p covariables que dependen del espacio y $\boldsymbol{\beta}$ es el vector de coeficientes asociados a dichas covariables,

$$\mathbf{Z}(\mathbf{s}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta}(\mathbf{s}), \tag{2.23}$$

siendo $\mathbf{Z}(\mathbf{s}) = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))^\top$.

Además, nótese que

$$E\left[2\widehat{\gamma}(h)\right] = E(Z(\mathbf{s}+h) - Z(\mathbf{s}))^2 = 2\gamma(h) + (\mu(\mathbf{s}+h) - \mu(\mathbf{s}))^2,$$

En efecto, se ve que para funciones medias no constantes, la esperanza del estimador es el variograma más el cuadrado de la diferencia de la función de medias en los puntos $\mathbf{s}$ y $\mathbf{s}+h$. En otras palabras, es sesgado. De conocerse la función media, podríamos estimar el variograma a partir de $\delta(\mathbf{s}) := Z(\mathbf{s}) - \mu(\mathbf{s})$. En la práctica, sin embargo, la función media generalmente se desconoce y debe estimarse. A continuación presentamos algunos procedimientos para este tipo de situaciones.

- *Usar el modelo lineal*

  En el modelo lineal general, estimamos en primer lugar $\boldsymbol{\beta}$, usando el método de mínimos cuadrados ordinarios (OLS), obteniendo $\widehat{\boldsymbol{\beta}}_{OLS}$. A continuación estiamos el variograma, $2\gamma(h)$, a partir de los residuos

  $$\hat{\boldsymbol{\delta}}(\mathbf{s}) = \mathbf{Z}(\mathbf{s}) - \mathbf{X}\widehat{\boldsymbol{\beta}}_{OLS}. \tag{2.24}$$

  En caso de ser necesario, podemos volver a estimar $\boldsymbol{\beta}$, empleando el método de minimos cuadrados generalizados (GLS), con la matriz de covarianza del variograma estimado. Las predicciones finales son entonces

  $$\widehat{Z}(\mathbf{s}_0) = \mathbf{x}_0^T \hat{\boldsymbol{\beta}}_{GLS} + \hat{\boldsymbol{\delta}}(\mathbf{s}_0), \tag{2.25}$$

  donde $\mathbf{x}_0^T = (X_0(\mathbf{s}_0), \ldots, X_p(\mathbf{s}_0))^T$ y $\hat{\boldsymbol{\delta}}(\mathbf{s}_0)$ es el predictor de $\boldsymbol{\delta}(\mathbf{s}_0)$, basado en los residuos de (2.24).

- *Ignorar la media no constante*

  Para funciones medias relativamente suaves tenemos que la diferencia $\mu(\mathbf{s}+h) - \mu(\mathbf{s})$ es pequeña, para h pequeña. En esta situación, puede ser apropiado ignorar la media no constante. Por lo tanto, si dicha media $\mu(\mathbf{s})$ es constante en los subdominios del dominio de datos D, podemos estimar el varigorama en cada subdominio, utilizando un estimador empírico y, posteriormente, combinar dichas estimaciones.

- *Datos gaussianos*

  Si los datos son conjuntamente gaussianos, a menudo es preferible encontrar los estimadores de máxima verosimilitud, maximizando la verosimilitud de los datos en los parámetros de media y covarianza.

Capítulo 3

Predicción espacial

3.1. Introducción

La predicción espacial o kriging es el proceso de estimar valores de un proceso espacial en ubicaciones no muestreadas dentro de un área geográfica, utilizando información de ubicaciones previamente muestreadas. Se basa en la idea de que existen patrones espaciales y correlaciones entre las observaciones, lo que permite hacer inferencia sobre valores en lugares no muestreados. La predicción espacial se aplica en diversas disciplinas, como la geología, la epidemiología y la planificación urbana, para comprender fenómenos que varían en el espacio.

Su enfoque se fundamenta en el principio de que los valores en ubicaciones cercanas en el espacio tienen una mayor similitud que aquellos ubicados lejos entre sí. Esta propiedad de correlación espacial se utiliza para obtener predicciones más precisas y fiables.

En este capítulo abordaremos dos de las variantes principales dentro del marco del kriging: el kriging ordinario y el kriging universal. El kriging ordinario es la forma más básica y se basa en la suposición de que la media del proceso es constante. Por otro lado, el kriging universal es una extensión del ordinario que incorpora una mayor flexibilidad al permitir la inclusión de covariables o variables auxiliares en la modelización de la media del proceso espacial.

Para el desarrollo de este capítulo la bibliografía utilizada ha sido [5, cap. 2] y [7, cap. 4].

3.2. Modelo para la predicción óptima.

Sea $Z(\mathbf{s})$ un proceso espacial, supongamos que se hacen mediciones en los puntos $\mathbf{s}_i$, $i = 1,\dots,n$, de la región de estudio D, es decir, se tienen realizaciones de las variables $Z(\mathbf{s}_1),\dots,Z(\mathbf{s}_n)$ y se desea predecir $Z(\mathbf{s}_0)$, en un punto $\mathbf{s}_0$ donde no hubo medición. Comenzaremos con la estructura más simple para la función media, es decir, supondremos que es constante. Concretamente, el modelo inicial es

$$Z(\mathbf{s}) = \mu + \delta(\mathbf{s}) \quad \mathbf{s} \in D \quad \mu \in \mathbb{R}, \tag{3.1}$$

donde μ es una constante desconocida y $\delta(\mathbf{s})$ es un proceso aleatorio estacionariamente intrínseco de media nula. Consideramos entonces el punto $\mathbf{s}_0$ no muestreado donde deseamos conocer el valor del proceso $Z(\mathbf{s}_0)$ no observado. El objetivo es obtener la 'mejor' predicción de $Z(\mathbf{s}_0)$ en base a las observaciones $Z(\mathbf{s}_1),\dots,Z(\mathbf{s}_n)$. Para ello, buscamos la predicción que minimiza el error cuadrático medio, es decir, queremos encontrar el $\widehat{Z}(\mathbf{s}_0)$ que minimiza

$$E\left\{\left[Z(\mathbf{s}_0) - \widehat{Z}(\mathbf{s}_0)\right]^2\right\}. \tag{3.2}$$

Por el principio de mínimos cuadrados sabemos que el valor que minimiza el error cuadrático medio $E(Y - g(X))^2$ es la esperanza condicional $g(X) = E(Y \mid X)$. Por lo tanto obtenemos que la expresión (3.2) se minimiza cuando:

$$\widehat{Z}(\mathbf{s}_0) = E\left[Z(\mathbf{s}_0) \mid Z(\mathbf{s}_1),\dots,Z(\mathbf{s}_n)\right]. \tag{3.3}$$

El desarollo completo se puede ver en [5, pág. 24].
La *varianza de predicción* se define como

$$E\left(\{Z(\mathbf{s}_0) - E[Z(\mathbf{s}_0) \mid \mathbf{Z}_n]\}^2\right), \tag{3.4}$$

donde $\mathbf{Z}_n := [Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)]$ es el vector del valor del proceso espacial en los puntos observados $\mathbf{s}_1, \ldots, \mathbf{s}_n$.

La dificultad surge al intentar calcular el predictor $\widehat{Z}(\mathbf{s}_0) = E[Z(\mathbf{s}_0) \mid \mathbf{Z}_n]$. La esperanza condicional depende de la distribución conjunta de las observaciones $Z(\mathbf{s}_0), Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)$. Por lo tanto, el cálculo de esta media condicional requiere el conocimiento de la densidad conjunta (si la distribución de $Z(\mathbf{s})$ es continua) y el cálculo de una integral $(n+1)$ dimensional. Dado que solo tenemos n datos no hay forma de estimar esa distribución conjunta.

