Revista de Educación Matemática
Volumen 39, N° 3 (2024), páginas 5 – 40
Unión Matemática Argentina - Famaf (UNC)

*Artículo de Matemática*

# DATOS CORRELACIONADOS ESPACIALMENTE: ANÁLISIS ESTRUCTURAL

Rodrigo García Arancibia, Pamela Llop, Mariel Lovatto

Resumen. En este artículo introducimos y estudiamos el marco teórico subyacente para datos espacialmente correlacionados. Más precisamente, definimos el proceso que genera los datos, estudiamos sus diferentes estructuras de covarianza y caracterizamos las diferentes clases de estacionariedad consideradas habitualmente para este tipo de datos. Además, estudiamos en profundidad el semivariograma teórico y empírico, la herramienta tal vez más utilizada para medir correlación espacial. Consideramos que este trabajo puede ser un material útil para el estudio y la enseñanza de los datos espaciales y sus principales características, que potencialmente pueden introducirse en un curso moderno de estadística.

Abstract. We introduce and investigate the underlying theoretical framework for spatially correlated data in this article. More specifically, we characterise the mechanism that generates the data, investigate its various covariance structures, and characterise the many stationarity classes that are typically considered for this type of data. Furthermore, we investigate the theoretical and empirical semivariogram, which is likely the most extensively used tool for measuring spatial correlation. We believe that this work can be a valuable resource for the study of spatial data and its primary properties, which might be integrated into a modern statistics course.

## §1. Introducción

En la era del Big Data, la disponibilidad y variedad de los datos es cada vez mayor, especialmente en lo que respecta a los datos georrefenciados. Esto se debe tanto al rápido avance tecnológico actual (como la agricultura de precisión y las imágenes satelitales) como al interés social en considerar la espacialidad como fuente de variabilidad de diferentes fenómenos. Por lo cual, creemos que incluir el estudio de los datos espaciales en un curso introductorio de estadística puede ser motivador a partir de aplicaciones reales más actuales. Para ello, este trabajo brinda herramientas para comprender los fundamentos teóricos subyacentes del fenómeno espacial en el contexto de procesos estocásticos como generador de este

*Palabras clave*: Proceso estocástico, Datos Espacialmente Correlacionados, Estacionariedad.
*Keywords*: Stochastic Process, Spatially Correlated Data, Stationarity.

tipo de datos. Los datos espaciales son aquellos cuya particularidad es que cada observación está asociada a una unidad espacial, como ser coordenadas o áreas geográficas. Por ello es imprescindible contar no solo con el dato en sí mismo (i.e. el valor observado de la/s variable/s de interés) sino también con información sobre la posición o referencia geográfica indexada a cada observación. Con esta caracterización de los datos espaciales surge, de forma natural, el concepto de lo que llamaremos valor regionalizado, definido como el valor de la variable de interés en el lugar o punto geográfico donde fue medido.

Dentro del contexto estadístico, la estadística espacial (o geoestadística) es, en parte, la aplicación de la teoría de procesos estocásticos a datos espaciales o, dicho de otra forma, a datos correlacionados espacialmente. Según Wackernagel (2006), la necesidad de tener en cuenta al espacio como fuente de variabilidad se originó a mediados del siglo XX en el campo de la minería y la geología, al tratar de analizar la distribución de las reservas de oro en un depósito mineral a partir de muestras tomadas en algunas ubicaciones espaciales, la cual resultó ser sesgada, revelando la omisión del espacio como fuente de variabilidad. A pesar de que las aplicaciones de la geoestadística han estado tradicionalmente más relacionadas con estudios de fenómenos geológicos y ambientales, recientemente se ha incorporado como una herramienta más de la geografía, la economía y otras ciencias sociales, sea con el objetivo de predecir y completar mapas de datos regionales, como también para modelar la variación espacial de una variable de interés (R. Haining, 2013; R. P. Haining, Kerry, y Oliver, 2010). En particular, los métodos geoestadísticos han tenido un importante alcance en la denominada ciencia regional (Atkinson y Lloyd, 2014). Así, por ejemplo, para la construcción de índices socio-económicos a nivel regional (Montes-Rojas, 2012), para el mapeo de precios de tierras (Derdouri y Murayama, 2020; Morales, Stein, Flacke, y Zevenbergen, 2020; Tsutsumi y Seya, 2008) y análisis de precios de vivienda (Chica-Olmo y Cano-Guervos, 2020; Dubin, 1992; Gámez, Montero, y Rubio, 2000; García Arancibia, Llop, y Lovatto, 2023; Valente, Wu, Gelfand, y Sirmans, 2005), para el análisis de índices de criminalidad (Fernández-Avilés, 2009; Kerry, Goovaerts, Haining, y Ceccato, 2010), o bien para el mapeo de niveles de pobreza o enfermedades utilizando datos regionales (Berke, 2004; Goovaerts, 2008; Vasan y Alcantara, 2016), entre otros. De esta manera, existen múltiples aplicaciones en diferentes ramas de la ciencias aplicadas, que motivan el estudio de datos correlacionados espacialmente.

En este trabajo definimos algunos conceptos y definiciones claves en el estudio de datos espaciales, como así también el marco teórico sobre el cual suelen modelarse. Con este análisis buscamos introducir las herramientas teóricas y metodológicas básicas para el tratamiento de datos espaciales, especificando los supuestos teóricos sobre los que descansan mayormente las tareas de inferencia y predicción estadística, restringiendo la atención en los denominados procesos estacionarios. Asimismo,

sobre la base de este marco teórico, se derivan herramientas de visualización y descripción para el estudio del fenómeno espacial de interés.

Siempre que sea posible, se utilizarán los clásicos datos del río Meuse (Rikken y Van Rijn, 1993) como caso de estudio para ilustrar definiciones y conceptos. Este conjunto de datos proporciona, entre otras variables, mediciones de cuatro tipos diferentes de metales en la capa superior del suelo en una llanura aluvial del río Meuse, cerca del pueblo de Stein (Países Bajos), medida en diferentes localizaciones espaciales, las cuales están identificadas con coordenadas geográficas (latitud y longitud). Nuestro interés será estudiar el comportamiento de uno de esos cuatro metales, el zinc, y en particular su logaritmo. En general, este conjunto de datos se utiliza para predecir el valor de alguno de esos metales en una locación espacial donde el mismo no fue medido. Particularmente, estos datos son utilizados para ejemplificar el método de predicción espacial clásico, denominado *Kriging* (Bivand, Pebesma, Gómez-Rubio, y Pebesma (2008)). A través de todo el trabajo nos referiremos a este ejemplo como *Ejemplo del río Meuse*.

Este trabajo está organizado como sigue: en la Sección 2 introducimos algunos aspéctos básicos que caracterizan a los datos espaciales mientras que en la Sección siguiente mostramos algunas herramientas para un primer análisis exploratorio de los mismos. En la Sección 4 definimos el proceso espacial, sus momentos y los tipos de estacionariedad que pueden presentarse. En la Sección 5 establecemos el modelo que asumiremos para el proceso, definiendo la estructura de correlación que suele asumirse en los modelos paramétricos, el semivariograma. Finalmente, la Sección 6 fue dedicada a las conclusiones del trabajo. Todas las demostraciones y resultados auxiliares se encuentran en el Apéndice.

## §2. Introducción a los datos espaciales

La forma más directa e intuitiva de visualizar a los datos espaciales es a través de los mapas, y nuestro acercamiento a los mismos empieza en muy temprana edad con la escolaridad elemental. En primer lugar, la maestra de primaria nos pide llevar un cierto tipo de mapa (en general, físico y/o político) de una región en particular, como ser el mapa del país o provincia que habitamos. Luego, ya en clases, la maestra nos enseña a colorear o rellenar el mapa en cuestión con algunos atributos que se estén estudiando, delimitando regiones (provincias o localidades), asignando un nombre particular a lo que se delimita, marcando con puntos, líneas o áreas un lugar con algún aspecto que lo hace especial, como ser las capitales o los ríos más importantes. Es así que desde la infancia entramos en contacto con el *dato espacial*, en formato físico y palpable, siendo incluso generadores de nuestra *propia base de datos espaciales* al momento en que al mapa lo cargamos con determinados atributos que nos informan sobre algún aspecto de interés (natural-geológico o político social) en cierto punto o región dentro del mismo. De esta

manera, en su forma más primitiva, un dato espacial es aquel que vincula una ubicación geográfica con una cierta propiedad o atributo descriptivo (Fischer y Wang, 2011). Por ello, marcar con un punto el centro de una ciudad capital de una cierta provincia argentina, es un dato espacial donde en una cierta ubicación (latitud y longitud) se describe un atributo particular (e.g., ciudad capital).

Hoy en día la georreferenciación forma parte de lo cotidiano, haciendo que los datos espaciales cumplan todo tipo de funciones, desde las más simples, como ser la comunicativa (e.g. mandar ubicación para pasar una dirección domiciliaria), a las más complejas, como ser aquellas abocadas a predecir fenómenos de diversa naturaleza, sean delitos, fenómenos ambientales, rendimientos de cultivos, precios de inmuebles o impactos u alcance de ciertas políticas sociales, entre otros. A su vez, en cierta ubicación geográfica podemos tener información de un solo o de varios atributos. En el primer caso, estamos ante la presencia de un dato espacial **univariado**, mientras que en el segundo el dato es **multivariado**. Por ejemplo, para los datos del río Meuse, podríamos analizar solamente la concentración de zinc (problema univariado), o bien considerar concentraciones de varios sólidos como ser de zinc, cadmio, plomo, cobre y sodio, entre otros (problema multivariado). Además, tales atributos pueden ser de naturaleza continua (e.g., niveles de concentración de sólidos en el agua, rendimientos de cultivos, precios de viviendas, salarios medios, etc.) o bien discreta (e.g. categorías de uso de la tierra, tipo de cultivo, presencia de un tipo de sólido, partido o corriente política gobernante en cierta localidad, cantidad de delitos reportados, etc.).

Los datos espaciales pueden clasificarse de varias maneras. Una de ellas, está relacionada con la representación geográfica del **dominio** $D$ en el espacio $\mathbb{R}^d$. Esto es, la forma en que las localizaciones son indexadas en el espacio (Longley, 2005). Cuando se asume un dominio **continuo**, el número de localizaciones o sitios donde podemos observar un atributo es no numerable (Zhang, Atkinson, y Goodchild, 2014). Esto es, el atributo puede estar medido (continuamente) en todas partes del espacio. El caso del ejemplo del río Meuse se corresponde con esta caracterización, dado que entre dos mediciones de concentración de zinc en lugares diferentes, potencialmente existen infinitos valores de tal atributo que podrían medirse. Estos datos continuamente espaciados son los comúnmente denominados **datos geoestadísticos** y, en general, el principal interés en este tipo de datos es la interpolación espacial. No obstante, en la práctica se cuenta con una versión discretizada del domino, pues en general se tienen puntos muestreados en ubicaciones discretas pero que representa un campo de variación continuo (Fischer y Wang, 2011).

Por otra parte, para un dominio $D$ **discreto**, se asume que el número de sitios o localizaciones para el cual podemos observar un atributo es a lo sumo numerable (finito o infinito numerable). En general, estos datos resultan de agregar la medición

de una/s variable/s en localizaciónes finitas, y entre dos consecutivas no existe espacio con puntos que puedan ser medidos o muestrados. Este tipo de datos se denomina **datos reticulares** (en inglés, *lattice data*). Ejemplo de datos recticulares son las tasas de fecundidad o mortalidad por localidad/departamento, tasa de deserción en escuelas/universidades, producción de una región, el rendimiento en un área o el color de un pixel de una imagen satelital, entre otros.

Es interesante observar que los datos geoestadísticos pueden ser transformados a datos reticulares, por ejemplo, generando un teselado del dominio continuo y tomando como dato reticular el centro de la celda, comúnmente denominado **centroide**. No obstante, debe notarse que la elección de los centroides podría influir en la medición de la distancia (Zhao y Wall, 2004) pudiendo incluso modificar la estructura de dependencia espacial existente en los datos. El reverso de esta operación ya no es posible, pues la interpolación espacial podría no tener sentido partiendo datos reticulares.

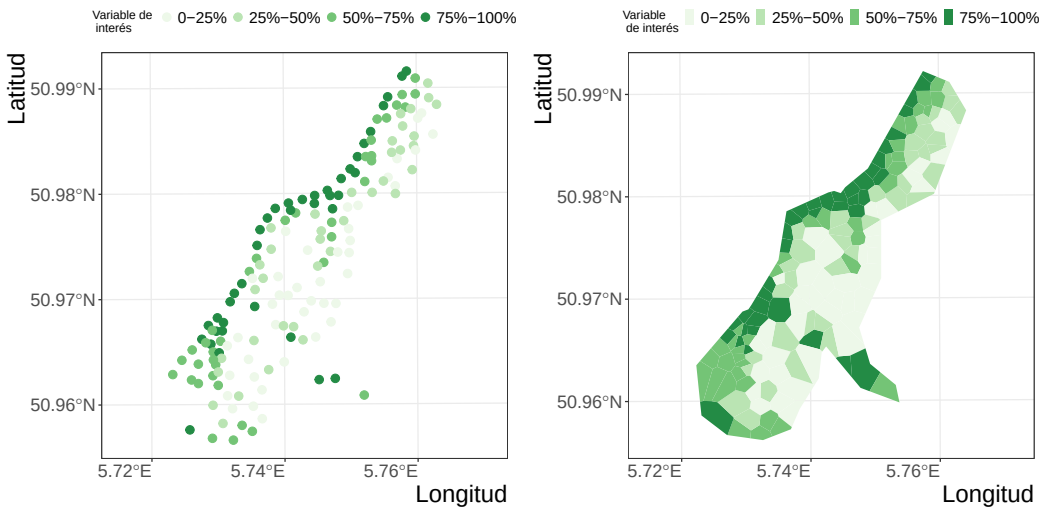
Tanto para datos geoestadísticos como reticulares, el domino $D$ es fijo. Si $D$ es aleatorio se está en presencia de los denominados **patrones puntuales** (en inglés, *point patterns data*). En este caso las localizaciones son generadas por algún proceso aleatorio, como ser la ubicación de árboles o nidos para estudiar una especie, epicentros de terremotos o caídas de rayos.

