

Parte II

Geoestadística. Análisis estructural

Supondremos que $Z(x)$ es un proceso aleatorio, $x \in D$ denota una localización en un espacio euclídeo d -dimensional, con $d = 1, 2, 3$ y $D \subset \mathbb{R}^d$ es un conjunto fijo (no aleatorio) formado por un intervalo.

Tal como señalamos en la introducción, en Geoestadística $Z(x)$ se supone proceso infinitodimensional de forma que las observaciones muestrales $\{x_1, x_2, \dots, x_i, \dots, x_n\} \subset D$ son una realización discreta del proceso.

$Z(x_1), \dots, Z(x_n)$ se consideran realizaciones de una misma variable aleatoria. Es la marginal del proceso aleatorio.

Referencias teóricas importantes para el estudio de este tipo de datos espaciales:

- Bardossy, A. Introduction to Geostatistics. Institute of Hydraulic Engineering. University of Stuttgart.
- Chilès, J.P. y Delfiner, P. (1999). Geostatistics. Modeling spatial uncertainty. Wiley, New York.
- Cressie, N. (1993). Statistics for spatial data. Wiley, New York.
- Henao, R.G. Introducción a la Geoestadística. Accesible en www.reddeagricultura-precision.unal.edu.co/doc/Publicaciones/LIBRO_%20DE%20GEOESTADISTICA-R%20Giraldo.pdf

Bajo la etiqueta de “Bibliografía adicional” de la presente secuencia de aprendizaje podéis encontrar parte de los textos de Bardossy y Henao. He incluido sólo los capítulos que tienen relación con el análisis estructural. Para consultar los otros dos libros tendríais que solicitarlos en la biblioteca.

3. Introducción

3.1. Variable regionalizada

Es un proceso aleatorio (función aleatoria o variable aleatoria georeferenciada) distribuido en el espacio. Representa un fenómeno natural. Lo más habitual es que presente una parte determinista y otra errática.

$$Z(x) = m(x) + Y(x); \quad x \in D \subset \mathbb{R}^d, d = 1, 2, 3,$$

donde $m(x)$ es la componente determinista, no aleatoria, y representa los cambios o evolución a gran escala e $Y(x)$ es la componente aleatoria (errática) y representa el comportamiento local o evolución a pequeña escala.

- Localmente, para un $x \in \mathbb{R}^d$, $Z(x)$ es una variable aleatoria.
- $Z(x)$ y $Z(x+h)$ no son independientes, sino que están correlacionadas por la estructura espacial del fenómeno.
- Sin pérdida de generalidad, se asume que $E[Z(x)] = m(x)$.
- $Cov(Z(x), Z(x+h)) = Cov(Y(x), Y(x+h))$.

3.2. Momentos

Dada la variable regionalizada $Z(x)$ y las localizaciones x_1, x_2, \dots, x_k existe el correspondiente conjunto de variables aleatorias $\{Z(x_1), Z(x_2), \dots, Z(x_k)\}$. Este conjunto está totalmente caracterizada por la función de distribución conjunta:

$$F_{x_1, x_2, \dots, x_k}(z_1, z_2, \dots, z_k) = P(Z(x_1) \leq z_1, \dots, Z(x_k) \leq z_k).$$

Si conocemos F_{x_1, x_2, \dots, x_k} para todo $k \in \mathbb{N}$ y para todo $x_i \in D$, entonces conocemos la “ley espacial” de $Z(x)$.

3.3. Hipótesis de estacionariedad

Esta hipótesis establece el grado de homogeneidad espacial del fenómeno.

Puesto que sólo disponemos de una realización discreta de la variable regionalizada $Z(x)$, es inevitable que haya que asumir ciertas hipótesis para poder llevar a cabo algún tipo de estudio estadístico. El tipo de estacionariedad asumido indica qué tipo de inferencia estadística puede realizarse con el modelo probabilístico.

3.3.1. Estrictamente estacionario

Una función aleatoria es estrictamente estacionaria si y sólo si la ley espacial es invariante a traslaciones, es decir:

$$F_{x_1, x_2, \dots, x_k}(z_1, z_2, \dots, z_k) = F_{x_1+h, x_2+h, \dots, x_k+h}(z_1, z_2, \dots, z_k), \quad \forall h \in \mathbb{R}^d, \forall k \in \mathbb{N}.$$

Trabajando con datos reales, pocas veces podrá comprobarse esta condición. Habitualmente se asumirá alguna de las dos siguientes.

3.3.2. Estacionario de segundo orden

- Existe $E[Z(x)] = m, \forall x \in D$.
- Existe $Cov(Z(x), Z(x+h)) = C(h), \forall x \in D \text{ y } \forall h \in \mathbb{R}^d$.