Planteamos entonces un objetivo más sencillo, buscar un predictor $\widehat{Z}(\mathbf{s}_0)$ que sea la mejor función lineal de los valores observados, es decir,

$$\widehat{Z}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i), \tag{3.5}$$

con $\lambda_i \in \mathbb{R}$, $i = 1, \ldots, n$ constantes. Esto ha reducido el problema de estimar una distribución $(n+1)$ dimensional a estimar n constantes. Antes de comenzar a buscar el mejor estimador lineal deberíamos preguntarnos si existe alguna restricción en los coeficientes λ_i. Puesto que hemos asumido anteriormente que la media es constante, resulta lógico exigir $E[\widehat{Z}(\mathbf{s}_0)] = \mu$. Esta condición implica que la esperanza del predictor es igual a la esperanza de la variable observada, es decir, el predictor será insesgado. Para que dicha condición se cumpla, es necesario que $\sum_{i=1}^{n} \lambda_i = 1$. Por lo tanto, nuestro objetivo principal va a ser minimizar

$$E\left\{\left[Z(\mathbf{s}_0) - \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i)\right]^2\right\} \quad \text{sujeto a} \quad \sum_{i=1}^{n} \lambda_i = 1.$$

En algunas situaciones, pueden ser necesarias restricciones adicionales sobre los pesos. Por ejemplo, en el caso que $Z(\mathbf{s})$ sean variables positivas, podría pensarse en imponer $\lambda_i > 0$ con $i = 1, \ldots, n$. No obstante, si imponemos la condición de que los pesos sean positivos, estamos eliminando la posibilidad de que las predicciones sean mayores que los valores observados, es decir, cualquier predicción del proceso espacial estará acotada por el máximo de los valores observados ya que

$$\widehat{Z}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i) \leq \sum_{i=1}^{n} \lambda_i \max_{1 \leq i \leq n} |Z(\mathbf{s}_i)| = \max_{1 \leq i \leq n} |Z(\mathbf{s}_i)|.$$

3.2.1. Cálculo del predictor lineal óptimo

Una vez definido lo que entendemos por mejor predictor lineal, vamos a calcularlo. En primer lugar desarrollamos el cuadrado:

$$\left[Z(\mathbf{s}_0) - \widehat{Z}(\mathbf{s}_0)\right]^2 = Z^2(\mathbf{s}_0) - 2\sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_0) Z(\mathbf{s}_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j Z(\mathbf{s}_i) Z(\mathbf{s}_j).$$

A continuación, completamos el cuadrado de los dos primeros términos de la ecuación sumando $\sum_{i=1}^{n} \lambda_i Z^2(\mathbf{s}_i)$. Restando esa misma cantidad en el tercer término y usando la condición $\sum_{i=1}^{n} \lambda_i = 1$ tenemos que

$$\left[Z(\mathbf{s}_0) - \widehat{Z}(\mathbf{s}_0)\right]^2 = \sum_{i=1}^{n} \lambda_i [Z(\mathbf{s}_i) - Z(\mathbf{s}_0)]^2 - \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \frac{[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)]^2}{2}.$$

Tomando la esperanza en ambas partes de la ecuación y usando la definición de la función semivariograma, $\gamma(h) = \frac{1}{2}E[Z(\mathbf{s}+h) - Z(\mathbf{s})]^2$ para todo h, obtenemos:

$$E\left[Z(\mathbf{s}_0) - \widehat{Z}(\mathbf{s}_0)\right]^2 = 2\sum_{i=1}^{n}\lambda_i\gamma(\mathbf{s}_0 - \mathbf{s}_i) - \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(\mathbf{s}_i - \mathbf{s}_j). \tag{3.6}$$

Considerando el error cuadrático medio definido en la ecuación (3.6) como una función de los pesos $F(\lambda_1,\ldots,\lambda_n)$, el cálculo del predictor de tipo kriging se reduce a minimizar $F(\lambda_1,\ldots,\lambda_n)$ sujeto a $G(\lambda_1,\ldots,\lambda_n) := \sum_{i=1}^{n}\lambda_i - 1 = 0$. Para resolverlo se emplea el método de los multiplicadores de Lagrange. Se puede consultar el desarrollo completo en [5, pág. 26]. Tenemos $(n+1)$ incógnitas, los $\lambda_i s$ y el multiplicador de Lagrange, m. Se obtiene entonces una única solución dada por:

$$\sum_{j=1}^{n}\lambda_j\gamma(\mathbf{s}_i - \mathbf{s}_j) + \frac{m}{2} = \gamma(\mathbf{s}_0 - \mathbf{s}_i), \quad i = 1,\ldots,n,$$

y

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

Esto es un sistema lineal y empleando la notación $\gamma_{ij} = \gamma(\mathbf{s}_i - \mathbf{s}_j)$ obtenemos:

$$\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n} & 1 \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2n} & 1 \\ \vdots & & & & \vdots \\ \gamma_{n1} & \gamma_{n2} & \ldots & \gamma_{nn} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ m/2 \end{bmatrix} = \begin{bmatrix} \gamma_{01} \\ \gamma_{02} \\ \vdots \\ \gamma_{0n} \\ 1 \end{bmatrix}.$$

Denotaremos $\widehat{\boldsymbol{\lambda}}_k$ a la solución del sistema $\boldsymbol{\Gamma}\boldsymbol{\lambda}_k = \boldsymbol{\gamma}$, donde $\boldsymbol{\Gamma}$ es la matriz $(n+1)\times(n+1)$ y $\boldsymbol{\gamma}$ el vector de dimensión $(n+1)$ definidos en la ecuación anterior. Es decir, que si $\boldsymbol{\Gamma}$ es invertible tenemos,

$$\widehat{\boldsymbol{\lambda}}_k = \boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}.$$

Los primeros n elementos del vector $\widehat{\boldsymbol{\lambda}}_k$ nos dan los pesos para el mejor predictor lineal de $Z(\mathbf{s}_0)$. El predictor $\widehat{Z}_k(\mathbf{s}_0) = \sum_{i=1}^{n}\widehat{\lambda}_{k_i}Z(\mathbf{s}_i)$ con estos pesos óptimos se denomina predictor *kriging* de $Z(\mathbf{s}_0)$. El caso en el que μ se considera constante pero desconocida se conoce como *kriging ordinario*. Por otro lado, el caso en el que se asume que μ es conocida se denomina *kriging simple*. Nótese que, al usar la función variograma, no es necesario conocer μ para obtener el predictor kriging. Este no es el caso si usamos la función de covarianza, que requiere explícitamente μ.

Nótese que la expresión (3.6) es la varianza de predicción para una predictor lineal usando cualquier peso λ_i, $i = 1,...,n$. Luego la varianza asociada al predictor kriging viene dada por la *varianza kriging*:

$$\sigma^2_{z_0} = E\left[Z(\mathbf{s}_0) - \widehat{Z}_k(\mathbf{s}_0)\right]^2 = 2\sum_{i=1}^{n}\widehat{\lambda}_{k_i}\gamma(\mathbf{s}_0 - \mathbf{s}_i) - \sum_{i=1}^{n}\sum_{j=1}^{n}\widehat{\lambda}_{k_i}\widehat{\lambda}_{k_j}\gamma(\mathbf{s}_i - \mathbf{s}_j) = \sum_{i=1}^{n}\widehat{\lambda}_{k_i}\gamma(\mathbf{s}_0 - \mathbf{s}_i) + \frac{m}{2} = \left\{\widehat{\boldsymbol{\lambda}}_k\right\}^{\mathrm{T}}\boldsymbol{\gamma},$$

donde la tercera igualdad se sigue del sistema lineal anterior.