Otra clasificación útil de los datos espaciales es aquella que los encuadra de acuerdo a diferentes formatos u objetos geométricos; esto es, puntos, líneas o polígonos. Para el primero se tiene un punto en el espacio medido por coordenadas geográficas (latitud y longitud). Para el segundo se tiene un conjunto ordenado de puntos, conectados o no, que forma un segmento, medido por la longitud del mismo. Por último, los polígonos son representados por medio de puntos conectados que encierran una cierta área. Estos diferentes formatos se corresponden con las estructuras cartográficas básicas para los datos vectoriales del sistema de información geográfica GIS.

Más allá de la clasificación dada previamente, en este trabajo consideraremos datos espaciales para un dominio $D$ fijo. Para este tipo de dominios, nos centraremos en el análisis de un sólo atributo o variable medida en una ubicación representada por una coordenada $\mathbf{s} \in D \subset \mathbb{R}^d$. Esto quiere decir que para algún punto espacial $\mathbf{s}_0$ fijo, trabajaremos sobre el atributo $Z(\mathbf{s}_0, \upsilon)$ univariado para todo $\upsilon$ en un espacio muestral $\Omega$. Este enfoque se sitúa en el contexto de **procesos estocásticos espaciales** que será definido formalmente en la Sección 4.

## §3. Análisis exploratorio

En los últimos años, con el desarrollo de software y la posibilidad de manejar grandes cantidades de datos es posible realizar una primera visualización gráfica de los mismos. Como es bien sabido, el análisis exploratorio de datos es la primera




(a) (b)

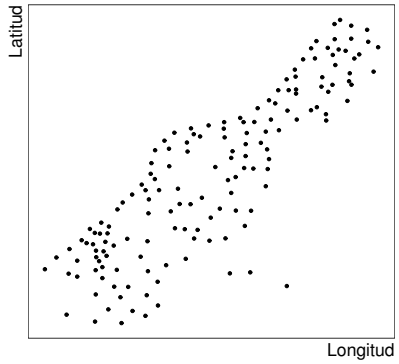
FIGURA 1. Mapas de caja: (a) para coordenadas geográficas; (b) para polígonos.

herramienta que nos permite conocer y resumir los datos de la muestra y con ello poder descubrir errores en la codificación, determinar casos atípicos, comprobar supuestos (e.g. estacionariedad) e incluso encontrar pistas para la modelización. Nuestro principal objetivo en el análisis exploratorio de datos será identificar, si existe, una estructura de correlación en los datos. Más precisamente, queremos cotejar si los valores cercanos son parecidos, y si dicha similitud desaparece a medida que los mismos se alejan. Como veremos en la Sección 4.2, las funciones de semivariograma o covariograma del proceso son la forma de representar esta asociación, estimando relaciones espaciales en lo que es llamado análisis estructural (Sección 4). Para ello, es necesario ver qué nos dicen los datos sobre la existencia o no de una estructura de correlación, de algún patrón estacionario o de tendencias.

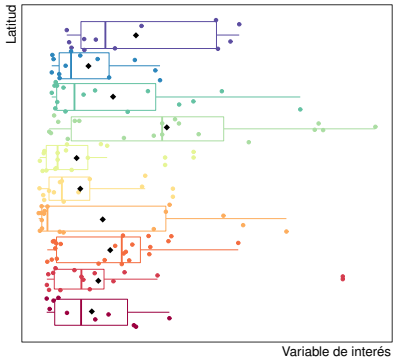
En esta sección presentamos, a partir del ejemplo del río Meuse, algunos gráficos y medidas que podrían ser útiles en esta parte del análisis. Los gráficos nombrados en esta sección son solo algunos de los que se pueden utilizar para un análisis exploratorio espacial, sin embargo no son los únicos (ver por ejemplo los libros Banerjee, Carlin, y Gelfand (2004); Chiles y Delfiner (2012); Cressie (1993)).

### 3.1. Tipos de gráficos exploratorios

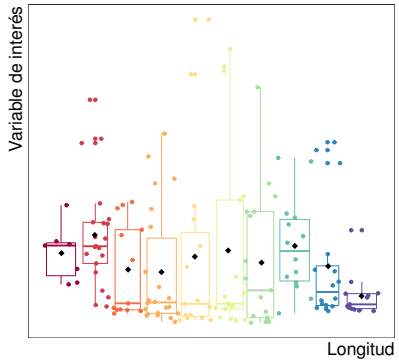
Como primera etapa en el estudio exploratorio se puede realizar un gráfico de los datos muestreados sobre mapas o graficando solamente los mismos sobre las unidades espaciales, ya sean puntos o polígonos. Los **gráficos de puntos** están definidos por sus coordenadas geográficas (latitud y longitud) y pueden ser, por ejemplo, estaciones meteorológicas, inmuebles, un punto en la tierra para medir




(a) Ubicaciones donde fue medida la variable de interés.




(b) Variación de la variable de interés en dirección norte-sur.




(c) Variación de la variable de interés en dirección este-oeste.

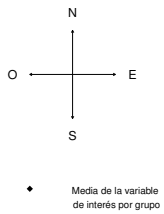
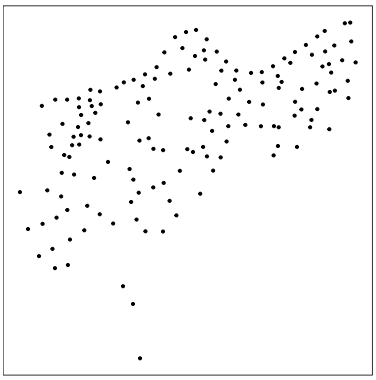



FIGURA 2. Distribución de la variable concentración de zinc de los Datos Meuse en función de las coordenadas (Cressie, 1993, pág. 37).

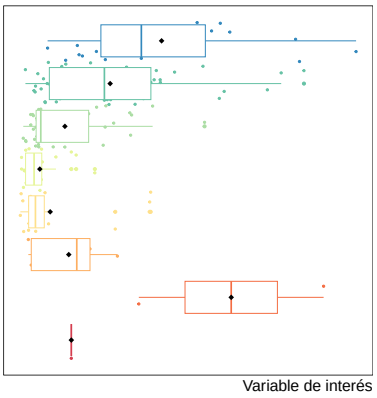
diferentes componentes de la misma como es el caso del ejemplo del río Meuse donde se mide la concentración de zinc (Figura 1 (a)). Los **polígonos** son áreas delimitadas por líneas como por ejemplo países, departamentos, estados, radios censales, o algún teselado sobre el espacio como observamos en la Figura 1 (b).

En este tipo de gráficos, para el análisis de la distribución espacial de la variable de interés, los valores de la misma se pueden categorizar de acuerdo a diferentes intervalos que se calculan a partir de percentiles y luego colorear cada polígono o punto en función de la categoría a la que pertenece. Si dichos intervalos están particionados a partir de los cuartiles, el gráfico se llama **mapa de caja**. En la Figura 1 (a) y (b) podemos observar cada caso, detectando cómo los colores se concentran en ciertas zonas, lo cual nos da una pista de la posible correlación espacial existente, esto quiere decir que los valores de la variable de interés son parecidos si están cerca espacialmente.

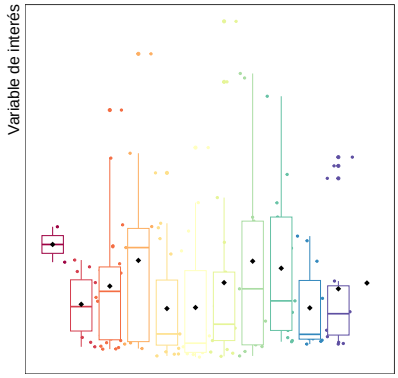
## 3.2. Tendencia



(a) Ubicaciones donde fue medida la variable de interés girados 45 grados en sentido horario.




(b) Variación de la variable de interés en dirección sureste - noroeste.




(c) Variación de la variable de interés en dirección suroeste - noreste.

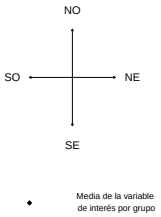



FIGURA 3. Distribución de la variable concentración de zinc de los Datos Meuse en función de las coordenadas (Cressie, 1993, pág. 37).

Si bien los gráficos exploratorios nos dan una idea de cómo es la distribución espacial de los datos, puede tornarse difícil detectar cambios en el comportamiento del promedio o en la variabilidad de los mismos. Para los datos del río Meuse, en términos de tendencia nos interesa estudiar el comportamiento del promedio del logaritmo del zinc como función de las coordenadas. Como para estos datos contamos con una sola medición de la variable de interés por ubicación, no podemos realizar un gráfico del promedio por ubicación espacial. En este caso, lo que hacemos es agrupar los datos en ambas direcciones geográficas (oeste-este y norte-sur) de manera tal de poder calcular el promedio y poder graficar diagrama de cajas que nos muestren el comportamiento de los datos en ambas direcciones. Tales gráficos fueron realizados en las Figuras 2 (b) y 2 (c), respectivamente. Los puntos negros sobre cada diagrama de caja representan la media según el agrupamiento de los

datos por latitud o longitud. De esta visualizacion particular no observamos tendencia en ninguna dirección lo cual nos estaría indicando que el comportamiento promedio de estos datos es constante en las direcciones consideradas.

No obstante, si rotamos el mismo gráfico en 45° en sentido horario, podemos observar cierta tendencia. En particular, como se observa en la Figura 3 (b), la concentración de zinc decrece de noroeste a sureste, revelando así una mayor concentración del metal a orillas del río la cual se disipa a medida que nos alejamos del mismo. Este comportamiento también se puede apreciar en la Figura 1. De este análisis se desprende la importancia de realizar un análisis exploratorio exhaustivo para obtener conclusiones más robustas. No obstante, como veremos más adelante, existen herramientas complementarias que permiten enriquecer este análisis.

### 3.3. Variabilidad espacial

A partir de los aportes de Waldo Tobler con la primera ley de la geografía que establece que todo está relacionado con todo lo demás, pero las cosas cercanas están más relacionadas que las lejanas (Siabato y Guszmán-Manrique, 2019), podemos inferir que dos puntos cercanos asumen valores similares porque estos se generaron en condiciones similares. Por el contrario, a grandes distancias las condiciones son diferentes y se esperan mayores variaciones.

La herramienta más utilizada para medir esta variabilidad es el conocido **semivariograma** empírico. Este gráfico es fundamental a la hora de describir la variación espacial de los datos. Dada la importancia del semivariograma en la geoestadística, tanto la versión empírica como luego la versión teórica del mismo serán introducidas y estudiadas en detalle en las Secciones 4 y 5.

En este apartado introduciremos otra herramienta también utilizada para visualizar la variabilidad espacial presente en los datos, los **diagramas de dispersión** $h$. Aquí (ver luego Sección 5) $h$ indica una distancia y dichos diagramas describen la relación de una misma variable medida en sitios que se encuentran separados por cierta distancia $h$.

La Figura 4 muestra varios paneles donde hemos graficado el logaritmo del zinc para el ejemplo del río Meuse y su asociación para diferentes rangos de la distancia $h$. Como puede observarse, para distancias chicas la asociación es mayor y la misma desaparece a medida que $h$ aumenta. La recta observada es la recta de regresión cuya pendiente tiende a cero cuando la distancia entre sitios aumenta.

## §4. Análisis estructural

En el apartado anterior mostramos, a través del ejemplo del río Meuse, cómo es posible obtener información a partir del análisis exploratorio de los datos espaciales. Dicho análisis sólo genera información sobre la muestra de los datos

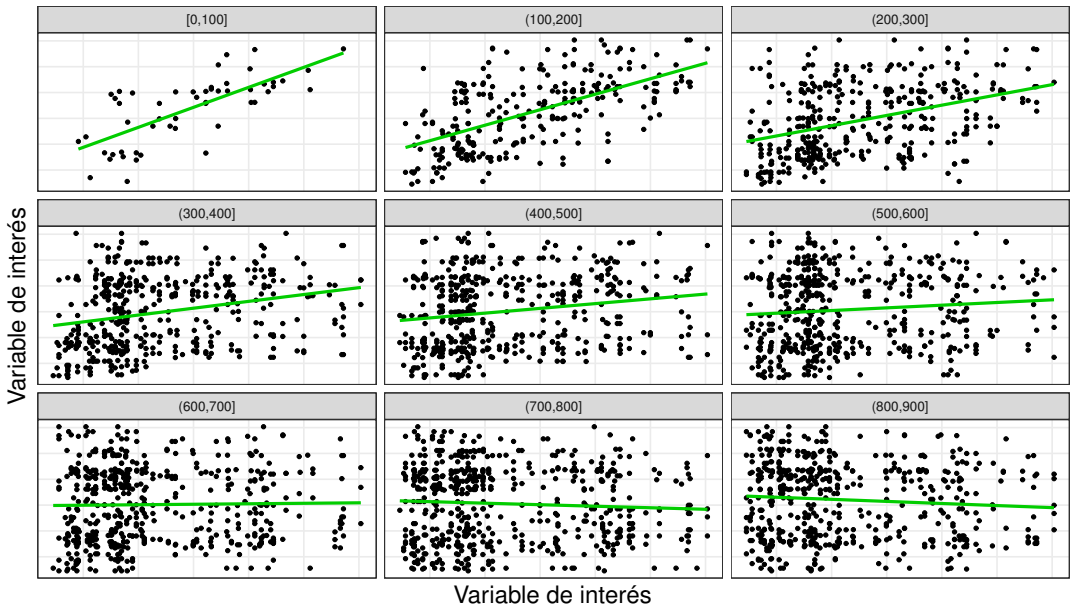



FIGURA 4. Gráfico de dispersión $h$ para los datos Meuse.

observados los cuales, como veremos en esta sección, se denominan **valores (datos) regionalizados.** Unos de los objetivos de la estadística en general y nuestro en particular, es poder obtener información general acerca del fenómeno (social, económico, ambiental, geológico) que genera los datos y así obtener información independiente de la muestra obtenida. Para ello consideramos un enfoque probabilístico, esto quiere decir que el conjunto de los datos regionalizados con los que contamos pueden considerarse como el resultado de un mecanismo aleatorio, al cual llamaremos **proceso estocástico espacial**. En la bibliografía también podemos hallarlo con el nombre de función aleatoria o campo aleatorio (Montero, Fernández-Avilés, y Mateu (2015), Wackernagel (2003)). Para comprender mejor las características de este proceso, los supuestos subyacentes y sus implicancias sobre el tratamiento de datos espaciales nos focalizamos, en esta sección, en la formalización del mismo y de sus propiedades.