La función $C(h)$ recibe el nombre de **covariograma**. Como consecuencia de la definición tenemos que:

1. $\sigma_Z^2 = C(0), \forall x \in D$
2. C es una función simétrica: $C(h) = C(-h)$.
3. Aplicando la desigualdad de Cauchy-Schwarz puede demostrarse que

$$|C(h)| \leq C(0)$$

4. $C(h)$ es una función definida positiva.
5. Una función relacionada es el correlograma: $\rho(h) = C(h)/C(0)$, verificando que

$$|\rho(h)| \leq \rho(0) = 1$$

3.3.3. Hipótesis intrínseca

Muchos fenómenos físicos poseen una capacidad infinita de dispersión, lo que quiere decir que no existe la varianza de la correspondiente variable regionalizada.

- Existe $E[Z(x) - Z(x+h)] = m(h)$, $\forall x \in D$ y $\forall h \in \mathbb{R}^d$.
- Existe $\frac{1}{2}Var(Z(x) - Z(x+h)) = \gamma(h)$, $\forall x \in D$ y $\forall h \in \mathbb{R}^d$.

La función $\gamma(h)$ recibe el nombre de **semivariograma**, aunque cada vez es más frecuente referirse a ella como **variograma**.

1. γ es una función simétrica: $\gamma(h) = \gamma(-h)$
2. $|\gamma(h)| \geq \gamma(0) = 0$.
3. Variograma estandarizado: $\gamma_s(h) = \frac{\gamma(h)}{\sigma_Z^2}$, $\gamma_s(0) = 0$.

3.3.4. Observaciones

1. Estrictamente estacionario \implies Estacionario de segundo orden \implies Hipótesis intrínseca. Los recíprocos no se verifican.
2. Si se verifica la condición de estacionario de segundo orden, entonces:
 $\gamma(h) = C(0) - C(h) = \sigma_Z^2 - C(h)$. Como consecuencia, ambas funciones son equivalentes para caracterizar la dependencia espacial de la variable regionalizada $Z(x)$. La figura 1 muestra una representación gráfica en la cual se aprecia la relación entre funciones.
3. Si se verifica la hipótesis intrínseca, entonces:
 - a) $\gamma(h)$ es una función condicionalmente definida negativa.
 - b) $\lim_{|h| \rightarrow +\infty} \frac{\gamma(h)}{|h|^2} = 0$.
4. En principio asumiremos que la tendencia m es constante. El caso general en el cual la tendencia depende del vector h se abordará en el capítulo del kriging universal.

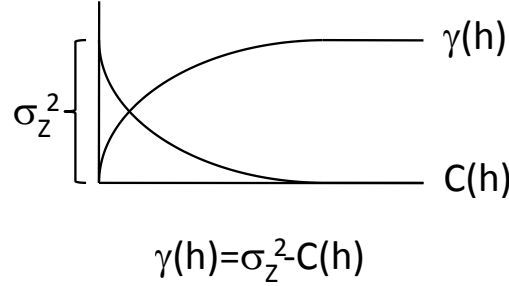


Figura 1: Relación entre semivariograma y covariograma para un proceso estacionario de segundo orden.

4. Estudio del semivariograma

El estudio del semivariograma, o variograma, es más habitual que el estudio del covariograma. Las razones son dos:

- La hipótesis de intrínsecamente estacionaria incluye la condición de estacionaria de segundo orden.
- Para el cálculo del semivariograma no es necesario conocer la media de la variable.

Por otra parte, si la variable es estacionaria de segundo orden, siempre puede obtenerse la función de covariograma a partir de la función de semivariograma: $\gamma(h) = \sigma_Z^2 - C(h)$

4.1. Nube de semivariograma y semivariograma experimental

Dada una variable regionalizada $\{Z(x)/x \in D\}$ de la cual tenemos, en n puntos o localizaciones diferentes, una muestra $\{Z(x_1), Z(x_2), \dots, Z(x_n)\}$ y una realización muestral $\{z(x_1), z(x_2), \dots, z(x_n)\}$ interesa aproximar el semivariograma teórico en una dirección determinada u , $|u| = 1$:

$$\gamma(h) = \frac{1}{2} \text{Var}(Z(x) - Z(x+h)), \quad h = |h|u.$$

Asumiendo que $E[Z(x) - Z(x+h)] = 0$, entonces

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

Para aproximar esta función, para un h fijo, suele utilizarse el semivariograma experimental o muestral basado en el **método de los momentos**:

$$\hat{\gamma}(h) = \frac{1}{2N_h} \sum_{|x_i - x_j| \simeq |h|} (Z(x_i) - Z(x_j))^2.$$

Donde N_h es el número de pares (x_i, x_j) verificando $x_i - x_j = \pm h$.