El predictor kriging es un interpolador exacto ya que la expresión del predictor en los valores observados coincide exactamente con los valores observados, es decir, $\widehat{Z}_k(\mathbf{s}_i) = Z(\mathbf{s}_i)$, para $i = 1,\ldots,n$. Esto se puede comprobar observando que el error cuadrático medio es idénticamente 0 cuando todo el peso se pone en la ubicación observada, es decir, $\lambda_i = 1$ con el resto de los pesos igual a 0. Nótese que este es el menor error cuadrático medio ya que el valor es la esperanza de una variable positiva.

Existen otras formas de predecir el valor $Z(\mathbf{s}_0)$. Sin embargo, el predictor kriging es óptimo entre todos los predictores lineales. A pesar de ello, es necesaria una buena estimación del variograma para obtener predicciones de calidad.

3.2.2. Intervalos de predicción

Los intervalos de predicción para $Z(\mathbf{s}_0)$ pueden ser difíciles de encontrar en general. Sin embargo, si las observaciones tienen una distribución aproximadamente normal, entonces los intervalos de predicción son sencillos. Un resultado estándar del análisis multivariante establece que si $Z(\mathbf{s}_1),\ldots,Z(\mathbf{s}_n)$ tienen una distribución normal multivariante conjunta, $\mathbf{Z}\sim N(\mu,\Sigma)$, entonces $\sum_{i=1}^{n}\lambda_i Z(\mathbf{s}_i)$ tiene una distribución normal para cualquier peso. Esto significa que, bajo normalidad multivariante,

$$\left[\widehat{Z}(\mathbf{s}_0)\pm 1{,}96\sqrt{\sigma^2_{z_0}}\right] \tag{3.7}$$

es un intervalo de predicción del 95 % para $Z(\mathbf{s}_0)$. Además, si las observaciones espaciales tienen una distribución normal multivariante, entonces el predictor que minimiza el error cuadrático medio es lineal. En otras palabras, bajo la hipótesis de normalidad, el predictor kriging no solo es el mejor predictor lineal, sino el mejor predictor que minimiza el error cuadrático medio (lineal o no lineal).

3.3. Kriging universal

Hasta ahora hemos supuesto que la función media es constante, es decir, que $E\left[Z(\mathbf{s})\right]=\mu$. En muchos casos, un modelo más apropiado es

$$Z(\mathbf{s})=\mu(\mathbf{s})+\delta(\mathbf{s}), \tag{3.8}$$

donde estamos suponiendo que la media es heterogénea espacialmente, es decir, que depende de la localizacón $\mathbf{s}$.

Como ya se comentó en la Subsección 2.3.3, dicho modelo se puede reescribir utilizando notación matricial como:

$$\mathbf{Z}(\mathbf{s})=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\delta}(\mathbf{s}). \tag{3.9}$$

donde $\boldsymbol{\beta}=(\beta_0,\ldots,\beta_p)^T$ es un vector desconocido, $\mathbf{X}$ es un vector de p covariables con $X_{ij}=X_{j-1}(\mathbf{s}_i)$ y $\delta(\mathbf{s})$ es un proceso espacial de media 0. Supondremos también que $X_0\equiv 1$, de esta forma, el caso particular $p=0$ se corresponderá con el modelo de kriging ordinario.

Podemos observar que el modelo que siguen las observaciones $Z(\mathbf{s}_i)$, con $i=1,\ldots,n$, es equivalente a un modelo de regresión. En los modelos de regresión siempre se supone que las observaciones son independientes, y por tanto también los errores. Sin embargo, en el modelo en el cual se va a basar el kriging universal, no vamos a imponer dicha hipótesis. En este caso, los errores $\delta(\mathbf{s}_i)$ se asumen estacionariamente intrínsecos.

La variable que buscamos predecir ahora es $Z(\mathbf{s}_0)=\mathbf{x}_0^{\mathbf{T}}\boldsymbol{\beta}+\delta(\mathbf{s}_0)$ donde $\mathbf{x}_0^{\mathbf{T}}$ denota las covariables p en el punto donde se desea hacer la predicción. Se asume también que el predictor lineal tiene la forma $\widehat{Z}(\mathbf{s}_0)=\sum_{i=1}^{n}\lambda_i Z(\mathbf{s}_i)$. Además, que el predictor $\widehat{Z}(\mathbf{s}_0)$ sea insesgado requiere en este caso que

$$E\left[\widehat{Z}(\mathbf{s}_0)\right]=\sum_{i=1}^{n}\lambda_i\mu(\mathbf{s}_i)=\sum_{i=1}^{n}\lambda_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}=\mu(\mathrm{s}_0)=\mathbf{x}_0^{\mathrm{T}}\boldsymbol{\beta}, \tag{3.10}$$

para todo $\boldsymbol{\beta}$. En otras palabras, se requiere $\sum_{i=1}^{n}\lambda_i\mathbf{x}_i^{\mathrm{T}}=\mathbf{x}_0^{\mathrm{T}}$, lo cual es un sistema de p ecuaciones. Notar que en el caso de kriging ordinario ($p=1$), $\mathbf{x}_i=1$, para todo i, y por lo tanto esta restricción se reduce a $\sum_{i=1}^{n}\lambda_i=1$. Ahora, sin embargo, la suposición de que $\sum_{i=1}^{n}\lambda_i\mathbf{x}_i^{\mathrm{T}}=\mathbf{x}_0^{\mathrm{T}}$ induce p restricciones. Sean $m_1,\ldots,m_p$ los p multiplicadores de Lagrange necesarios para hacer cumplir estas p restricciones. Procediendo de manera completamente análoga al caso de kriging ordinario, obtenemos que

$$\widehat{\boldsymbol{\lambda}}_k=\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}, \tag{3.11}$$

donde $\widehat{\boldsymbol{\lambda}}_k=(\lambda_1,\ldots,\lambda_n,m_1,\ldots,m_p)^T$, $\boldsymbol{\gamma}=\left[\gamma(\mathbf{s}_0-\mathbf{s}_1),\ldots,\gamma(\mathbf{s}_0-\mathbf{s}_n),1,X_1(\mathbf{s}_0),\cdots,X_{p-1}(\mathbf{s}_0)\right]$ y $\boldsymbol{\Gamma}$ es una matriz de tamaño $(n+p)\times(n+p)$.

Por lo tanto, el predictor $\widehat{Z}_k(\mathbf{s}_0) = \sum_{i=1}^{n} \widehat{\lambda}_{k_i} Z(\mathbf{s}_i)$ con los pesos óptimos que acabamos de calcular se denominará predictor kriging universal de $Z(\mathbf{s}_0)$.

3.4. Validación cruzada

El método de validación cruzada es la técnica normalmente utilizada en modelos espaciales para diagnosticar si un modelo describe adecuadamente la variabilidad espacial de los datos. La idea básica es eliminar algunos de los datos y utilizar los restantes para predecir las observaciones eliminadas. Entonces, el error de predicción se puede calcular a partir de la diferencia entre los valores previstos menos los reales. Si repetimos este proceso sobre muchos subconjuntos eliminados, podremos evaluar la variabilidad del error de predicción. Su versión más simple, la validación cruzada dejando uno fuera (*Leave-one-out cross-validation*, LOOCV), consiste en obtener una predicción para cada observación de la muestra empleando el resto de observaciones. En el caso de datos punto-referenciados no sólo interesa analizar las predicciones, sino también las estimaciones del error cuadrático de predicción (varianza kriging).

Supongamos que $\hat{Z}_{-j}(\mathbf{s}_j)$ es un predictor de $Z(\mathbf{s}_j)$ obtenido utilizando alguno de los métodos de predicción espacial, a partir de $\{Z(\mathbf{s}_i) : i \neq j\}$ y el variograma ajustado $2\gamma(\cdot, \hat{\boldsymbol{\theta}})$ (calculado utilizando todos los datos). Su error de predicción asociado será $\sigma^2_{Z_{-j}}(\mathbf{s}_j)$, que depende, entre otras cosas, del modelo de variograma ajustado.