### 4.1. Proceso estocástico espacial

Con el objetivo de poder estudiar estos datos de manera analítica y desde un marco estadístico formal, asumimos entonces que los mismos son el resultado de un mecanismo aleatorio. Específicamente, dado $d \in \mathbb{N}$, $D \subset \mathbb{R}^d$ y $(\Omega, \mathcal{A}, P)$ un espacio de probabilidad, se llama **proceso estocástico espacial** al conjunto

$$\{Z(\mathbf{s}, \upsilon) : \mathbf{s} \in D, \upsilon \in \Omega\},$$

con $Z$ una función a valores reales tal que $Z : D \times \Omega \longrightarrow \mathbb{R}$.

Tomando al valor regionalizado como resultado de un proceso estocástico, obtenemos dos caracterizaciones del mismo. En particular, si fijamos una localización

$\mathbf{s}_0 \in D$, tenemos una variable aleatoria $Z(\mathbf{s}_0, \cdot)$ con su variabilidad asociada. Si en su lugar fijamos $\upsilon_0 \in \Omega$, entonces tenemos la realización $z : D \subset \mathbb{R}^d \longrightarrow \mathbb{R}$, denominada comúnmente como **variable regionalizada o regionalización** (e.g. Wackernagel (2006)). Esta doble caracterización nos permite tener en cuenta las diferentes fuentes de variabilidad del proceso. Por un lado, la variabilidad del fenómeno en cada sitio, entender que el dato existe con cierta probabilidad y que existen otros valores posibles para cada localización. Por otro lado, formalizar la variabilidad entre sitios, considerando las distancias entre ellos y la forma de la dependencia con la variable de interés.

En el siguiente diagrama ilustramos estas definiciones junto con la notación a ser utilizada. Cabe aclarar que el símbolo $\doteq$ se utiliza, y se utilizará a lo largo del artículo, para introducir nueva notación.

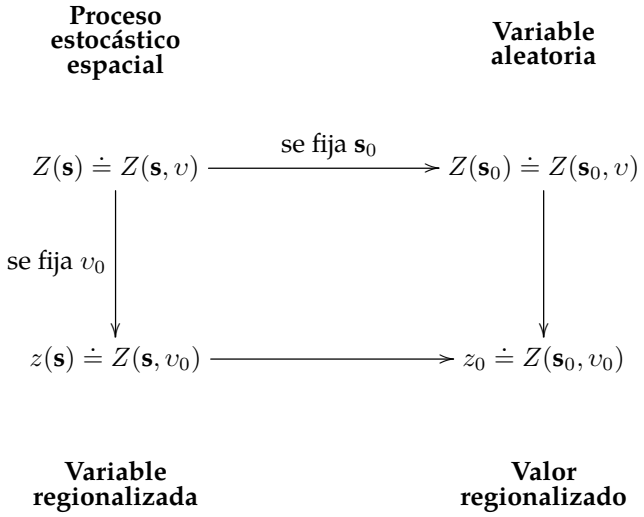


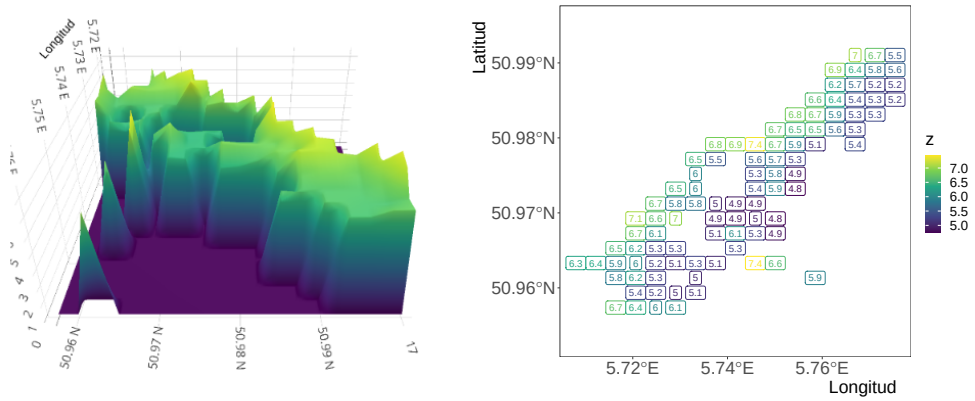

FIGURA 5. Diagrama resumen de las definiciones junto con la notación a ser utilizada a lo largo de todo el trabajo (Wackernagel, 2003, pág. 28.)

.

Con esta caracterización, si fijamos un conjunto de localizaciones $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$ y un $\upsilon_0 \in \Omega$, tenemos un vector $(z_1, \ldots, z_n)^T = (Z(\mathbf{s}_1, \upsilon_0), \ldots, Z(\mathbf{s}_n, \upsilon_0))^T$ de **valores regionalizados** los cuales representan los **datos empíricos**.

Para el ejemplo del río Meuse, $d = 2$ ya que las mediciones de la concentración de zinc fueron tomadas para determinadas coordenadas geográficas, esto es, latitud y longitud. La región $D$ es el dominio del proceso, en este caso representado por la región sobre la cual tenemos información. Más precisamente, en la Figura 6 presentamos una adaptación de los datos Meuse en la región $D$. En el panel izquierdo de la misma observamos una variable regionalizada $Z(\mathbf{s}, \upsilon_0)$ para un $\upsilon_0$ fijo y

para todo $\mathbf{s} \in D$ y en el panel derecho un conjunto de los valores regionalizados nombrados anteriormente y corresponden a los datos con los que contamos.




(a) (b)

FIGURA 6. Para el ejemplo del río Meuse: (a) variable regionalizada; (b) conjunto de valores regionalizados.

Dado que un proceso estocástico espacial queda determinado por sus distribuciones finito dimensionales, resulta fundamental definir su función de distribución para así poder estudiar al proceso en el contexto estadístico.

Específicamente, podemos definir la **distribución del proceso estocástico** $Z(\mathbf{s})$ como la colección de distribuciones conjuntas de dimensión finita de vectores aleatorios $(Z(\mathbf{s}_1), \cdots, Z(\mathbf{s}_n))^T$, para conjuntos finitos de sitios $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$,

$$\begin{aligned} F_{\mathbf{s}_1,\cdots,\mathbf{s}_n}(z_1, \cdots, z_n) &\doteq \mathbb{P}(\{\upsilon : Z(\mathbf{s}_1, \upsilon) \leq z_1, \cdots, Z(\mathbf{s}_n, \upsilon) \leq z_n\}) \\ &= \mathbb{P}(Z(\mathbf{s}_1) \leq z_1, \cdots, Z(\mathbf{s}_n) \leq z_n). \end{aligned} \tag{4.1}$$

De esta manera, con la especificación (4.1) para todos los posibles conjuntos de puntos de muestreo $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$ y para todo $n$ tenemos una caracterización completa de un proceso estocástico. Un caso particular de especial importancia es el denominado **proceso gaussiano** donde las distribuciones de dimensión finita definidas en la Ecuación (4.1) son normales multivariadas. Para denotar la **función de densidad o función de probabilidad puntual conjunta** de las variables aleatorias obtenidas en localizaciones específicas $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$ , usaremos $p_{\mathbf{s}_1,\ldots,\mathbf{s}_n}(z_1, \ldots, z_n)$ y $p_{\mathbf{s}}(z)$ para la densidad o probabilidad puntual de $Z(\mathbf{s})$, para cada $\mathbf{s} \in D$.

Es bien sabido que al estudiar una variable o un conjunto de variables, determinar su función de distribución es un problema en sí mismo. Dicha complejidad reside en que la medición experimental de las distribuciones de probabilidad multidimensionales (e incluso unidimensionales) es una tarea muy complicada,

resultando excesivamente costoso estimar dicha función en todo su dominio. Por ejemplo, conocer la distribución completa del proceso estocástico que dio lugar a la concentración de zinc a orillas del río Meuse, puede ser muy complejo. Sin embargo, este problema suele simplificarse estudiando algunas medidas que caracterizan a la distribución y que de hecho están involucradas con las medidas de centralidad y variabilidad de los datos. Estas medidas son los momentos.

El primer momento, **valor esperado**, **función media** o simplemente **media**, denotado por $\mu(\mathbf{s})$, es la función que describe la variación no aleatoria (o tendencia) del proceso $Z(\mathbf{s})$. De esta manera, para $\mathbf{s} \in D$ la media de $Z(\mathbf{s})$ está dada por

$$\mu(\mathbf{s}) = \mathbb{E}\left[Z(\mathbf{s})\right] = \int_{-\infty}^{\infty} z p_{\mathbf{s}}(z) dz.$$

Para un proceso de valores discretos, la integral se reemplaza por la suma y en este caso $p_{\mathbf{s}}(z)$ es la función de probabilidad puntual de $Z(\mathbf{s})$. Aunque no lo volveremos a mencionar, este comentario se replicará cada vez que aparezcan integrales.

En la Sección 3.2 observamos que, exploratoriamente, no existe tendencia en la concentración de zinc a orillas del río Meuse, y a esto lo hemos observado a partir de los promedios, es decir, los primeros momentos medidos en cada dirección.

Si bien la media es una medida que caracteriza de una manera simple la centralidad del proceso, no es completamente informativa para describir el comportamiento del mismo. Por ejemplo, podríamos tener dos regiones con la misma concentración media de zinc pero con una dispersión muy diferente como se muestra en la Figura 7. En la misma hemos graficado datos simulados sobre una cuadrícula en $D$ para el ejemplo del río Meuse, utilizando para ello un proceso con media constante 6 para todo $\mathbf{s}$. En el panel izquierdo observamos que los datos simulados se alejan de la media de una manera homogénea sobre la región de interés. En cambio, en el panel derecho, los datos se comportan de forma diferente, esto es, a medida que aumenta la longitud (coordenada $x$) los mismos se alejan de la media.

Para modelar esta variabilidad es necesario definir los tres momentos de segundo orden más utilizados en geoestadística que son la varianza, la covarianza (o la covarianza normalizada, que comúnmente se conoce como correlación) y el variograma. Para $\mathbf{s} \in D$, la **varianza** de $Z(\mathbf{s})$ está dada por

$$\text{Var}\left(Z(\mathbf{s})\right) \doteq \mathbb{E}[(Z(\mathbf{s}) - \mu(\mathbf{s}))^2] = \mathbb{E}\left[Z^2(\mathbf{s})\right] - \mu^2(\mathbf{s}),$$

donde,

$$\mathbb{E}\left[Z^2(\mathbf{s})\right] = \int_{-\infty}^{\infty} z^2 p_{\mathbf{s}}(z) dz.$$

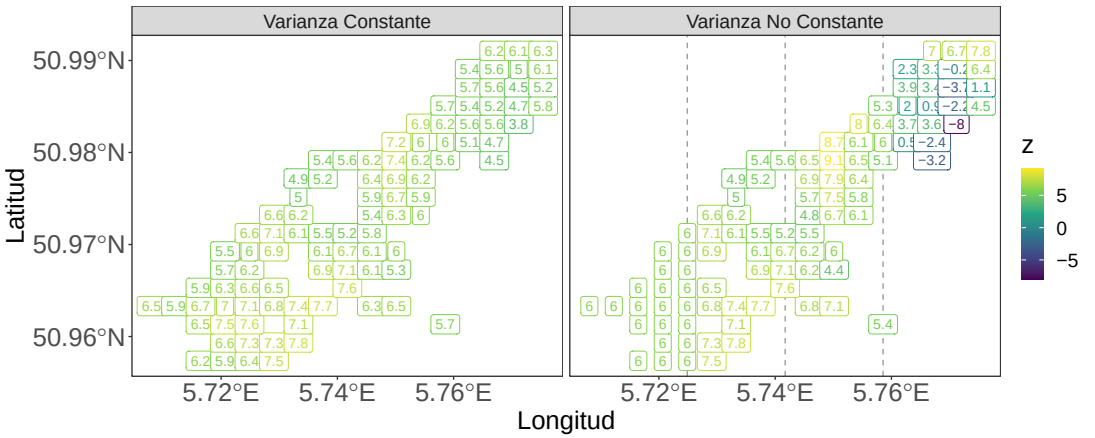



FIGURA 7. Datos simulados sobre las coordenadas de los datos Meuse. Panel izquierdo: varianza constante. Panel derecho: varianza no constante.

La varianza es un caso particular de la **covarianza o autocovarianza** $C : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, la cual queda definida, para todo $\mathbf{s}_i, \mathbf{s}_j \in D$, como

$$\begin{aligned} C(\mathbf{s}_i, \mathbf{s}_j) &\doteq \text{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) \\ &= \mathbb{E}\left[(Z(\mathbf{s}_i) - \mu(\mathbf{s}_i))(Z(\mathbf{s}_j) - \mu(\mathbf{s}_j))\right] \\ &= \mathbb{E}\left[Z(\mathbf{s}_i)Z(\mathbf{s}_j)\right] - \mu(\mathbf{s}_i)\mu(\mathbf{s}_j), \end{aligned}$$

donde,

$$\mathbb{E}\left[Z(\mathbf{s}_i)Z(\mathbf{s}_j)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} z_i z_j p_{\mathbf{s}_i,\mathbf{s}_j}(z_i, z_j) dz_i dz_j.$$

La covarianza mide el grado de dependencia espacial existente en los datos. Por ejemplo, para los datos del río Meuse, mide cuánto afecta la concentración de zinc medida en cierta localización a los valores de la misma medidos en localizaciones vecinas. En este sentido, podría ser de interés modelar si a distancias mayores los niveles de zinc tienen relación, lo cual veremos con más detalle en la Sección 5.