Si los datos siguen una distribución normal, es bien sabido que el métodos de los momentos coincide con el de máxima verosimilitud. Esto indicaría una buena aproximación de la función $\gamma(h)$. Si los datos no presentan un comportamiento “cercano” a la distribución normal, puede ser aconsejable la utilización de un estimador robusto de la función de semivariograma. La mayoría de estos estimadores están basados en el cálculo de medianas:

Ejemplos de **estimador robusto**:

$$\hat{\gamma}(h) = \frac{1}{2} \frac{\left\{ \frac{1}{N_h} \sum_{N_h} |Z(x_i) - Z(x_j)|^{1/2} \right\}^4}{0,457 + \frac{0,494}{N_h}}$$

$$\hat{\gamma}(h) = \frac{1}{2} \frac{\left\{ \text{mediana} \left[|Z(x_i) - Z(x_j)|^{1/2} / (x_i, x_j) \in N_h \right] \right\}^4}{0,457 + \frac{0,494}{N_h}}$$

El cálculo del semivariograma experimental direccional (en la dirección u) comprende el cálculo de una serie de valores

$$\{\hat{\gamma}(h_1), \dots, \hat{\gamma}(h_J)\},$$

donde $|h_j - h_{j-1}| = \Delta h$ suele ser constante.

En el caso del estimador por el método de los momentos, se denomina nube de semivariograma a la representación de los $\binom{n}{2} = n(n-1)/2$ valores muestrales $(Z(x_i) - Z(x_j))^2$ frente a $|h| = |x_i - x_j|$.

En el caso del estimador robusto, la nube de semivariograma es la representación gráfica de los $n(n-1)/2$ valores $|Z(x_i) - Z(x_j)|^{1/2}$ frente a $|h| = |x_i - x_j|$.

Especialmente en el caso en que $d = 2$ suele ser muy útil la representación gráfica de la denominada superficie de semivariograma. Éste es un gráfico tridimensional, como los que se muestran en las figuras 8 y 9. En estos gráficos se representan, en el plano X, Y , las posibles direcciones u y, en el eje Z , los semivariogramas experimentales en las distintas direcciones u .

4.1.1. Regiones de tolerancia

Es evidente que para el cálculo del semivariograma experimental, $\hat{\gamma}(h)$, no es buena idea restringirse a elegir los pares de localizaciones que verifiquen de forma estricta la condición $x_i - x_j = \pm h$. Esta condición suele relajarse permitiendo regiones de tolerancia.

- Superficie de semivariograma.

Para el cálculo de la superficie de semivariograma se utilizan regiones de tolerancia en forma de segmento ($d = 1$), cuadrado ($d = 2$) o cubo ($d = 3$) tal como se muestra en la figura 2.

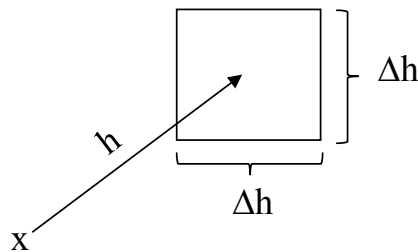


Figura 2: Región de tolerancia para el cálculo de la superficie de semivariograma. Caso $d = 2$.

- Semivariograma direccional.

Son los semivariogramas experimentales calculados en una dirección u determinada

La figura 3 representa la región de tolerancia que se utiliza para calcular un semivariograma experimental en una dirección determinada ($d = 2$). En estos casos deben elegirse un ángulo α y una distancia d_i de tolerancia. Es habitual utilizar $d_i = \Delta h$.

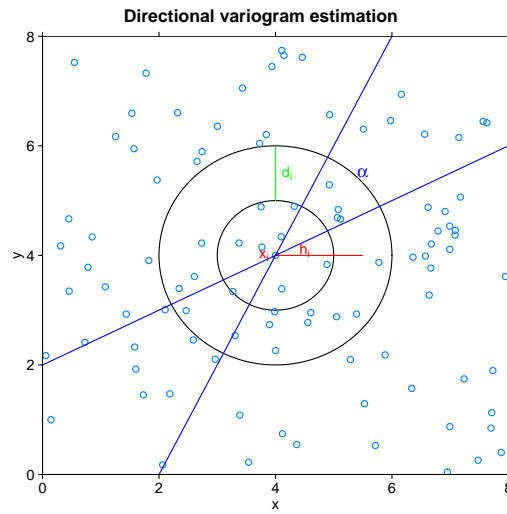


Figura 3: Región de tolerancia para el cálculo de semivariogramas direccionales. Caso $d = 2$.