Hay varias formas de medir la aproximación de las predicciones a los verdaderos valores, por ejemplo:

i) Raíz del error cuadrático medio (*Root-mean-square error*, RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Z(\mathbf{s}_i) - \widehat{Z}_{-i}(\mathbf{s}_i))^2}. \tag{3.12}$$

ii) Error absoluto medio (*Mean absolute error*, MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Z(\mathbf{s}_i) - \widehat{Z}_{-i}(\mathbf{s}_i)\right|. \tag{3.13}$$

iii) Coeficiente de determinación (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Z(\mathbf{s}_i) - \widehat{Z}_{-i}(\mathbf{s}_i))^2}{\sum_{i=1}^{n}(Z(\mathbf{s}_i) - \bar{Z})^2}. \tag{3.14}$$

donde $\bar{Z}$ representa la media muestral de los valores observados.

Para valores pequeños del RMSE y MAE, mejor será el modelo. En el caso del coeficiente de determinación, para datos dentro de la muestra, el valor de R^2 variaría entre 0 y 1. En el caso de validación cruzada, el R^2 también puede tomar valores negativos, que indican un peor ajuste que utilizar solo la media, mientras que un valor cercano a 1 indica un mejor ajuste del modelo a los datos.

La validación cruzada puede utilizarse para elegir el modelo con las mejores métricas, o para explorar si hay datos atípicos entre los errores de predicción. Además, permite valorar si la capacidad de predicción del modelo satisface los requerimientos para una aplicación en particular.

Capítulo 4

Análisis espacial de datos de temperatura

El objetivo del capítulo es ilustrar mediante dos conjuntos de datos de temperatura mensual, algunos de los procedimientos vistos en las secciones anteriores. Para facilitar el análisis se emplearán algunas funciones de estas librerías con sus opciones por defecto, *gstat* [8], *lattice* [9] y *sp* [10]. Esto incluye estimar los coeficientes de regresión por medio de mínimos cuadrados, o en la estimación del variograma considerar una distancia máxima en la inclusión de pares de puntos de un tercio de la distancia máxima entre observaciones y considerar 15 agrupaciones de pares de puntos. Utilizar mínimos cuadrados generalizados y valorar otras opciones para el variograma por medio visual o validación cruzada podría mejorar el ajuste del modelo, pero un análisis más exhaustivo se aleja de los objetivos ilustrativos de la sección.

4.1. Datos y análisis exploratorio

Datos

Los datos de los que se dispone son la media mensual de la temperatura máxima diaria, medida en grados Celsius (°C), de los meses de enero y agosto del año 2021, en varias localidades españolas. Estas temperaturas están proporcionadas por la European Climate Assessment & Dataset (ECA&D) [11].

La región de estudio es la parte peninsular de España, la cual está conectada por los Pirineos y cuya costa noroeste está bordeada por el Océano Atlántico y la sudeste por el Mar Mediterráneo. Es por ello que su climatología es bastante diversa. En gran parte del territorio, predomina el clima mediterráneo que se caracteriza por tener veranos cálidos frente a inviernos suaves. Por otro lado, las regiones del norte de España, especialmente en la costa atlántica y en Galicia, tienden a tener un clima suave durante todo el año. En el interior del país, los inviernos suelen ser fríos y los veranos muy calurosos. Por último, las zonas montañosas, como los Pirineos o Sierra Nevada, presentan inviernos fríos y nevados y veranos frescos.

Debido a la gran variabilidad geográfica del país, las localizaciones escogidas presentan una gran diversidad de elevaciones, desde 3 metros en Gijón hasta 1894 metros de elevación con respecto al nivel del mar en Navacerrada. Además, 14 de las 40 observaciones se encuentran a menos de 20 km de la costa, mientras que el resto son puntos de interior.

Un dato de interés sobre nuestro análisis es que en agosto de ese año hubo una gran ola de calor que afectó a casi toda España y que batió numerosos récords de temperatura. Además, en enero tuvo lugar el temporal 'Filomena', que provocó frío intenso y nevadas significativas en varias regiones del país. Para conocer más detalles sobre estos eventos se puede consultar el artículo [12].

En la Figura 4.1 podemos observar, dentro de la España peninsular, a la izquierda los gráficos de temperaturas de enero y agosto, en el centro un mallado de elevaciones (en metros), y a la derecha un mallado de la distancia a la costa (en km). Los mallados considerados tienen una resolución de 10 km $\times$ 10 km. Podemos apreciar un posible efecto de la elevación y la distancia, así como una clara dependencia espacial ya que los puntos contiguos tienen colores similares.

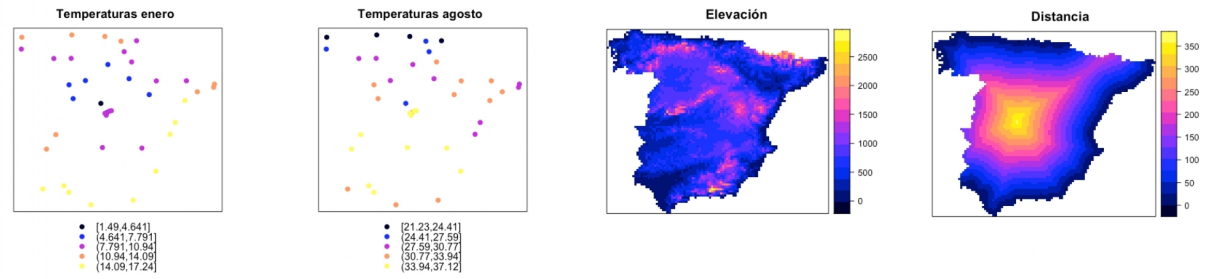

Figura 4.1: Los gráficos muestran de izquierda a derecha las temperaturas de enero y las de agosto en las localizaciones observadas, y la elevación y la distancia a la costa de un mallado de la España peninsular.

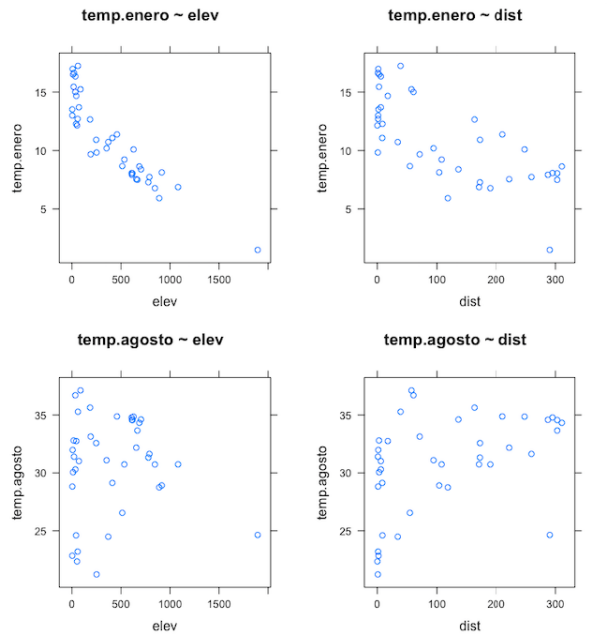

Figura 4.2: Representación gráfica de las temperaturas de enero (parte superior) y de agosto (parte inferior) frente a las covariables de elevación (en metros) y distancia a la costa (en km).

Efecto de covariables

Las dos gráficas superiores de la Figura 4.2 muestran la relación de la covariables elevación y distancia a la costa en las temperaturas de enero, mientras que las gráficas inferiores representan dicha relación en las temperaturas de agosto. En el análisis espacial también puede ser habitual utilizar la latitud y longitud. Sin embargo, en nuestro estudio, para ver el efecto de la dependencia espacial y para utilizar un modelo más parsimonioso, no hemos utilizado dichas covariables.