Asociado a la covarianza tenemos el **variograma** del proceso el cual, para todo $\mathbf{s}_i, \mathbf{s}_j \in D$, está definido como

$$\begin{aligned} 2\gamma(\mathbf{s}_i, \mathbf{s}_j) &\doteq \text{Var}\left(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\right) \\ &= \text{Var}(Z(\mathbf{s}_i)) + \text{Var}(Z(\mathbf{s}_j)) - 2\text{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j)). \end{aligned}$$

La función $\gamma : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ es conocida como el **semivariograma** del proceso $Z(\mathbf{s})$ y es una herramienta clásica muy utilizada ya que cuantifica la disimilaridad del proceso para diferentes locaciones. Como lo hemos mencionado anteriormente, tanto $Z(\mathbf{s}_i)$ como $Z(\mathbf{s}_j)$ son variables aleatorias y por lo tanto la diferencia $Z(\mathbf{s}_i) - Z(\mathbf{s}_j)$ también lo es. Por lo tanto, dado que el variograma es la varianza de dicha diferencia, el mismo proporciona una medida de dependencia espacial.

Dados dos sitios a cierta distancia, si las variables $Z(\mathbf{s}_i)$ y $Z(\mathbf{s}_j)$ toman valores similares, la variabilidad de la diferencia será pequeña, lo que resultará en valores bajos del semivariograma. Valores altos se observarán cuando exista una mayor diferencia entre los valores de la variable. Si pensamos el variograma en función de las distancias entre sitios, y suponemos dependencia espacial, lo que esperamos es que para distancias cercanas a cero los valores del semivariograma sean pequeños y aumenten a medida que las distancias entre los sitios crezcan. Por ello el semivariograma es el primer paso en todo análisis geoestadístico ya que nos permite visualizar la existencia o no de dicha dependencia, aunque no es su único uso. Por ejemplo, en predicción espacial, más precisamente en *Kriging*, la estimación del semivariograma es imprescindible para el cálculo de los pesos utilizados en la combinación lineal pesada que lo define.

Finalmente, se define la **función de correlación o autocorrelación** que no es más que la covarianza estandarizada, normalizada o libre de unidades. La misma se define, para todo $\mathbf{s}_i, \mathbf{s}_j \in D$ como

$$\rho(i,j) \doteq \frac{C(\mathbf{s}_i, \mathbf{s}_j)}{\sqrt{\text{Var}\left(Z(\mathbf{s}_i)\right) \text{Var}\left(Z(\mathbf{s}_j)\right)}}.$$

### 4.2. Estacionariedad

Supongamos que para el ejemplo del río Meuse nos interesa estudiar si la función de distribución del proceso que generó los datos es exactamente la misma para cualquier locación, o podría considerarse alguna relajación de este supuesto. En esta dirección, detallamos tres tipos de estacionariedad conmúnmente descriptos en la literatura (Banerjee y cols. (2004); Chiles y Delfiner (2012); Cressie (1993); Montero y cols. (2015); Wackernagel (2003)). Los supuestos de estacionariedad intentan describir el comportamiento de una amplia gama de fenómenos imponiendo condiciones a los momentos del proceso. En general, la estacionariedad es adoptada por un amplio abanico de modelos y métodos para datos dependientes, ya que permite simplificar el análisis y hacer inferencias más sólidas sobre el comportamiento del proceso.

El tipo de estacionariedad más fuerte que podemos considerar para un proceso $Z(\mathbf{s})$ es la **estacionariedad en sentido estricto** o simplemente **estacionariedad estricta**. La misma implica que dado un proceso $Z(\mathbf{s})$ con esperanza y varianza finita y un conjunto de localizaciones $\{\mathbf{s}_1, \ldots, \mathbf{s}_k\}$, para cualquier vector de separación $\mathbf{h} \in \mathbb{R}^d$, $(Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_k))$ y $(Z(\mathbf{s}_1 + \mathbf{h}), \ldots, Z(\mathbf{s}_k + \mathbf{h}))$ tienen la misma función de distribución conjunta para todo $k \in \mathbb{N}$.

Un tipo de estacionariedad más débil que la estricta es la conocida como **estacionariedad débil o de segundo orden**. Más precisamente, diremos que un proceso $Z(\mathbf{s})$ es **estacionario de segundo orden** o **débilmente estacionario**, si tiene

momento de segundo orden finito, es decir, $\mathbb{E}\left[Z^2(\mathbf{s})\right] < \infty$ y verifica las siguientes dos condiciones:

(i) Para todo $\mathbf{s} \in D$, la esperanza de $Z(\mathbf{s})$ existe y es constante. Es decir,

$$\mathbb{E}[Z(\mathbf{s})] = \mu. \tag{4.2}$$

(ii) Para todo $\mathbf{s}_i, \mathbf{s}_j \in D$ y cada par de variables $Z(\mathbf{s}_i), Z(\mathbf{s}_j)$ medidas en esos sitios, la covarianza existe y solo depende del vector de separación $\mathbf{s}_i - \mathbf{s}_j$ entre los sitios $\mathbf{s}_i, \mathbf{s}_j$. Esto es,

$$\text{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) \doteq C(\mathbf{s}_i - \mathbf{s}_j). \tag{4.3}$$

En este caso la función $C(\cdot)$ es llamada **covariograma** o **función de covarianza estacionaria**. Para simplificar la notación, definimos a dicho vector de separación como $\mathbf{h} \doteq \mathbf{s}_i - \mathbf{s}_j$, a partir de lo cual, para todo $\mathbf{s} \in D$, tenemos que

$$C(\mathbf{s} + \mathbf{h}, \mathbf{s}) = \text{Cov}(Z(\mathbf{s} + \mathbf{h}), Z(\mathbf{s})) \doteq C(\mathbf{h}).$$

En lo que resta del artículo utilizamos esta notación donde $\mathbf{s} + \mathbf{h}$ y $\mathbf{s}$ son sitios que están a separación $\mathbf{h}$.

Como lo mencionamos anteriormente, la estacionariedad de segundo orden es más débil que la estacionariedad estricta en el sentido de que si un proceso es estrictamente estacionario entonces es débilmente estacionario.

Dado que los momentos de segundo orden de un proceso débilmente estacionario son finitos, la varianza del mismo existe, es finita y constante,

$$\text{Var}(Z(\mathbf{s})) = \text{Cov}(Z(\mathbf{s}), Z(\mathbf{s})) = C(\mathbf{0}) \doteq \sigma^2 < \infty, \quad \forall\, \mathbf{s} \in D.$$

En este contexto el variograma está dado por

$$\begin{aligned} 2\gamma(\mathbf{s}, \mathbf{s} + \mathbf{h}) &= \text{Var}(Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s})) \\ &= \text{Var}(Z(\mathbf{s} + \mathbf{h})) + \text{Var}(Z(\mathbf{s})) - 2\text{Cov}(Z(\mathbf{s} + \mathbf{h}), Z(\mathbf{s})) \\ &= C(\mathbf{0}) + C(\mathbf{0}) - 2C(\mathbf{h}) \qquad \text{(por (4.3))} \\ &= 2(C(\mathbf{0}) - C(\mathbf{h})), \end{aligned} \tag{4.4}$$

por lo que el semivariograma de un proceso débilmente estacionario solo depende del vector de separación $\mathbf{h}$ y queda expresado como

$$\gamma(\mathbf{s}, \mathbf{s} + \mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}) \doteq \gamma(\mathbf{h}). \tag{4.5}$$

Otra propiedad importante de la función de covarianza $C$, en el marco de los procesos estacionarios de segundo orden, es que está acotada por su valor en el origen. Esto es,

$$|C(\mathbf{h})| \leq C(\mathbf{0}). \tag{4.6}$$

En efecto, de $0 \leq \text{Var}(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})) = 2(C(\mathbf{0}) - C(\mathbf{h}))$ se tiene que $C(\mathbf{h}) \leq C(\mathbf{0})$ y, análogamente, de $0 \leq \text{Var}(Z(\mathbf{s}+\mathbf{h}) + Z(\mathbf{s})) = 2(C(\mathbf{0}) + C(\mathbf{h}))$ se tiene que $-C(\mathbf{0}) \leq C(\mathbf{h})$ de donde se sigue (4.6).

Observemos además que si el proceso $Z(\mathbf{s})$ es estacionario de segundo orden, $\gamma(\mathbf{h})$ también está acotada pues, de (4.5) y (4.6) se tiene que,

$$|\gamma(\mathbf{h})| = |C(\mathbf{0}) - C(\mathbf{h})| \leq |C(\mathbf{0})| + |C(\mathbf{h})| \leq 2C(\mathbf{0}).$$

Esto nos dice que si $\gamma(\mathbf{h})$ no está acotada, el proceso $Z(\mathbf{s})$ no puede ser estacionario de segundo orden pues en ese caso no existiría la covarianza.

De la Ecuación (4.5) se revela que un semivariograma queda determinado a partir de una función de covarianza. Sin embargo, el recíproco no es cierto ya que el semivariograma puede crecer indefinidamente mientras que la covarianza no. A modo de ejemplo, (Wackernagel, 2003, Ejemplo 7.1, pág. 52) el proceso llamado movimiento browniano fraccional tiene un variograma de la siguiente forma,

$$\gamma(\mathbf{h}) = b\|\mathbf{h}\|^p, \qquad \text{con } 0 < p < 2 \text{ y } b > 0. \tag{4.7}$$

Dicha función no podría dar lugar a una función de covarianza ya que crece sin límites. Esta función de semivariograma supera el marco de los procesos débilmente estacionarios; esto es, $\gamma(\mathbf{h})$ existe pero $C(\mathbf{h})$ no. Esto significa que existe un conjunto de procesos que no son débilmente estacionarios pero que sin embargo podemos modelar su autocorrelación a partir de la función de semivariograma $\gamma(\mathbf{h})$. Describimos dicho conjunto a continuación.

Dado que las hipótesis de media constante y de varianza finita, asumidas por la estacionariedad débil, pueden resultar muy restrictivas en varios contextos, es que suele considerarse un tercer (y más débil) tipo de estacionariedad que es llamada **estacionariedad intrínseca**. Esto permite, por ejemplo, tratar con procesos con capacidad de variación infinita, a parir de suponer la finitud de la varianza de sus diferencias $Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})$. Formalmente, diremos que un proceso $Z(\mathbf{s})$ es **intrínsecamente estacionario** o simplemente **intrínseco** si, para todo $\mathbf{s} \in D$ verifica que

(i) La esperanza del proceso diferencia es nula. Es decir,

$$\mathbb{E}[Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})] = 0. \tag{4.8}$$

(ii) El variograma depende solo del vector de separación $\mathbf{h}$. Esto es,

$$2\gamma(\mathbf{s}, \mathbf{s}+\mathbf{h}) = \text{Var}(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})) \doteq 2\gamma(\mathbf{h}). \tag{4.9}$$

Una identidad muy utilizada para procesos intrínsecos se desprende de las Ecuaciones (4.8) y (4.9) y es la siguiente:

$$\gamma(\mathbf{h}) = \frac{1}{2}\text{Var}\left(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})\right) = \frac{1}{2}\mathbb{E}[(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s}))^2]. \tag{4.10}$$

A partir de todo lo anterior, es sencillo probar que todo proceso estacionario de segundo orden es intrínsecamente estacionario pues, la condición (4.8) es una consecuencia inmediata de (4.2) y la condición (4.9) de (4.4).

El recíproco de este resultado es falso. Es decir, existen procesos intrínsecamente estacionarios que no son estacionarios de segundo orden. Por ejemplo, hemos visto que la función de semivariograma dada en (4.7) no es acotada y como consecuencia no cuenta con una función de covarianza asociada, lo que hace que supere el marco de los procesos débilmente estacionarios. Sin embargo, esta función de semivariograma podría estar asociada a un proceso intrínseco. Otro ejemplo es el proceso discreto de Wiener-Levy (Montero y cols., 2015, pág. 17) dado por $Z_{k+1} = Z_k + \epsilon_k$, con $\epsilon_k \sim N(0,1)$ variables aleatorias independientes, indexado en los naturales $k$ con $Z_0 = 0$, el cual puede probarse (ver Lema 8.1 en el Apéndice) que es un proceso intrínsecamente estacionario pero no débilmente estacionario.

Es importante destacar que, a lo largo de todo el artículo, cuando consideramos que la función de semivariograma $\gamma(\mathbf{h})$ o el covariograma $C(\mathbf{h})$ existen, estamos asumiendo que las mismas están asociadas a algún proceso intrínsecamente estacionario o débilmente estacionario, respectivamente.

## §5. Modelo espacial

Luego de realizar el análisis exploratorio de la concentración de zinc en la llanura aluvial del río Meuse, podemos hacer varias afirmaciones acerca de la tendencia y la dependencia espacial existente en los datos, y con ello, poder encontrar un modelo teórico para el proceso espacial subyacente. Como es bien sabido, modelar correctamente un experimento nos permite luego poder infererir y predecir eventos de manera eficiente. En esta dirección es usual descomponer al proceso $Z(\mathbf{s})$ como una suma de dos componentes, esto es,

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + e(\mathbf{s}), \tag{5.1}$$

donde la función media $\mu(\mathbf{s}) = \mathbb{E}[Z(\mathbf{s})]$ es continua y determinística y el proceso de error $e(\mathbf{s})$ es estocástico con media cero y cierta estructura de dependencia espacial:

- Si $e(\mathbf{s})$ es débilmente estacionario, para todo $\mathbf{s}_i, \mathbf{s}_j \in D$ tendremos que
$$\text{Cov}(e(\mathbf{s}_i), e(\mathbf{s}_j)) = C(\mathbf{s}_i - \mathbf{s}_j).$$
- Si $e(\mathbf{s})$ es intrínsecamente estacionario, para todo $\mathbf{s}_i, \mathbf{s}_j \in D$ tendremos que
$$\text{Var}(e(\mathbf{s}_i) - e(\mathbf{s}_j)) = 2\gamma_e(\mathbf{s}_i - \mathbf{s}_j),$$
donde $2\gamma_e(\mathbf{s}_i, \mathbf{s}_j)$ es el **variograma del residuo** el cual es llamado así ya que, en la práctica, en lugar de usar los datos originales para hallarlo, se

utilizan los residuos obtenidos a partir de la estimación de la función media (Wackernagel, 2006, pag. 181).