En el caso de enero, se observa una gran correlación entre las variables de temperatura y elevación. En efecto, esta tiene un valor de $-0{,}902$, que indica una correlación negativa fuerte. En la gráfica se observa claramente que, a medida que la variable elevación aumenta, la temperatura decrece considerablemente. Además, las temperaturas máximas se alcanzan en los puntos de elevación próxima a 0. Sin embargo, a pesar de que existe también una correlación considerable, en el caso de la distancia esta es ligeramente menor, con un valor de $-0{,}716$. En el gráfico podemos observar que las temperaturas más altas se encuentran en las ubicaciones más próximas a la costa, pero una vez superados los 100 km, dicha distancia afecta menos a la temperatura.

Por otro lado, en el caso de las temperaturas de agosto, la correlación con la variable elevación presenta un valor muy cercano al 0, lo que sugiere una relación prácticamente nula o muy débil. Finalmente,

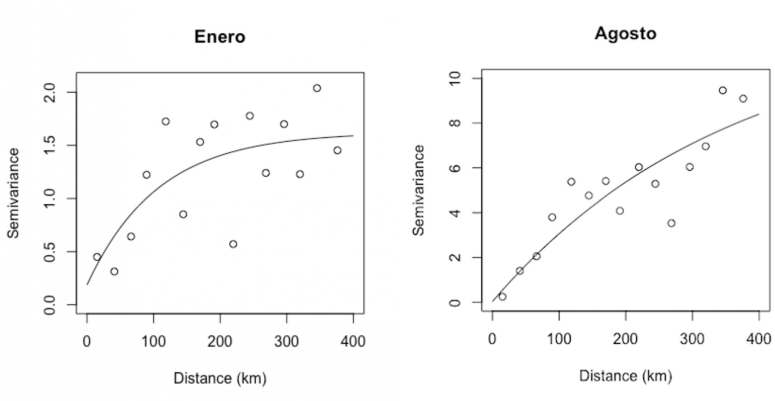

Figura 4.3: Representación del variograma no paramétrico de las temperaturas de enero (izquierda) y agosto (derecha), y de los mejores modelos de variograma paramétricos seleccionados mediante validación cruzada.

las variables temperatura y distancia tienen una correlación positiva moderada de 0,436, que indica que en general, cuando una variable aumenta, la otra tiende a aumentar también. Estos valores tan bajos de correlación en las temperaturas de agosto se deben en parte a la ola de calor que afectó a gran parte de la península ese mismo mes.

Nótese que una interacción de las covariables de elevación y distancia puede resultar interesante puesto que no es lo mismo un punto con una elevación considerable en una zona próxima a la costa, que otro con la misma elevación, pero situado en el interior.

4.2. Modelo espacial

Con el fin de comparar ambos meses, se propone el mismo modelo espacial para ambos conjuntos. Definimos $Z(\mathbf{s})$ como la temperatura media mensual con $\mathbf{s} \in D$, donde D es la España peninsular. Modelizamos entonces dichas temperaturas como

$$Z(\mathbf{s}) = \beta_0 + \beta_{elev} \cdot elev(\mathbf{s}) + \beta_{dist} \cdot dist(\mathbf{s}) + \beta_{elev:dist} \cdot elev(\mathbf{s}) \times dist(\mathbf{s}) + \delta(\mathbf{s}), \tag{4.1}$$

donde β_0 es un intercepto, $elev(\mathbf{s})$ es la elevación en la localización $\mathbf{s}$ y $dist(\mathbf{s})$ es la distancia a la costa en $\mathbf{s}$. Además, se ha incluido un efecto de interacción entre ambas, denotado por $elev(\mathbf{s}) \times dist(\mathbf{s})$. Por último, $\delta(\mathbf{s})$ es un proceso de media nula que captura la dependencia espacial. En este caso se dispone de $n = 40$ observaciones y el objetivo es poder hacer predicciones para cualquier $\mathbf{s}_0 \in D$.

Variograma no paramétrico

Como hemos comentado a lo largo del trabajo, en los modelos de datos espaciales, el variograma es utilizado para modelar la dependencia espacial de una variable en función de la distancia entre diferentes puntos de observación. En la Figura 4.3 aparecen aparecen representados mediante puntos los variogramas muestrales de las temperaturas de enero y agosto respectivamente.

Sin embargo, debido a la baja densidad de muestreo, el variograma no paramétrico puede llevar a estimaciones poco fiables y limitar la capacidad para capturar la estructura de la correlación espacial. Además, sabemos que los variogramas deben ser condicionalmente semidefinidos negativos, una propiedad que estos estimadores no paramétricos pueden no verificar. Tradicionalmente esto se remedia ajustando un modelo paramétrico válido al estimador muestral.

Covariables	Estimación	Variograma	RMSE	MAE	R^2
elev*dist	WLS	Lineal	1,343	1,070	0,861
elev*dist	OLS	Lineal	1,372	1,098	0,854
elev*dist	WLS	Exponencial	**1,158**	**0,920**	**0,896**
elev*dist	OLS	Exponencial	1,190	0,954	0,890
elev*dist	WLS	Esférico	1,318	1,052	0,866
elev*dist	OLS	Esférico	1,338	1,067	0,862
elev*dist	WLS	Gaussiano	1,256	1,010	0,878
elev*dist	OLS	Gaussiano	1,290	1,031	0,871
elev*dist	-	-	1,525	1,262	0,820
1	WLS	Gaussiano	1,911	1,447	0,718

Cuadro 4.1: Modelos de variogramas para las temperaturas medias de enero, estimados mediante OLS y WLS y validación cruzada con las métricas RMSE, MAE y R^2. Se pone en negrita el modelo con los mejores valores. La penúltima fila representa un modelo sin introducir dependencia espacial, es decir, el modelo de regresión lineal habitual, y la última el mejor modelo obtenido pero sin covariables.

Variograma paramétrico

Como se explica en la Sección 2.3.2, el procedimiento habitual para el modelado de la dependencia consiste en obtener una estimación inicial del semivariograma utilizando algún tipo de estimador empírico (en nuestro caso el variograma muestral, Figura 4.3) y posteriormente ajustar un modelo paramétrico válido de semivariograma a las estimaciones iniciales obtenidas en el primer paso. En la Sección 2.2 se presentan algunos de los modelos de variograma isotrópicos tradicionalmente utilizados en modelos con datos espaciales. En nuestro caso, emplearemos los modelos lineal, exponencial, esférico y gaussiano y se compararán después. Sin embargo, no usaremos el modelo Matérn debido a que presenta muchos problemas de convergencia. Para ajustar dichos modelos paramétricos, se emplean diversos métodos de bondad de ajuste, también mencionados en la Sección 2.3.2. Además, vamos a emplear los métodos de mínimos cuadrados ordinarios (OLS) y mínimos cuadrados ponderados (WLS). Por último, una vez ajustados dichos modelos se realiza la validación cruzada (Sección 3.4) que permite elegir el modelo ajustado que proporciona las mejores estimaciones y que posteriormente emplearemos para el kriging.

Las Tablas 4.1 y 4.2 muestran la validación cruzada realizada mediante las métricas RMSE, MAE y R^2 a los distintos modelos de variogramas estimados mediante los métodos OLS y WLS, casi siempre obteniendo mejores resultados el segundo como cabía esperar. Podemos observar (se indica en la tabla en negrita) que el mejor modelo de variograma, tanto en las temperaturas de enero como de agosto, es el exponencial. Este modelo es utilizado frecuentemente de manera estándar en la práctica en modelos para temperaturas [13], y es el que posteriormente emplearemos en el kriging universal. Estos variogramas se pueden ver representados en la Figura 4.3.

Además, para ilustrar más adelante con el kriging ordinario, se han repetido los mismos análisis sin considerar covariables. El mejor modelo obtenido para el mes de enero es el gaussiano con estimación WLS, mientras que para agosto es el exponencial, de nuevo con estimación WLS.

Interpretación del modelo

A continuación, en la Tabla 4.3, están representadas las estimaciones de los parámetros del modelo.

- *Efecto de la elevación en la temperatura*

  En el caso de las temperaturas medias mensuales del mes de enero, podemos observar a través de las estimaciones de los parámetros que, para ubicaciones muy próximas a la costa, si aumentamos 1km la elevación, la temperatura media disminuye en casi 10°C. Por otro lado, si dichos puntos se encuentran en el interior, pongamos a 300km de la costa, al aumentar de nuevo 1km la elevación, la temperatura media disminuye solo en 5°C.

  Si nos centramos en la temperaturas medias mensuales del mes de agosto, se observa que al aumentar 1km la elevación, en ubicaciones situadas en la costa, la temperatura disminuye en $3,5$°C.

Covariables	Estimación	Variograma	RMSE	MAE	R^2
elev*dist	WLS	Lineal	1,630	1,211	0,844
elev*dist	OLS	Lineal	2,069	1,462	0,749
elev*dist	WLS	Exponencial	**1,392**	**1,003**	**0,886**
elev*dist	OLS	Exponencial	1,442	1,079	0,878
elev*dist	WLS	Esférico	1,407	1,058	0,884
elev*dist	OLS	Esférico	1,471	1,113	0,873
elev*dist	WLS	Gaussiano	2,074	1,430	0,748
elev*dist	OLS	Gaussiano	2,011	1,384	0,763
elev*dist	-	-	3,497	2,677	0,331
1	WLS	Exponencial	2,469	1,886	0,643

Cuadro 4.2: Modelos de variogramas para las temperaturas medias de agosto, estimados mediante OLS y WLS y validación cruzada con las métricas RMSE, MAE y R^2. Se pone en negrita el modelo con los mejores valores. La penúltima fila representa un modelo sin introducir dependencia espacial, es decir, el modelo de regresión lineal habitual, y la última el mejor modelo obtenido pero sin covariables.

Temperatura	$\hat{\beta}_0$(°C)	$\hat{\beta}_{elev}$(°C/km)	$\hat{\beta}_{dist}$(°C/100 km)	$\hat{\beta}_{elev:dist}$([°C/[km· 100km])	$\hat{\tau}^2$	$\hat{\sigma}^2$	$1/\hat{\phi}$(km)
Enero	14,89	-9,84	-0,95	1.47	0,18	1.43	106,27
Agosto	29,01	-3,45	4,17	-1.98	0,03	12.40	356,22

Cuadro 4.3: Estimación de los parámetros del modelo.

Sin embargo, si nos alejamos 300km de la costa, dicha temperatura media disminuye aproximadamente 9°C.

Los eventos extremos opuestos sufridos en ambos meses pueden explicar un efecto tan distinto de la elevación en ambos meses.

- *Efecto de la distancia en la temperatura*

  En el caso de las temperaturas medias mensuales de enero, podemos observar a partir de las estimaciones de la Tabla 4.3 que cuando nos encontramos en puntos con elevación 0, entonces la temperatura disminuye $0,95$°C de media cada 100 km que nos alejamos de la costa. Por otro lado, si nos encontramos a 1 km de elevación respecto al nivel del mar, entonces la temperatura aumenta $0,5$°C de media por cada 100 km que nos alejamos de la costa.

  Sin embargo, si nos centramos en las temperaturas medias mensuales del mes de agosto, cuando nos ubicamos en puntos de elevación 0, entonces la temperatura aumenta $4,17$°C cada 100 km que nos alejamos de la costa. Por último, si de nuevo nos encontramos a 1 km de elevación, la temperatura media aumenta $2,18$°C cada 100 km que nos alejamos.

- *Parámetros del variograma*

  Nótese que en las temperaturas de enero, tenemos un valor de τ^2 pequeño, es decir, existe poco error puramente aleatorio y el σ^2 que mide la dependencia espacial, es unas ocho veces mayor. En el caso de agosto, el valor de τ^2 es de nuevo muy pequeño, mientras que el valor de σ^2 es inusualmente grande, lo que indica que existe una enorme dependencia espacial. El rango efectivo en enero es de $318,81$ km mientras que en agosto es de $1068,66$ km.

  La interpretación de dichos parámetros se observa claramente en las Tablas 4.1 y 4.2, en la fila donde únicamente se considera el efecto de las covariables de elevación y distancia a la costa, y no se tiene en cuenta la dependencia espacial. Para enero, se obtienen muy buenos resultados de las métricas, lo que nos indica que, en general, las covariables explican muy bien la variable respuesta, quedando únicamente un $1,43$ de variabilidad espacial. Sin embargo, si nos fijamos en el mes de agosto, se obtienen peores valores. Esto nos indica que las covariables en este caso tienen menos capacidad predictiva, pero existe una gran dependencia espacial, es decir, en localizaciones cercanas, las temperaturas serán similares, sin importar en gran medida la elevación o la distancia. Esta fuerte dependencia espacial puede explicarse por los factores que provocaron la ola de calor que afectó a todo el territorio.

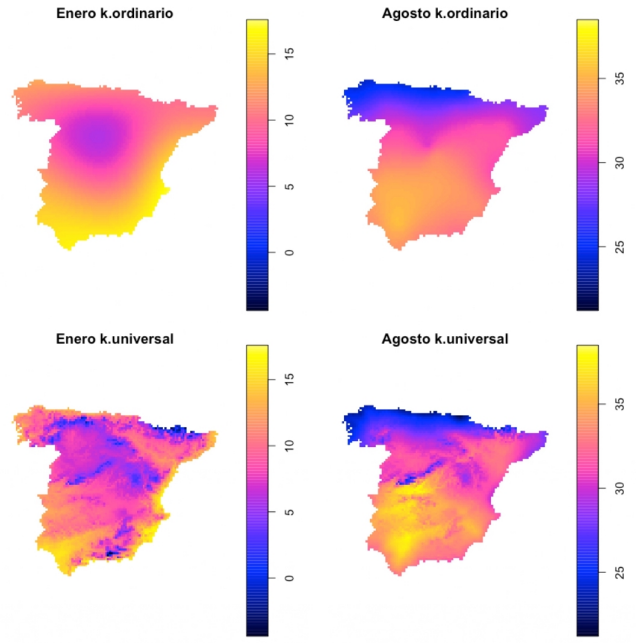

Figura 4.4: Representación del kriging ordinario (parte superior) y kriging universal (parte inferior) de las temeperaturas mensuales de los meses de enero (izquierda) y agosto (derecha).

En la Tabla 4.3 podemos observar también el valor del intercepto β_0, que es el valor medio de la variable respuesta si las covariables valen 0. En nuestro caso, si se considera elevación y distancia a la costa 0, la temperatura media en el mes de enero es de $14,89°$C mientras que en agosto es de 29°C.

4.3. Kriging

Una vez han sido escogidos los modelos con las mejores métricas tras realizar la validación cruzada, procedemos a calcular las predicciones de temperatura utilizando el kriging.

En la Figura 4.4, en la parte superior aparecen representadas las predicciones de temperatura media de los meses de enero (izquierda) y agosto (derecha), realizadas mediante el kriging ordinario. Para ello se ha empleado el mejor modelo obtenido sin considerar covariables, que recordemos es el modelo gaussiano con estimación WLS para el mes de enero y el modelo exponencial con estimación WLS para el mes de agosto.