El proceso de error $e(\mathbf{s})$ además de capturar la variación espacial a pequeña escala explica, en parte, el error de medición que puede ocurrir en el proceso de recopilación de datos. Este componente generalmente no tiene estructura espacial, por lo tanto, para algunos propósitos puede ser deseable separarlo explícitamente de la componente espacialmente dependiente. Es decir, podemos escribir

$$e(\mathbf{s}) = \eta(\mathbf{s}) + \epsilon(\mathbf{s}), \tag{5.2}$$

donde $\eta(\cdot)$ y $\epsilon(\cdot)$ son independientes, $\eta(\cdot)$ es la componente que mide la dependencia espacial y $\epsilon(\cdot)$, comúnmente conocido como **efecto pepita** o proceso de ruido blanco, es un proceso con media cero, espacialmente no correlacionado y que modela el error de medición. Más precisamente, su función de covarianza está dada por

$$\text{Cov}(\epsilon(\mathbf{s}), \epsilon(\mathbf{s}+\mathbf{h})) = \begin{cases} \sigma_\epsilon^2 \geq 0, & \text{si } \mathbf{h} = \mathbf{0}, \\ 0, & \text{si } \mathbf{h} \neq \mathbf{0}. \end{cases}$$

Conociendo los supuestos sobre el proceso de error $e(\mathbf{s})$ y la función media $\mu(\mathbf{s})$, podemos deducir el comportamiento del proceso $Z(\mathbf{s})$. Más precisamente, si $\mu(\mathbf{s})$ es constante, entonces el proceso $Z(\mathbf{s})$ hereda las propiedades de $e(\mathbf{s})$. Esto es, si el proceso $e(\mathbf{s})$ es estacionario de segundo orden (intrínsecamente estacionario), entonces $Z(\mathbf{s})$ es estacionario de segundo orden (intrínsecamente estacionario). Formalmente estos enunciados se encuentran sintetizados en los Lemas 8.2 y 8.3, respectivamente, cuyas demostraciones están dadas en el Apéndice.

En la práctica, el semivariograma y el covariograma se estiman empíricamente a partir de los datos como lo veremos en la próxima sección.

## 5.1. Semivariograma y covariograma empíricos

En muchos casos, incluso en aquellos donde se cuenta con información suficiente para suponer que el proceso es débilmente estacionario, se prefiere utilizar el semivariograma antes que el covariograma pues este último requiere conocer la media del proceso $Z(\mathbf{s})$. En la práctica, ésta es desconocida y debe estimarse a partir de los datos, lo que introduce un sesgo en la estimación del covariograma. Para demostrar este hecho, definimos primero las versiones empíricas del covariograma y semivariograma de un proceso $Z(\mathbf{s})$.

Bajo las condiciones de estacionariedad intrínseca (4.8) y (4.9) o las condiciones de estacionariedad de segundo orden (4.2) y (4.3), los estimadores empíricos del semivariograma y el covariograma son construidos a partir del método de los momentos. Específicamente, dado un conjunto de sitios $\mathbf{s}_1, \ldots, \mathbf{s}_n$ y una dirección $\mathbf{h}$, un estimador del semivariograma está dado por,

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} (Z(\mathbf{s}_i + \mathbf{h}) - Z(\mathbf{s}_i))^2 \tag{5.3}$$

y un estimador del covariograma por (Smith, 2014),

$$\hat{C}(\mathbf{h}) = \frac{1}{N(\mathbf{h}) - 1} \sum_{i=1}^{N(\mathbf{h})} (Z(\mathbf{s}_i + \mathbf{h}) - \bar{Z}_{\mathbf{h}})(Z(\mathbf{s}_i) - \bar{Z}_{(\mathbf{h})}), \tag{5.4}$$

donde

$$\bar{Z}_{\mathbf{h}} \doteq \frac{1}{N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} Z(\mathbf{s}_i + \mathbf{h}) \qquad \text{y} \qquad \bar{Z}_{(\mathbf{h})} \doteq \frac{1}{N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} Z(\mathbf{s}_i), \tag{5.5}$$

y $N(\mathbf{h})$ es la cantidad de pares de puntos que están a una distancia $\mathbf{h}$ los cuales, y sin pérdida de generalidad, suponemos que se presentan en el siguiente orden $\{(\mathbf{s}_1, \mathbf{s}_1 + \mathbf{h}), \ldots, (\mathbf{s}_{N(\mathbf{h})}, \mathbf{s}_{N(\mathbf{h})} + \mathbf{h})\}$.

Observemos que el covariograma empírico (5.4) a diferencia del variograma, utiliza la media muestral haciendo que $\hat{C}(\mathbf{h})$ sea sesgado. Por otro lado, podemos probar que el semivariograma empírico $\hat{\gamma}(\mathbf{h})$ definido en (5.3) es insesgado. Estos resultados se exponen en el siguiente lema, cuya demostración se encuentra en el Apéndice.

**Lema 5.1.** *El semivariograma empírico definido en* (5.3) *satisface que*

$$\mathbb{E}[\hat{\gamma}(\boldsymbol{h})] = \gamma(\boldsymbol{h}),$$

*y el covariograma empírico definido en* (5.4) *satisface que*

$$\mathbb{E}(\hat{C}(\boldsymbol{h})) = C(\boldsymbol{h}) - \frac{1}{N(\boldsymbol{h})(N(\boldsymbol{h}) - 1)} \sum_{i=1}^{N(\boldsymbol{h})} \sum_{j \neq i}^{N(\boldsymbol{h})} Cov(Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j + \boldsymbol{h}))$$

Entonces, a partir de los datos disponibles podemos calcular el semivariograma empírico (5.3), el cual resume la relación espacial presente en los datos.

Dado el vector de separación $\mathbf{h}$, para cada dirección podemos graficar los llamados **semivariogramas direccionales**, los cuales juegan un papel muy importante en el análisis de datos espaciales pues brindan información acerca de la variabilidad del proceso en cada dirección. Si éstos son diferentes para cada una de ellas, entonces el semivariograma (5.3) depende de $\mathbf{h}$ no sólo en magnitud sino también en dirección, dando lugar a lo que se conoce como comportamiento **anisotrópico**. Por otro lado, si dicho gráfico se mantiene constante para cada dirección decimos que el semivariograma depende solo de la magnitud $h \doteq ||\mathbf{h}||$ que es la **distancia** a la cual se encuentran los datos, y es llamado **isotrópico**. En este caso construimos un único semivariograma al cual llamamos **semivariograma omnidireccional** (o

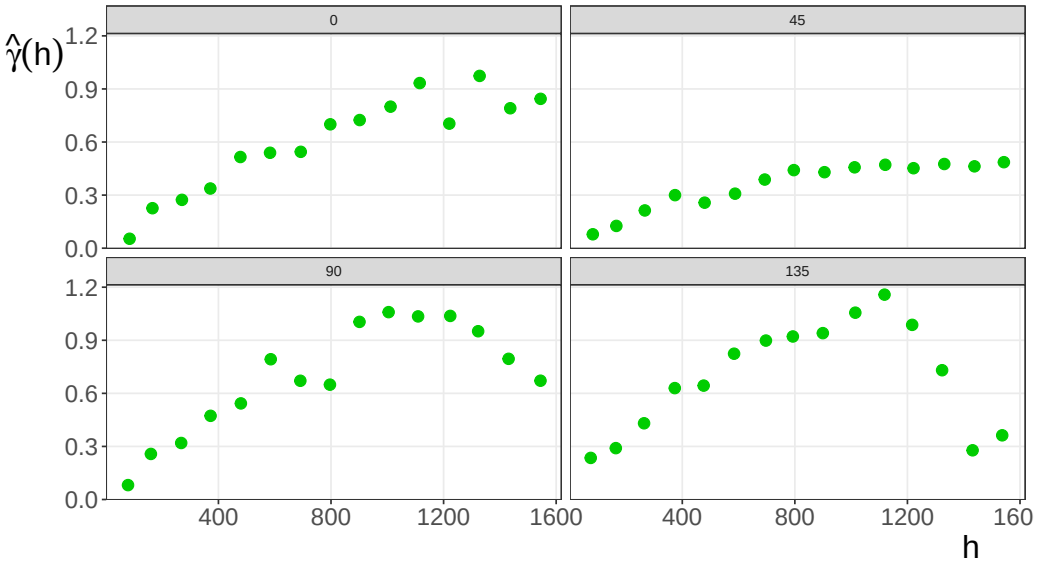



FIGURA 8. Semivariograma empírico para los datos del río Meuse.

simplemente semivariograma) y que se construye considerando todos los puntos que se encuentran a cierta distancia sin fijar una dirección.

En la Figura 8 graficamos los semivariogramas empíricos para los datos del río Meuse para diferentes direcciones (para cada panel), fijando una distancia máxima $h_{max}$ la cual, generalmente, se toma como un tercio de la distancia espacial máxima existente entre lo datos. Para realizar cada uno de estos gráficos se divide el intervalo $[0, h_{max}]$ en subintervalos $[h_i, h_j]$ de forma tal que en cada uno de ellos exista una cantidad mínima de pares de sitios $N(\mathbf{h})$ cuyas distancias se encuentren en ese intervalo. Luego, para cada subintervalo se calcula $\hat{\gamma}(h)$ tal que $h \in [h_i, h_j]$. En primer lugar se hallan diferentes semivariogramas para distintas direcciones (Figura 8) con el objetivo de explorar sobre la anisotropía o no del proceso subyacente a los datos. En el caso que sea isotrópico calculamos el semivariograma omnidireccional $\hat{\gamma}(h)$ que observamos en la Figura 9.

En la Figura 8 podemos observar que tanto los valores del semivariograma para distancias cercanas a cero como la distancia a partir de la cual dichos valores se estabilizan son aproximadamente similares. No obstante, en la orientación noreste-suroeste (45º), no solo se alcanzan valores máximos menores en comparación con otras direcciones, sino que también se registra una tasa de crecimiento más moderada. Este último detalle sugiere una correlación inferior entre las variables en esta dirección. Este hecho también puede apreciarse en la Figura 1, en la cual se observa que en la dirección noreste-suroeste, los valores son similares, indicando una variabilidad reducida. En contraste, en la dirección sureste-noroeste, puede notarse un aumento de los valores del logaritmo del zinc indicando que la variabilidad espacial explicada en la dirección noreste-suroeste es menor que

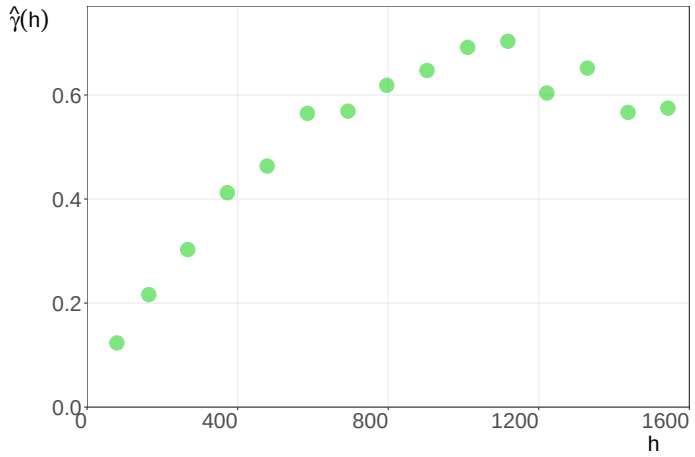



FIGURA 9. Semivariograma empírico omnidireccional para los datos del río Meuse.

en las otras direcciones. Como primer acercamiento al análisis de estos datos y para simplificar el mismo, concluiremos que, dado que tres de los cuatro semivariogramas observados son similares y el cuarto contribuye con poco peso a la variabilidad espacial explicada no consideraremos un comportamiento anisotrópico en el proceso subyacente de los datos. Entonces, en este caso, construimos un único semivariograma omnidireccional el cual puede apreciarse en la Figura 9. Sin embargo, existen herramientas que permitirían hacer un análisis específico de anisotropía que excede este trabajo.

### 5.2. Modelos teóricos de semivariograma

Como puede observarse en la Figura 9, la versión empírica del semivariograma no nos brinda información sobre la correlación espacial de variables definidas en sitios cuyas distancias no hayan sido observadas (i.e. no es continuo en $h$). Por lo tanto, es necesario asumir la existencia de un modelo teórico continuo de semivariograma. En esta dirección, si bien existen métodos de estimación no paramétrica de los mismos (García Soidán, Febrero Bande, y González Manteiga, 2004), los modelos más utilizados son los paramétricos. En este trabajo nos centraremos en estos últimos, y para ello, utilizaremos la notación $\gamma(\mathbf{h}, \boldsymbol{\theta})$ la cual hace hincapié en que el semivariograma no es más que una función paramétrica con una expresión analítica simple que depende del vector de separación $\mathbf{h}$ y de un vector de parámetros $\boldsymbol{\theta}$.

Para ser considerada una función de semivariograma válida de un proceso intrínsecamente estacionario, la función $\gamma(\mathbf{h}, \boldsymbol{\theta})$, debe satisfacer las siguientes propiedades. Para cada $\boldsymbol{\theta} \in \Theta$, donde $\Theta$ es el espacio de parámetros, $\gamma(\mathbf{h}, \boldsymbol{\theta})$ debe satisfacer:

1) $\gamma(\mathbf{0}, \boldsymbol{\theta}) = 0$;

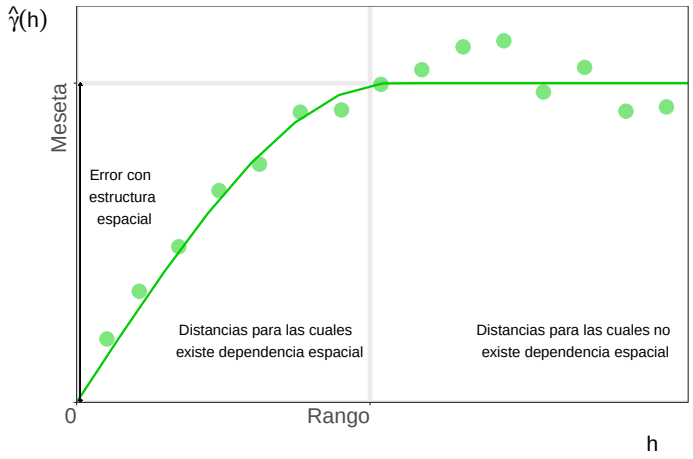



FIGURA 10. Semivariograma empírico y semivariograma estimado, a partir de un modelo esférico para los datos del río Meuse.

II) $\gamma(\mathbf{h}, \boldsymbol{\theta}) \geq 0$;
III) $\gamma(-\mathbf{h}, \boldsymbol{\theta}) = \gamma(\mathbf{h}, \boldsymbol{\theta})$;
IV) $\gamma$ es condicionalmente definido negativo. Esto es, para cualquier vector $\mathbf{a}^T = (a_1, \ldots, a_n),\ a_i \in \mathbb{R}$ que verifique $\mathbf{a}^T\mathbf{1} = 0$, se tiene que $\mathbf{a}^T\boldsymbol{\Gamma}\mathbf{a} \leq 0$, donde $\boldsymbol{\Gamma} \in \mathbb{R}^{n\times n}$ es la matriz semivariograma entre localizaciones tal que $\Gamma_{i,j} \doteq \gamma(\mathbf{s}_i - \mathbf{s}_j, \boldsymbol{\theta})$ para $i, j = 1, \ldots, n$,

ya que estas son las propiedades que cumple el semivariograma teórico de un proceso intrínsecamente estacionario. Las tres primeras propiedades se prueban fácilmente mediante la definición de $\gamma$. En el Apéndice se puede encontrar el detalle de la prueba de la Propiedad IV.

Si bien existe una gran variedad de modelos que satisfacen los requisitos de validez antes mencionados, en la práctica es común utilizar los modelos isotrópicos ya que simplifican el análisis. Recordemos que en este caso el semivariograma empírico solo depende de la distancia $h = \|\mathbf{h}\|$ y no de la dirección por lo que solo es necesario ajustar un único semivariograma teórico el cual denotaremos $\gamma^* : \mathbb{R} \to \mathbb{R}$.

Una característica que tienen los datos espaciales es que, generalmente, la dependencia decrece con la distancia, por lo que el covariograma es una función decreciente de la misma. Por lo tanto, a partir de la Ecuación (4.5), tenemos que el semivariograma es una función creciente respecto a $\mathbf{h}$, y es lo que se tiene en cuenta a la hora de proponer una función de semivariograma válida. Por ejemplo, para los datos de río Meuse, en la Figura 10 hemos graficado el semivariograma empírico y su ajuste paramétrico a partir de un modelo esférico para $\boldsymbol{\theta} = (\theta_1, \theta_2)$

con $\theta_1 > 0$, $\theta_2 > 0$, dado por.

$$\gamma^*(h, \boldsymbol{\theta}) = \begin{cases} \theta_1 \left[ \frac{3h}{2\theta_2} - \frac{1}{2} \left( \frac{h}{\theta_2} \right)^3 \right], & \text{si } 0 < h \leq \theta_2, \\ \theta_1, & \text{si } h > \theta_2. \end{cases} \tag{5.6}$$

En dicho gráfico ilustramos los parámetros asociados al comportamiento del semivariograma a grandes distancias, estos son la meseta y el rango , los cuales definiremos a continuación, y en este caso coinciden con $\theta_1$ y $\theta_2$. La tasa de crecimiento del semivariograma refleja el grado de disimilaridad entre valores cada vez más distantes, la misma puede mantenerse si la variabilidad del fenómeno no tiene límite a grandes distancias o puede tender a desaparecer a partir de cierto valor de $h$ en cuyo caso el semivariograma se estabiliza en lo que llamamos meseta. El valor a partir del cual el semivariograma se estabiliza es el rango y ese valor constante que alcanza o se aproxima el semivariograma es la meseta. Formalmente se llama **meseta** de $\gamma^*(h, \boldsymbol{\theta})$ al límite

$$\lim_{h \to \infty} \gamma^*(h, \boldsymbol{\theta}),$$

siempre que el mismo exista. Los semivariogramas de procesos estacionarios de segundo orden alcanzan este límite en los que permanecen a partir de un cierto $h$ si la covarianza tiende a cero cuando $h$ tiende a infinito. Esto es, si

$$\lim_{h \to \infty} C(h) = 0.$$

Por lo tanto, y a partir de la Ecuación (4.5), tenemos que

$$\lim_{h \to \infty} \gamma^*(h, \boldsymbol{\theta}) = C(\mathbf{0}),$$

por lo que a priori y en muchos casos la meseta está determinada por la varianza del proceso. El caso en el que el semivariograma empírico sea no acotado puede ser un indicio de la invalidez del supuesto de estacionariedad débil como ya lo hemos mencionado anteriormente.

Si la meseta se alcanza, como es el caso de la Figura 10, entonces el **rango** de $\gamma^*(h, \boldsymbol{\theta})$ es el valor más pequeño de $h$ para el cual $\gamma^*(h, \boldsymbol{\theta})$ es igual a su meseta. Formalmente,

$$\text{mín} \left\{ h_0 : \gamma^*(h_0, \boldsymbol{\theta}) = \lim_{h \to \infty} \gamma^*(h, \boldsymbol{\theta}) \right\}.$$

En este contexto, el modelo esférico (5.6) tiene por definición un rango igual a $\theta_2$. Si la meseta no se alcanza, como es el caso de la Figura 11 para los datos del río Meuse, se dice que el rango no existe (en el sentido estricto). No obstante, cuando ello ocurre, se tiene una medida relacionada, llamada **rango efectivo**, que se define como el mínimo valor de $h$ para el cual $\gamma^*(h, \boldsymbol{\theta})$ es igual al $95\,\%$ de la meseta. Puede

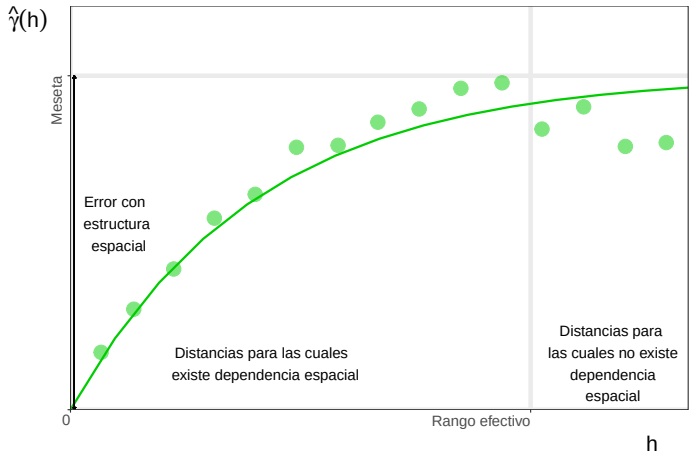



FIGURA 11. Semivariograma empírico y semivariograma estimado a partir de un modelo exponencial para los datos del río Meuse.

probarse que el modelo exponencial (Figura 11) dado por

$$\gamma^*(h, \boldsymbol{\theta}) = \theta_1 \left[1 - \exp\left(-\frac{h}{\theta_2}\right)\right],$$

donde $\boldsymbol{\theta} = (\theta_1, \theta_2)$ con $\theta_1 > 0$, $\theta_2 > 0$, tiene rango efectivo de aproximadamente 3 veces el rango $\theta_2$, esto es, $3\theta_2$. En efecto, si igualamos $\gamma^*(h, \boldsymbol{\theta})$ al $95\,\%$ de la meseta tenemos que,

$$\theta_1 \left[1 - \exp\left(-\frac{h}{\theta_2}\right)\right] = 0{,}95\theta_1,$$

de donde se sigue que,

$$h = -\ln(0{,}05)\theta_2 \approx 3\theta_2.$$

Con respecto al comportamiento del semivariograma cerca del origen, como mencionamos en la Propiedad I) una condición para que $\gamma^*(h, \boldsymbol{\theta})$ sea una función de semivariograma válida es que

$$\gamma^*(0, \boldsymbol{\theta}) = 0.$$

En la práctica esto muchas veces no sucede como por ejemplo en el semivariograma dado en la Figura 9, donde los valores del mismo cuando $h$ se acerca a 0 parecen acercarse a algún valor estrictamente positivo. Por esto, y a diferencia de la Figura 10, el ajuste que observamos en la Figura 12 es diferente ya que consideramos para el mismo combinar el modelo que elegimos para la estructura de covarianza de estos datos, por ejemplo el modelo esférico con el modelo pepita puro, dado por

$$\gamma^*(h, \theta_3) = \begin{cases} \theta_3, & \text{si } h \neq 0, \\ 0, & \text{si } h = 0, \end{cases} \tag{5.7}$$

donde $\theta_3 > 0$.

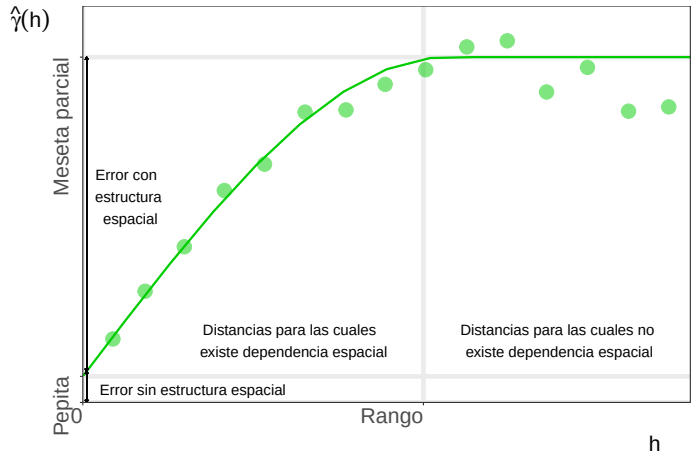



FIGURA 12. Semivariograma empírico y semivariograma estimado, a partir de una combinación entre el modelo pepita puro y un modelo esférico, para los datos del río Meuse.

Podemos formalizar esta decisión de combinar modelos a partir de la descomposición del error dada en (5.2), esto es

$$\begin{aligned}
\gamma_e^*(h,\boldsymbol{\theta}) &= \frac{1}{2}Var\left[e(\mathbf{s}+h)-e(\mathbf{s})\right] \\
&= \frac{1}{2}Var\left[\eta(\mathbf{s}+h)+\epsilon(\mathbf{s}+h)-\eta(\mathbf{s})-\epsilon(\mathbf{s})\right] \\
&= \frac{1}{2}Var\left[\eta(\mathbf{s}+h)-\eta(\mathbf{s})\right]+\frac{1}{2}Var\left[\epsilon(\mathbf{s}+h)-\epsilon(\mathbf{s})\right] \quad (\eta \text{ y } \epsilon \text{ independientes}) \\
&= \gamma_\eta^*(h,\boldsymbol{\theta})+\gamma_\epsilon^*(h,\boldsymbol{\theta}), \qquad (5.8)
\end{aligned}$$

donde $\gamma_\eta^*(h,\boldsymbol{\theta})$ es el semivariograma del proceso que captura la variabilidad espacial y $\gamma_\epsilon^*(h,\boldsymbol{\theta})$ es el semivariograma del proceso asociado al error de medición, el cual es el que hemos llamado **efecto pepita**. Si en el semivariograma empírico observamos que los valores del mismo no se acercan a cero cuando $h$ sí lo hace y además el comportamiento cercano a cero es aproximadamente lineal, una buena opción para el ajuste puede ser la combinación de modelos en términos de (5.8) donde $\gamma_\epsilon^*(h,\theta_3)$ está generalmente dado por el modelo pepita puro (5.7), quien modela la variablilidad del proceso $\epsilon(\mathbf{s})$ el cual suponemos que no cuenta con variabilidad espacial. Por otro lado, $\gamma_\eta^*(h,\boldsymbol{\theta})$ puede ser modelado a partir de cualquier función válida de variograma, por ejemplo, un modelo esférico (5.6). Luego a partir de (5.7) y (5.6), (5.8), el modelo para el semivariograma del residuo está dado por

$$\gamma_e^*(h,\boldsymbol{\theta}) = \begin{cases} 0, & \text{si } h=0, \\ \theta_3+\theta_1\left[\frac{3h}{2\theta_2}-\frac{1}{2}\left(\frac{h}{\theta_2}\right)^3\right], & \text{si } 0<h\leq\theta_2, \\ \theta_3+\theta_1, & \text{si } h>\theta_2. \end{cases}$$

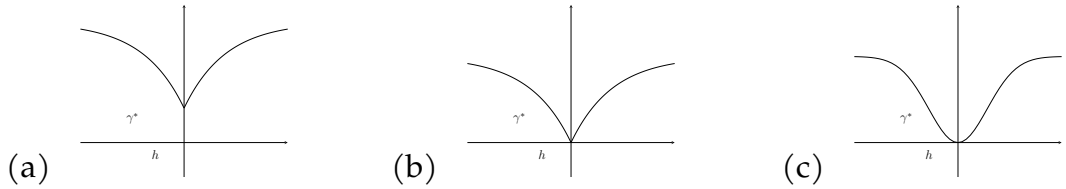



FIGURA 13. Comportamiento del semivariograma cerca al origen: (a) intersección positiva (pepita); (b) comportamiento lineal continuo, no diferenciable; (c) comportamiento parabólico, continuo y diferenciable.

donde $\theta_3$ es el efecto pepita, que no es más que el parámetro meseta en el modelo (5.7), y $\theta_1$ la meseta del modelo esférico, llamada **meseta parcial**, con lo cual la meseta del modelo $\gamma_e^*(h, \boldsymbol{\theta})$ es $\theta_3 + \theta_1$, esto es, la meseta se divide en meseta parcial y efecto pepita, lo cual podemos observar en la Figura 12. Por ello y si elegimos ajustar un semivariograma teórico que considere un efecto pepita, tendremos que estimar tres parámetros: la meseta parcial y el rango dados por $\theta_1$ y $\theta_2$, respectivamente, y el efecto pepita que parametrizamos como $\theta_3$.