Por otro lado, en la Figura 4.4, en la parte inferior podemos observar las mismas predicciones de temperatura media mensual, de los meses de enero (izquierda) y agostos (derecha), pero en este caso realizadas mediante kriging universal. Para ello, de nuevo se han empleado los mejores modelos, obtenidos en este caso considerando el mallado de las covariables de elevación y distancia a la costa. Recordar que el mejor modelo obtenido en ambos meses ha sido el exponencial con estimación WLS, cuyos gráficos se pueden consultar en la Figura 4.4.

Comparando ambos tipos de kriging, es fácil reconocer la importancia de introducir covariables en el modelo ya que, la dependencia espacial, salvo que se disponga de muchos datos, no captura la misma información. Ejemplo de ello puede ser el mes de enero en el cual el kriging ordinario no detecta varias zonas con temperaturas muy bajas como pueden ser los Pirineos, Sierra Nevada o la Meseta Central y que, por el contrario, el kriging universal muestra claramente. En la Figura 4.4 se observa también el efecto que tuvo la ola de calor durante el mes de agosto, afectando a toda la península, salvo Galicia, mostrando también la importancia de capturar la dependencia espacial más allá de las covariables.

Anexo A

Script utilizado en R

```
library(gstat)
library(lattice)
library(sp)
temp
temp.grid

#GRÁFICOS INTRODUCCIÓN
puntenero <- spplot(temp, "temp.enero", main = "Temperaturas enero")
puntagosto <- spplot(temp, "temp.agosto", main = "Temperaturas agosto")
elev <- spplot(temp.grid, "elev", main = "Elevación" )
dist <- spplot(temp.grid, "dist", main = "Distancia")

grid.arrange(puntos, elev, dist, ncol = 3)

#ANÁLISIS EXPLORATORIO

library(gridExtra)
plot1 <- xyplot(temp.enero ~ elev,
                data = as.data.frame(temp), main = "temp.enero ~ elev")
plot2 <- xyplot(temp.enero ~ dist,
                data = as.data.frame(temp), main = "temp.enero ~ dist")
plot3 <- xyplot(temp.agosto ~ elev,
                data = as.data.frame(temp), main = "temp.agosto ~ elev")
plot4 <- xyplot(temp.agosto ~ dist,
                data = as.data.frame(temp), main = "temp.agosto ~ dist")

grid.arrange(plot1, plot2, plot3, plot4, ncol = 2)

#Estudio de la correlación
cor1 <- cor(temp$temp.enero,temp$elev)
cor2 <- cor(temp$temp.enero,temp$dist)
cor3 <- cor(temp$temp.agosto,temp$elev)
cor4 <- cor(temp$temp.agosto,temp$dist)
```

```
#VARIOGRAMA NO PARAMÉTRICO

v1 = variogram(temp.enero ~ 1, temp)
plot11 <- plot(v1, plot.numbers = T, main="temp.enero ~ 1")

v2 = variogram(temp.enero ~ elev*dist, temp)
plot22 <- plot(v2, plot.numbers = T, main="temp.enero ~ elev*dist")

v3 = variogram(temp.agosto ~ 1, temp)
plot33 <- plot(v3, plot.numbers = T, main="temp.agosto ~ 1")

v4 = variogram(temp.agosto ~ elev*dist , temp)
plot44 <- plot(v4, plot.numbers = T, main="temp.agosto ~ elev*dist")

grid.arrange(plot11, plot22, plot33, plot44, ncol = 2)
```

```
#VARIOGRAMA PARAMÉTRICO

#Enero con covariables
v.fit11 <- fit.variogram(v2, vgm(model = "Lin", nugget = NA),fit.method = 2)
v.fit12 <-fit.variogram(v2, vgm(model = "Lin", nugget = NA),fit.method = 6)
v.fit13 <- fit.variogram(v2, vgm(model = "Exp", nugget = NA),fit.method = 2)
v.fit14 <- fit.variogram(v2, vgm(model = "Exp", nugget = NA),fit.method = 6)
v.fit15 <- fit.variogram(v2, vgm(model = "Sph", nugget = NA),fit.method = 2)
v.fit16 <- fit.variogram(v2, vgm(model = "Sph", nugget = NA),fit.method = 6)
v.fit17 <- fit.variogram(v2, vgm(model = "Gau", nugget = NA),fit.method = 2)
v.fit18 <- fit.variogram(v2, vgm(model = "Gau", nugget = NA),fit.method = 6)

#El mejor ajuste de modelo
enero <- plot(v.fit13, plot.numbers = T, cutoff=360000,
              main="Modelo Exponencial")

#Enero sin covariables
v.fit31 <- fit.variogram(v1, vgm(model = "Lin", nugget = NA),fit.method = 2)
v.fit32 <- fit.variogram(v1, vgm(model = "Lin", nugget = NA),fit.method = 6)
v.fit33 <- fit.variogram(v1, vgm(model = "Exp", nugget = NA),fit.method = 2)
v.fit34 <- fit.variogram(v1, vgm(model = "Exp", nugget = NA),fit.method = 6)
v.fit35 <- fit.variogram(v1, vgm(model = "Sph", nugget = NA),fit.method = 2)
v.fit36 <- fit.variogram(v1, vgm(model = "Sph", nugget = NA),fit.method = 6)
v.fit37 <- fit.variogram(v1, vgm(model = "Gau", nugget = NA),fit.method = 2)
v.fit38 <- fit.variogram(v1, vgm(model = "Gau", nugget = NA),fit.method = 6)

#El mejor ajuste de modelo
v.fit37 <- fit.variogram(v1, vgm(model = "Gau", nugget = NA),fit.method = 2)
```

```
#Agosto con covariables
v.fit21 <- fit.variogram(v4, vgm(model = "Lin", nugget = NA),fit.method = 2)
v.fit22 <- fit.variogram(v4, vgm(model = "Lin", nugget = NA),fit.method = 6)
v.fit23 <- fit.variogram(v4, vgm(model = "Exp", nugget = NA),fit.method = 2)
v.fit24 <- fit.variogram(v4, vgm(model = "Exp", nugget = NA),fit.method = 6)
v.fit25 <- fit.variogram(v4, vgm(model = "Sph", nugget = NA),fit.method = 2)
v.fit26 <- fit.variogram(v4, vgm(model = "Sph", nugget = NA),fit.method = 6)
v.fit27 <- fit.variogram(v4, vgm(model = "Gau", nugget = NA),fit.method = 2)
v.fit28 <- fit.variogram(v4, vgm(model = "Gau", nugget = NA),fit.method = 6)

#El mejor ajuste de modelo
agosto <- plot(v.fit23, plot.numbers = T, cutoff=1500000,
               main="Modelo Exponencial")

#Agosto sin covariables
v.fit41 <- fit.variogram(v3, vgm(model = "Lin", nugget = NA),fit.method = 2)
v.fit42 <- fit.variogram(v3, vgm(model = "Lin", nugget = NA),fit.method = 6)
v.fit43 <- fit.variogram(v3, vgm(model = "Exp", nugget = NA),fit.method = 2)
v.fit44 <- fit.variogram(v3, vgm(model = "Exp", nugget = NA),fit.method = 6)
v.fit45 <- fit.variogram(v3, vgm(model = "Sph", nugget = NA),fit.method = 2)
v.fit46 <- fit.variogram(v3, vgm(model = "Sph", nugget = NA),fit.method = 6)
v.fit47 <- fit.variogram(v3, vgm(model = "Gau", nugget = NA),fit.method = 2)
v.fit48 <- fit.variogram(v3, vgm(model = "Gau", nugget = NA),fit.method = 6)

#El mejor ajuste de modelo
v.fit43 <- fit.variogram(v3, vgm(model = "Exp", nugget = NA),fit.method = 2)

plotenero <- plot(variogram(temp.enero ~ elev * dist, temp)$dist,
     variogram(temp.enero ~ elev * dist, temp)$gamma,
     xlim = c(0, 400000), ylim = c(0, 2.1), main = "Enero",
     xaxt='n', xlab = "Distance (km)", ylab = "Semivariance")
axis(side = 1, at = 0:4 * 100000, labels = 0:4 * 100)
lines(variogramLine(fit.variogram

            (variogram(temp.enero ~ elev * dist,
            temp), vgm(model = "Exp", nugget = NA), fit.method = 2),
                    maxdist = 400000))

plotagosto <- plot(variogram(temp.agosto ~ elev * dist, temp)$dist,
     variogram(temp.agosto ~ elev * dist, temp)$gamma,
     xlim = c(0, 400000), ylim = c(0, 10), main = "Agosto",
     xaxt='n', xlab = "Distance (km)", ylab = "Semivariance")
axis(side = 1, at = 0:4 * 100000, labels = 0:4 * 100)
lines(variogramLine(fit.variogram
                    (variogram(temp.agosto ~ elev * dist, temp),
                    vgm(model = "Exp", nugget = NA), fit.method = 2),
                    maxdist = 400000))
```

```
#VALIDACIÓN CRUZADA

cv <- krige.cv(formula=temp.enero ~ elev*dist, locations = temp,
               model = v.fit11)
summary_cv <- function(cv.data,
                       tol = sqrt(.Machine$double.eps)) {
  err <- cv.data$residual
  obs <- cv.data$observed
  return(c(rmse=sqrt(mean(err^2)),mae=mean(abs(err)),
           r.squared=1-sum(err^2)/sum((obs-mean(obs))^2)))}
summary_cv(cv)

#Enero
v.fit13 <- fit.variogram(v2, vgm(model = "Exp", nugget = NA),fit.method = 2)
plot(v2, pl = T, v.fit13)
#(sin dependencia de parámetros)
train(temp.enero ~ elev * dist, method = "lm", data = data.frame(temp),
      trControl = trainControl(method = "LOOCV"))

#Agosto
v.fit23 <- fit.variogram(v4, vgm(model = "Exp", nugget = NA),fit.method = 2)
v.fit23
plot(v4, pl = T, v.fit23)
#(sin dependencia de parámteros)
train(temp.agosto ~ elev * dist, method = "lm", data = data.frame(temp),
      trControl = trainControl(method = "LOOCV"))
```

```
#RECUPERAR PARÁMETROS

#Enero
lm(temp.enero ~ elev*dist, temp)
summary(lm(temp.enero~elev*dist, temp))
#Agosto
lm(temp.agosto ~ elev*dist, temp)
summary(lm(temp.agosto~elev*dist, temp))

#Enero
nugget <- v.fit13$psill[1]
sill   <- nugget + v.fit13$psill[2]
range  <- v.fit13$range[2]

#Agosto
nugget <- v.fit23$psill[1]
sill   <- nugget + v.fit23$psill[2]
range  <- v.fit23$range[2]
```

```
#KRIGING ORDINARIO

#Enero
OKenero = krige(temp.enero~1, temp, temp.grid, v.fit37)
names(OKenero)
a <- plot(OKenero, zlim=c(-4.4,17.6), main = "Enero k.ordinario")
a
#Agosto
OKagosto = krige(temp.agosto~1, temp, temp.grid, v.fit43)
names(OKagosto)
b <- plot(OKagosto, zlim=c(21.2,38.5), main = "Agosto k.ordinario")
b
#KRIGING UNIVERSAL

#Enero
UKenero = krige(temp.enero~elev*dist, temp, temp.grid, v.fit13)
names(UKenero)
c <- plot(UKenero, zlim=c(-4.4,17.6), main= "Enero k.universal")
c

#Agosto
UKagosto = krige(temp.agosto~elev*dist, temp, temp.grid, v.fit23)
names(UKagosto)
d <- plot(UKagosto, zlim=c(21.2,38.5), main = "Agosto k.universal")
d

grid.arrange(a, b, c, d, ncol = 2)
```