Para elegir el mejor modelo que se ajuste a los datos es importante analizar el comportamiento del semivariograma cerca del origen, y particularmente en el origen, ya que esto nos brinda información sobre intervalos de distancias donde la dependencia espacial es grande. Dicho comportamiento es analizado a través de qué tan suave es la superficie generada por el proceso subyacente, la cual clásicamente se caracteriza a partir de la continuidad y diferenciabilidad del mismo. Por ejemplo, si el comportamiento de $\gamma^*(h, \boldsymbol{\theta})$ cerca de cero es lineal, estamos asumiendo una variable regionalizada menos suave que en el caso de considerar un comportamiento parabólico para las distancias cercanas a 0. Estos comportamientos se observan en las Figuras 13 (b) y (c). Si consideramos que el semivariograma tiene un efecto pepita, tendremos una discontinuidad en el origen como la observada en la Figura 13 (a). En este caso asumimos que la variable regionalizada presenta irregularidades para distancias pequeñas.

Un tipo de continuidad muy utilizada es la **continuidad en media cuadrática** que se define como

$$\mathbb{E}\left[(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s}))^2\right] \to 0, \qquad \text{cuando } \|\mathbf{h}\| \to 0,$$

en cuyo caso diremos que el proceso $Z(\mathbf{s})$ es continuo en media cuadrática. A partir de esta definición podemos afirmar que si el semivariograma $2\gamma(\cdot)$ es continuo en el origen entonces el proceso $Z(\mathbf{s})$ es continuo en media cuadrática, ya que, por (4.10), $\mathbb{E}\left[(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s}))^2\right] \to 0$ si y sólo si $2\gamma(\mathbf{h}) \to 0$ cuando $\|\mathbf{h}\| \to 0$. Los modelos más utilizados están definidos de manera tal que $\gamma^*(h, \boldsymbol{\theta}) \to 0$ cuando $h \to 0$. Además, por la Propiedad I) $\gamma^*(0, \boldsymbol{\theta}) = 0$, por lo que son funciones continuas. Esto

significa que el proceso subyacente asociado a dichas funciones de semivariograma genera superficies suaves.

**5.3. Semivariograma estimado** Por último es necesario determinar cómo estimamos los parámetros del semivariograma (pepita, meseta y rango) para lograr el mejor ajuste al semivariograma empírico. Como lo afirman los autores Webster y Oliver (2007, pág. 101), elegir modelos y ajustarlos a los valores del semivariograma empírico se encuentra entre los temas más controvertidos en geoestadística. Este trabajo se puede tornar difícil ya que la precisión de cada estimación varía debido a las diferencias de tamaños de muestra; la variabilidad puede no ser la misma en todas las direcciones; la gráfica del semivariograma empírico puede mostrar mucha fluctuación punto a punto; la mayoría de los modelos no son lineales en uno o más parámetros y se deben usar métodos iterativos para estimar los mismos.

A partir de todos estos problemas es que se sugiere (Montero y cols., 2015; Wackernagel, 2006; Webster y Oliver, 2007) dos formas de ajustar el modelo: la inspección y ajuste manual del mismo y los correspondientes métodos estadísticos. La forma manual consiste en utilizar métodos gráficos, los valores del semivariograma empírico, el conocimiento del fenómeno ya que la forma analítica específica del semivariograma no importa tanto como el hecho de que respete las características principales del fenómeno. El valor del efecto de pepita se puede obtener a partir de los primeros valores del semivariograma empírico extrapolando hasta que corten el eje de las ordenadas. Este valor, el cual representa el comportamiento cerca del origen (para distancias inferiores a la distancia mínima en la muestra), es otra decisión central en el ajuste manual porque tiene una gran influencia en los resultados de la predicción, por ejemplo. La pendiente del semivariograma en el origen puede derivarse también de los primeros valores del semivariograma empírico. En cuanto al comportamiento para distancias grandes, representado por el rango (o rango práctico en caso de tener una meseta asintótica) y la meseta, generalmente se detectan fácilmente, especialmente en el caso estacionario de segundo orden: la distancia más allá de la cual el semivariograma se estabiliza y el valor del semivariograma cuando se estabiliza, respectivamente. Estas cuestiones traen consigo el problema de la subjetividad por parte del investigador. La forma no manual tiene que ver con la utilización de métodos estadísticos conocidos:

- **Métodos de mínimos cuadrados**: ordinarios (OLS, Ordinary Least Square), generalizados (GLS, Generalized Least Square) y pesados (WLS, Weighted Least Square).
- **Métodos basados en la máxima verosimilitud**: incluyen los métodos de máxima verosimilitud (ML, Maximum Likelihood) y los de máxima verosimilitud restringida (REML, Restricted Maximum Likelihood).

Si el objetivo es encontrar el modelo que mejor se ajuste al semivariograma empírico para cada variable, y no tenemos información a priori del fenómeno, se utilizan dichos métodos estadísticos. Si contamos con información a priori del comportamiento de nuestra variable de interés, puede ser interesante realizar un ajuste manual de los modelos al semivariograma empírico. Gallardo (2006) sostiene que si el objetivo del trabajo es comparar los cambios en el semivariograma según los parámetros, la utilización de modelos diferentes resulta poco útil. Hay que tener en cuenta que, por ejemplo, los rangos del modelo esférico y el exponencial no son directamente comparables.

El modelo esférico es el más utlizado en geoestadística, ya que alcanza la meseta. Si observamos las Figuras 11, 10 y 12, donde hemos ajustado un modelo exponencial, esférico y esférico con efecto pepita, respectivamente, para los datos Meuse, vemos que el que mejor se ajusta es el esférico con efecto pepita. Considerando dicho ajuste, podemos concluir que a partir de 900 metros (valor del rango) no existe correlación espacial en los datos. Por otro lado, podemos afirmar que $\hat{\sigma}^2_\epsilon = 0{,}05$ que es la estimación de la varianza del efecto pepita indicando que posiblemente exista un error de medición en los datos.

## §6. Conclusiones

En este trabajo definimos el marco teórico subyacente a los datos espacialmente correlacionados, comenzando por el modelo considerado para el proceso estocástico que da lugar a los mismos e imponiendo diferentes funciones que modelan su estructura de covarianza. Definimos además herramientas exploratorias las cuales sirven como punto de partida para el análisis de variables con dependencia espacial, que permiten visualizar tendencia, variabilidad y asociación espacial de los datos. Todo lo expuesto se ejemplificó con los clásicos datos del río Meuse, los cuales contienen, entre otras variables, la concentración de zinc en la capa superior del suelo en una llanura aluvial del río Meuse, cerca del pueblo de Stein (Países Bajos), medida en diferentes coordenadas geográficas.

## §7. Agradecimientos

Este trabajo fue realizado en el marco del Proyecto PICT-2019-00301 financiado por la ANPCYT.

## §8. Apéndice: Resultados auxiliares y demostraciones

Para una lectura más autocontenida del trabajo, en este apéndice incluiremos resultados auxiliares y demostraciones del mismo.

**Lema 8.1.** *El proceso estocástico discreto de Wiener-Levy* $\{Z_k\}_{k\geq 0}$ *con* $k \in \mathbb{N}$*, cumple que*

$$Z_{k+1} = Z_k + \epsilon_k, \tag{8.1}$$

*donde* $\epsilon_k$ *tiene distribución normal* $N(0,1)$ *y son independientes para todo* $k$ *y* $Z_0 = 0$*. Dicho proceso es intrínsecamente estacionario pero no débilmente estacionario.*

*Demostración.* Comencemos probando que el proceso (8.1) satisface las condiciones de estacionariedad intrínseca (4.8) y (4.9). En primer lugar tenemos que

$$\begin{aligned} Z_{k+h} &= Z_{k+h-1} + \epsilon_{k+h-1} \\ &= Z_{k+h-2} + \epsilon_{k+h-2} + \epsilon_{k+h-1} \\ &\vdots \\ &= Z_k + \epsilon_k + \epsilon_{k+1} + \cdots + \epsilon_{k+h-1}. \end{aligned} \tag{8.2}$$

Luego

$$\mathbb{E}\left[Z_{k+h} - Z_k\right] = \mathbb{E}\left[\sum_{i=0}^{h-1} \epsilon_{k+i}\right] = 0,$$

$$\text{Var}\left(Z_{k+h} - Z_k\right) = \text{Var}\left(\sum_{i=0}^{h-1} \epsilon_{k+i}\right) = h. \tag{8.3}$$

Sin embargo, para este proceso, la covarianza no depende solo de $h$ pues,

$$\begin{aligned} \text{Var}\left(Z_{k+h}\right) &= \text{Var}\left(Z_k + \epsilon_k + \epsilon_{k+1} + \cdots + \epsilon_{k+h-1}\right) && \text{de (8.2)} \\ &= \text{Var}\left(Z_k\right) + \text{Var}\left(\sum_{i=0}^{h-1} \epsilon_{k+i}\right) && \text{por independencia} \\ &= \text{Var}\left(Z_k\right) + h. && \text{de (8.3)} \end{aligned} \tag{8.4}$$

Despejando la covarianza de la siguiente ecuación,

$$\text{Var}\left(Z_{k+h} - Z_k\right) = \text{Var}\left(Z_{k+h}\right) + \text{Var}\left(Z_k\right) - 2\text{Cov}\left(Z_{k+h}, Z_k\right),$$

y utilizando (8.4) y (8.3) se tiene que,

$$\begin{aligned} \text{Cov}\left(Z_{k+h}, Z_k\right) &= \frac{1}{2}\Big(\text{Var}\left(Z_{k+h}\right) + \text{Var}\left(Z_k\right) - \text{Var}\left(Z_{k+h} - Z_k\right)\Big) \\ &= \frac{1}{2}\Big(\text{Var}\left(Z_k\right) + h + \text{Var}\left(Z_k\right) - h\Big) \\ &= \text{Var}\left(Z_k\right), \end{aligned}$$

de donde se sigue que la covarianza no depende solo de $h$ y por lo tanto no puede ser un proceso débilmente estacionario. □

**Lema 8.2.** *Si el proceso $e(\boldsymbol{s})$ es estacionario de segundo orden y $\mu(\boldsymbol{s}) \doteq \mu$ es constante entonces el proceso $Z(\boldsymbol{s})$ también es estacionario de segundo orden.*

*Demostración.* A partir del modelo (5.1) tenemos que

$$\mathbb{E}(Z(\mathbf{s})) = \mu + \mathbb{E}(e(\mathbf{s})) = \mu,$$

y

$$\begin{aligned}\text{Cov}(Z(\mathbf{s}+\mathbf{h}), Z(\mathbf{s})) &= \text{Cov}(\mu(\mathbf{s}+\mathbf{h}) + e(\mathbf{s}+\mathbf{h}), \mu(\mathbf{s}) + e(\mathbf{s})) \\ &= \text{Cov}(e(\mathbf{s}+\mathbf{h}), e(\mathbf{s})) \\ &= C(\mathbf{h}),\end{aligned}$$

luego $Z(\mathbf{s})$ es estacionario de segundo orden. □

**Lema 8.3.** *Si $e(\boldsymbol{s})$ es intrínsecamente estacionario y $\mu(\boldsymbol{s}) \doteq \mu$ constante entonces el proceso $Z(\boldsymbol{s})$ también es intrínsecamente estacionario.*

*Demostración.* A partir del mismo razonamiento que en el Lema anterior, tenemos que

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

y

$$\begin{aligned}\text{Var}\left[Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})\right] &= \text{Var}([\mu(\mathbf{s}+\mathbf{h}) + e(\mathbf{s}+\mathbf{h}) - \mu(\mathbf{s}) - e(\mathbf{s})]) \\ &= \text{Var}([e(\mathbf{s}+\mathbf{h}) - e(\mathbf{s})]) = 2\gamma_e(\mathbf{h}),\end{aligned}$$

con lo cual se prueba que $Z(\mathbf{s})$ es intrínsecamente estacionario. □

*Demostración del Lema 5.1.* Partiendo de la Ecuación (5.3) tenemos que,

$$\begin{aligned}\mathbb{E}[\hat{\gamma}(\mathbf{h})] &= \mathbb{E}\left[\frac{1}{2N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2\right] \\ &= \frac{1}{N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}\frac{1}{2}\mathbb{E}\left[(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2\right] \\ &= \frac{1}{N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}\frac{1}{2}\text{Var}\left[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\right] \\ &= \frac{1}{N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}\gamma(\mathbf{h}) \\ &= \gamma(\mathbf{h}).\end{aligned}$$

Por otro lado, para el caso del covariograma consideremos $\tilde{C}(\mathbf{h}) = \frac{N(\mathbf{h})-1}{N(\mathbf{h})}\hat{C}(\mathbf{h})$ y tenemos que

$$\begin{aligned}
\mathbb{E}(\tilde{C}(\mathbf{h})) &= \mathbb{E}\left[\frac{1}{N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}(Z(\mathbf{s}_i+\mathbf{h})-\bar{Z}_{\mathbf{h}})\left(Z(\mathbf{s}_i)-\bar{Z}_{(\mathbf{h})}\right)\right] \\
&= \mathbb{E}\left[\frac{1}{N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}(Z(\mathbf{s}_i+\mathbf{h})Z(\mathbf{s}_i)-Z(\mathbf{s}_i+\mathbf{h})\bar{Z}_{(\mathbf{h})}-Z(\mathbf{s}_i)\bar{Z}_{\mathbf{h}}+\bar{Z}_{\mathbf{h}}\bar{Z}_{(\mathbf{h})})\right] \\
&= \frac{1}{N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}\mathbb{E}\left[Z(\mathbf{s}_i+\mathbf{h})Z(\mathbf{s}_i)\right]-\mathbb{E}\left[\bar{Z}_{(\mathbf{h})}\frac{1}{N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}Z(\mathbf{s}_i+\mathbf{h})\right] \\
&\quad -\mathbb{E}\left[\bar{Z}_{\mathbf{h}}\frac{1}{N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}Z(\mathbf{s}_i)\right]+\mathbb{E}\left[\bar{Z}_{\mathbf{h}}\bar{Z}_{(\mathbf{h})}\right] \qquad (8.5) \\
&= \frac{1}{N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}\mathbb{E}\left[Z(\mathbf{s}_i+\mathbf{h})Z(\mathbf{s}_i)\right]-\mathbb{E}\left[\bar{Z}_{\mathbf{h}}\bar{Z}_{(\mathbf{h})}\right]. \qquad \text{(por (5.5))}
\end{aligned}$$

Calculamos la esperanza $\mathbb{E}\left[\bar{Z}_{\mathbf{h}}\bar{Z}_{(\mathbf{h})}\right]$ considerando el proceso $Z(\mathbf{s})$ estacionario de segundo orden, esto es, $\mathbb{E}[Z(\mathbf{s})] = \mu$ y $\text{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_i+\mathbf{h})) = C(\mathbf{h})$,

$$\begin{aligned}
\mathbb{E}(\bar{Z}_{\mathbf{h}}\bar{Z}_{(\mathbf{h})}) &= \text{Cov}(\bar{Z}_{\mathbf{h}}, \bar{Z}_{(\mathbf{h})}) + \mathbb{E}(\bar{Z}_{\mathbf{h}})\mathbb{E}(\bar{Z}_{(\mathbf{h})}) \\
&= \frac{1}{N(\mathbf{h})^2}\sum_{i=1}^{N(\mathbf{h})}\sum_{j=1}^{N(\mathbf{h})}\text{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j+\mathbf{h})) + \mu^2 \\
&= \frac{1}{N(\mathbf{h})^2}\sum_{i=1}^{N(\mathbf{h})}\text{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_i+\mathbf{h})) \\
&\quad + \frac{1}{N(\mathbf{h})^2}\sum_{i=1}^{N(\mathbf{h})}\sum_{j\neq i}^{N(\mathbf{h})}\text{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j+\mathbf{h})) + \mu^2 \\
&= \frac{1}{N(\mathbf{h})}C(\mathbf{h}) + \frac{1}{N(\mathbf{h})^2}\sum_{i=1}^{N(\mathbf{h})}\sum_{j\neq i}^{N(\mathbf{h})}\text{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j+\mathbf{h})) + \mu^2. \qquad (8.6)
\end{aligned}$$

Luego reemplazamos (8.6) en (8.5) y, teniendo en cuenta que

$$\mathbb{E}\left[Z(\mathbf{s}_i+\mathbf{h})Z(\mathbf{s}_i)\right] = C(\mathbf{h}) + \mu^2,$$

obtenemos

$$\begin{aligned}\mathbb{E}(\tilde{C}(\mathbf{h})) &= \frac{1}{N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}\mathbb{E}\left[Z(\mathbf{s}_i+\mathbf{h})Z(\mathbf{s}_i)\right] - \mathbb{E}\left[\bar{Z}_{\mathbf{h}}\bar{Z}_{(\mathbf{h})}\right] \\ &= C(\mathbf{h}) + \mu^2 - \left(\frac{1}{N(\mathbf{h})}C(\mathbf{h}) + \frac{1}{N(\mathbf{h})^2}\sum_{i=1}^{N(\mathbf{h})}\sum_{j\neq i}^{N(\mathbf{h})}\mathrm{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j+\mathbf{h})) + \mu^2\right) \\ &= \left(1 - \frac{1}{N(\mathbf{h})}\right)C(\mathbf{h}) - \frac{1}{N(\mathbf{h})^2}\sum_{i=1}^{N(\mathbf{h})}\sum_{j\neq i}^{N(\mathbf{h})}\mathrm{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j+\mathbf{h})).\end{aligned}$$

Por último usando la definición de $\tilde{C}(\mathbf{h})$ obtenemos que

$$\begin{aligned}\mathbb{E}(\hat{C}(\mathbf{h})) &= \frac{N(\mathbf{h})}{N(\mathbf{h})-1}\mathbb{E}(\tilde{C}(\mathbf{h})) \\ &= C(\mathbf{h}) - \frac{1}{N(\mathbf{h})(N(\mathbf{h})-1)}\sum_{i=1}^{N(\mathbf{h})}\sum_{j\neq i}^{N(\mathbf{h})}\mathrm{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j+\mathbf{h})). \qquad (8.7)\end{aligned}$$

Con lo cual podemos observar que el estimador $\hat{C}(\mathbf{h})$ es sesgado. Solo en el caso particular de que las covarianzas sean nulas (datos independientes) o la suma de las mismas de cero por covarianzas positivas y negativas, el estimador no sería sesgado. Sin embargo, para el caso más clásico de dependencias espaciales no negativas, todas las covarianzas en el segundo término de (8.7) deben ser positivas o cero, con lo cual es sesgo es negativo (Smith, 2014). □

*Demostración de la Propiedad iv.* A partir de la condición de estacionariedad intrínseca (4.10) y que la sumatoria de los $a_i$ es cero, resulta que

$$\begin{aligned}\mathbf{a}^T\boldsymbol{\Gamma}\mathbf{a} = \sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_j\gamma(\mathbf{s}_i-\mathbf{s}_j) &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_j\mathbb{E}[(Z(\mathbf{s}_i)-Z(\mathbf{s}_j))^2] && \text{(de (4.10))} \\ &= \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{n}\left(\sum_{j=1}^{n}a_j\right)a_i(Z(\mathbf{s}_i))^2\right] \\ &\quad - \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_jZ(\mathbf{s}_i)Z(\mathbf{s}_j)\right] \\ &\quad + \frac{1}{2}\mathbb{E}\left[\sum_{j=1}^{n}\left(\sum_{i=1}^{n}a_i\right)a_j(Z(\mathbf{s}_j))^2\right] \\ &= -\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_jZ(\mathbf{s}_i)Z(\mathbf{s}_j)\right] && \left(\text{pues } \sum_{i=1}^{n}a_i = \sum_{j=1}^{n}a_j = 0\right) \\ &= -\mathbb{E}\left[\left(\sum_{i=1}^{n}a_iZ(\mathbf{s}_i)\right)^2\right] \leq 0.\end{aligned}$$

Lo cual prueba que $\gamma$ es condicionalmente definido negativo. □

## Bibliografía

Atkinson, P. M., y Lloyd, C. D. (2014). Geostatistical models and spatial interpolation. En M. M. Fischer y P. Nijkamp (Eds.), *Handbook of regional science* (pp. 1461–1476). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-642-23430-9_75

Banerjee, S., Carlin, B. P., y Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC.

Berke, O. (2004). Exploratory disease mapping: Kriging the spatial risk function from regional count data. *International journal of health geographics*, *3*, 18. doi: 10.1186/1476-072X-3-18

Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V., y Pebesma, E. J. (2008). *Applied spatial data analysis with r*. Springer.

Chica-Olmo, J., y Cano-Guervos, R. (2020). Does my house have a premium or discount in relation to my neighbors? a regression-kriging approach. *Socio-Economic Planning Sciences*, *72*, 100914. doi: 10.1016/j.seps.2020.100914

Chiles, J.-P., y Delfiner, P. (2012). *Geostatistics: modeling spatial uncertainty*. John Wiley & Sons.

Cressie, N. (1993). *Statistics for spatial data*. John Wiley and Sons, Inc.

Derdouri, A., y Murayama, Y. (2020, 05). A comparative study of land price estimation and mapping using regression kriging and machine learning algorithms across fukushima prefecture, japan. *Journal of Geographical Sciences*, *30*, 794-822. doi: 10.1007/s11442-020-1756-1

Dubin, R. A. (1992). Spatial autocorrelation and neighborhood quality. *Regional Science and Urban Economics*, *22*(3), 433-452. doi: https://doi.org/10.1016/0166-0462(92)90038-3

Fernández-Avilés, G. (2009). Spatial regression analysis vs. kriging methods for spatial estimation. *International Advances in Economic Research*, *15*, 44-58. doi: 10.1007/s11294-008-9189-0

Fischer, M. M., y Wang, J. (2011). *Spatial data analysis: models, methods and techniques*. Springer Science & Business Media.

Gallardo, A. (2006). Geostadística. *Ecosistemas*, *3*.

Gámez, M., Montero, J., y Rubio, N. (2000). Kriging methodology for regional economic analysis: Estimating the housing price in albacete. *International Advances in Economic Research*, *6*, 438-450. doi: 10.1007/BF02294963

García Arancibia, R., Llop, P., y Lovatto, M. (2023). Nonparametric prediction for univariate spatial data: Methods and applications. *pre-print*, 1-45.

García Soidán, P., Febrero Bande, M., y González Manteiga, W. (2004, 03). Non-parametric kernel estimation of an isotropic variogram. *Journal of Statistical Planning and Inference*, *121*, 65-92. doi: 10.1016/S0378-3758(02)00507-4

Goovaerts, P. (2008). Geostatistical analysis of health data: State-of-the-art and perspectives. En A. Soares, M. J. Pereira, y R. Dimitrakopoulos (Eds.), *Geostatistics for environmental applications: Proceedings of the sixth european conference on geostatistics for environmental applications* (pp. 3–22). Dordrecht: Springer Netherlands. doi: 10.1007/978-1-4020-6448-7_1

Haining, R. (2013). Spatial data and statistical methods: A chronological overview. En M. M. Fischer y P. Nijkamp (Eds.), *Handbook of regional science* (pp. 1277–1294). Springer Berlin Heidelberg. doi: 10.1007/978-3-642-23430-9_71

Haining, R. P., Kerry, R., y Oliver, M. A. (2010). Geography, spatial data analysis, and geostatistics: An overview. *Geographical Analysis*, *42*(1), 7-31. doi: 10.1111/j.1538-4632.2009.00780.x

Kerry, R., Goovaerts, P., Haining, R., y Ceccato, V. (2010). Applying geostatistical analysis to crime data: Car-related thefts in the baltic states. *Geographical analysis*, *42*, 53-77. doi: 10.1111/j.1538-4632.2010.00782.x

Longley, P. (2005). *Geographic information systems and science*. John Wiley & Sons.

Montero, J.-M., Fernández-Avilés, G., y Mateu, J. (2015). *Spatial and spatio-temporal geostatistical modeling and kriging*. John Wiley and Sons, Ltd.

Montes-Rojas, G. V. (2012). Optimal spatial prediction and the construction of regional indexes. *The Journal of Economic Asymmetries*, *9*(1), 1-21. doi: https://doi.org/10.1016/j.jeca.2012.01.001

Morales, J., Stein, A., Flacke, J., y Zevenbergen, J. (2020). Predictive land value modelling in guatemala city using a geostatistical approach and space syntax. *International Journal of Geographical Information Science*, *34*, 1-24. doi: 10.1080/13658816.2020.1725014

Rikken, M., y Van Rijn, R. (1993). *Soil pollution with heavy metals - an inquiry into spatial variation, cost of mapping and the risk evaluation of copper, cadmium, lead and zinc in the oodplains of the meuse west of stein*. Utrecht: Department of Physical Geography, Utrecht University.

Siabato, W., y Guszmán-Manrique, J. (2019). La autocorrelación espacial y el desarrollo de la geografía cuantitativa. *Cuadernos de Geografía: Revista Colombiana de Geografía*, *28*(1), 1-22.

Smith, T. (2014). *Notebook on spatial data analysis.* Descargado de https://www.seas.upenn.edu/~tesmith/NOTEBOOK/index.html

Tsutsumi, M., y Seya, H. (2008). Measuring the impact of large-scale transportation projects on land price using spatial statistical models. *Papers in Regional Science*, *87*(3), 385-401. doi: https://doi.org/10.1111/j.1435-5957.2008.00192.x

Valente, J., Wu, S., Gelfand, A., y Sirmans, C. (2005, 02). Apartment rent prediction using spatial modeling. *Journal of Real Estate Research*, *27*, 105-136. doi:

10.1080/10835547.2005.12091148

Vasan, S., y Alcantara, A. (2016). Gis-based methods for estimating missing poverty rates & projecting future rates in census tracts. *Review of Economics & Finance*, *3*, 1-13.

Wackernagel, H. (2003). *Multivariate geostatistics*. Springer.

Wackernagel, H. (2006). Geostatistics. En *Encyclopedia of statistical sciences.* American Cancer Society. Descargado de https://onlinelibrary.wiley.com/doi/abs/10.1002/0471667196.ess5085.pub2 doi: 10.1002/0471667196.ess5085.pub2

Webster, R., y Oliver, M. (2007). *Geostatistics for environmental scientists* (2th ed.). Wiley.

Zhang, J., Atkinson, P., y Goodchild, M. F. (2014). *Scale in spatial information and analysis*. CRC Press.

Zhao, Y., y Wall, M. M. (2004). Investigating the use of the variogram for lattice data. *Journal of Computational and Graphical Statistics*, *13*(3), 719–738.

Rodrigo García Arancibia
*Instituto de Economía Aplicada Litoral (IECAL-FCE)*
*Universidad Nacional del Litoral y CONICET*
(✉) *r.garcia.arancibia@gmail.com*

Pamela Llop
*Facultad de Ingeniería Química*
*Universidad Nacional del Litoral y CONICET*
(✉) *lloppamela@gmail.com*

Mariel Lovatto
*Facultad de Ingeniería Química*
*Universidad Nacional del Litoral y CONICET*
(✉) *marielguadalupelovatto@gmail.com*

Recibido: *7 de julio de 2023*.
Aceptado: *6 de diciembre de 2024*.
Publicado en línea: *20 de diciembre de 2024*.