Bibliografía

[1] SUDIPTO BANERJEE, BRADLEY P. CARLIN Y ALAN E. GELFAND, *Hierarchical Modeling and Analysis for Spatial Data, 2nd Edition*, CRC Press (2015).

[2] RICHARD WEBSTER Y MARGARET A. OLIVER, *Geostatistics for Environmental Scientists*, Wiley (2007).

[3] PAUL ELLIOTT, JONATHAN WAKEFIELD Y NICOLA BEST, *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, Oxford University Press (1996).

[4] NOEL A. C. CRESSIE, *Statistics for Spatial Data*, Wiley (1991).

[5] MICHAEL SHERMAN, *Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties*, John Wiley & Sons, Ltd (2011).

[6] MAXIMILIANO GARNIER-VILLARREAL, *Introduction to geostatistical analysis of data in geosciences: theory and aplication*, Universidad de Costa Rica (UCR). `https://revistas.ucr.ac.cr/index.php/geologica/article/download/51474/52401?inline=1`

[7] RUBÉN FERNÁNDEZ CASAL Y TOMÁS COTOS YÁÑEZ, *Estadística Espacial con R*, (2022).

[8] BENEDIKT GRÄLER, EDZER PEBESMA Y GERARD HEUVELINK, *Spatio-Temporal Interpolation using gstat*, The R Journal (2016).

[9] SARKAR, DEEPAYAN, *Lattice: Multivariate Data Visualization with R*, Springer, New York (2008).

[10] PEBESMA E Y BIVAND R, *"Classes and methods for spatial data in R."*, (2005).

[11] KLEIN TANK, A. M. G. AND WIJNGAARD, J. B. AND KÖNNEN, G. P. AND BÖHM, R. AND DEMARÉE, G. AND GOCHEVA, A. AND MILETA, M. AND PASHIARDIS, S. AND HEJKRLIK, L. AND KERN-HANSEN, C. AND HEINO, R. AND BESSEMOULIN, P. AND MÜLLER-WESTERMEIER, G. AND TZANAKOU, M. AND SZALAI, S. AND PÁLSDÓTTIR, T. AND FITZGERALD, D. AND RUBIN, S. AND CAPALDO, M. AND MAUGERI, M. AND LEITASS, A. AND BUKANTIS, A. AND ABERFELD, R. AND VAN ENGELEN, A. F. V. AND FORLAND, E. AND MIETUS, M. AND COELHO, F. AND MARES, C. AND RAZUVAEV, V. AND NIEPLOVA, E. AND CEGNAR, T. AND ANTONIO LÓPEZ, J. AND DAHLSTRÖM, B. AND MOBERG, A. AND KIRCHHOFER, W. AND CEYLAN, A. AND PACHALIUK, O. AND ALEXANDER, L. V. AND PETROVIC, P., *Daily Dataset of 20th-Century Surface Air Temperature and Precipitation Series for the European Climate Assessment*, International Journal of Climatology, 22, 1441-1453, (2002).

[12] DAVIDE FARANDA, STELLA BOURDIN, MIREIA GINESTA, MERIEM KROUMA, ROBIN NOYELLE, FLAVIO PONS, PASCAL YIOU Y GABRIELE MESSORI, *A climate-change attribution retrospective of some impactful weather extremes of 2021*, Weather Clim. Dynam., 3, 1311–1340, (2022). `https://doi.org/10.5194/wcd-3-1311-2022`

[13] JORGE CASTILLO-MATEO, MIGUEL LAFUENTE, JESÚS ASÍN, ANA C. CEBRIÁN, ALAN E. GELFAND AND JESÚS ABAURREA *Spatial modeling of day-within-year temperature time series:*

An examination of daily maximum temperatures in Aragón, Spain, Journal of Agricultural, Biological and Environmental Statistics, 27, 487-505, (2022)