- Semivariograma omnidireccional.

Son los semivariogramas calculados tomando pares (x_i, x_j) de localizaciones que distan $|x_i - x_j| \simeq |h|$ independientemente de la dirección.

La figura 4 representa la región de tolerancia que se utiliza para calcular un semivariograma experimental omnidireccional ($d = 2$).

Es de destacar la importancia de las regiones de tolerancia puesto que, si se toman muy amplias, “suavizan” la función $\hat{\gamma}(h)$.

Como consecuencia de estas regiones ocurre que, aún fijando un h determinado, se utilicen vectores aproximados $x_i - x_j \simeq h$.

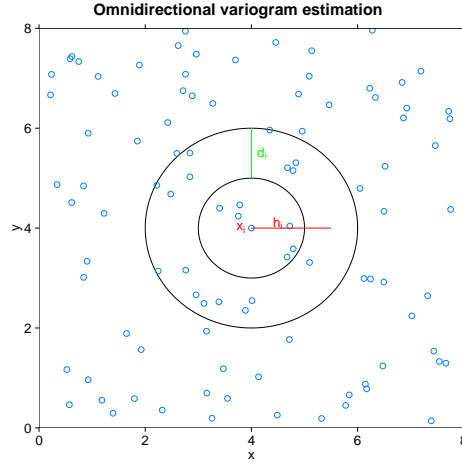


Figura 4: Región de tolerancia para el cálculo de semivariogramas omnidireccionales. Caso $d = 2$.

4.1.2. Partes del semivariograma

- Meseta. Si el semivariograma alcanza un valor límite, es decir $\exists \lim_{|h| \rightarrow +\infty} \gamma(h)$, éste recibe el nombre de meseta. No siempre existe.

Si la variable $Z(x)$ es estacionaria de segundo orden, la meseta coincide con la varianza σ_Z^2 .

- Rango o alcance. Es un valor real de forma que si $|h| \geq \text{alcance}$, entonces $\gamma(h) = \text{meseta}$. Es equivalente a decir que, en estos casos, $Z(x)$ y $Z(x+h)$ son incorreladas. No siempre existe.

Cuando el rango no existe pero la variable $Z(x)$ es estacionaria de segundo orden, puede calcularse el rango efectivo. Este es el valor real a partir del cual se verifica que si $|h| \geq \text{alcance efectivo}$, entonces $\gamma(h) \geq 0,95 \times \text{meseta}$.

- Pepita. Siempre se verifica que $\gamma(0) = 0$, pero puede ocurrir que $\lim_{|h| \rightarrow 0} \gamma(h) \neq 0$. A esta desigualdad de salto se le denomina pepita y puede deberse a diversas razones.
- Meseta parcial. Cuando el semivariograma tiene pepita, se denomina meseta parcial a la diferencia entre la meseta y la pepita.

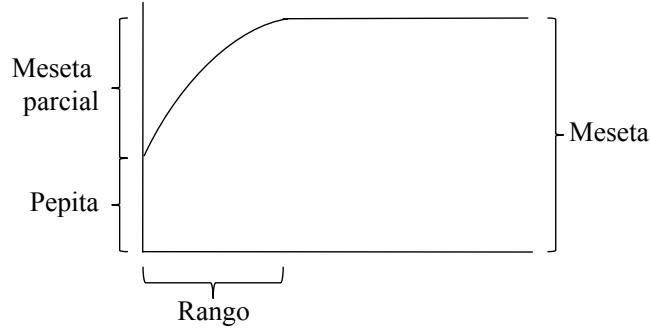


Figura 5: Partes de la función semivariograma.

Los modelos de transición o modelos con meseta son aquellos en los que la variable regionalizada $Z(x)$ es estacionaria de segundo orden. En este caso

$$\lim_{|h| \rightarrow +\infty} \gamma(h) = \sigma_Z^2 = C(0)$$

y el semivariograma está acotado superiormente.

4.1.3. Modelos de anisotropía

Si la dependencia espacial sólo depende de $|h|$ y no de su dirección, se dice que la variable $Z(x)$ tiene un comportamiento isotrópico. Obviamente, con $d = 1$ la variable sólo puede ser isotrópica.

Con $d > 1$ pueden presentarse varios tipos de anisotropía. Estos comportamientos deben detectarse calculando y estudiando los semivariogramas experimentales direccionales.

- Anisotropía zonal. Al variar la dirección varía el valor de la meseta. Su tratamiento suele ser complicado.
- Anisotropía elíptica o geométrica. Al variar la dirección varía el valor del rango, tal como puede verse en la figura 6.

Se corrige mediante un cambio de coordenadas. Con $d = 2$ el cambio es:

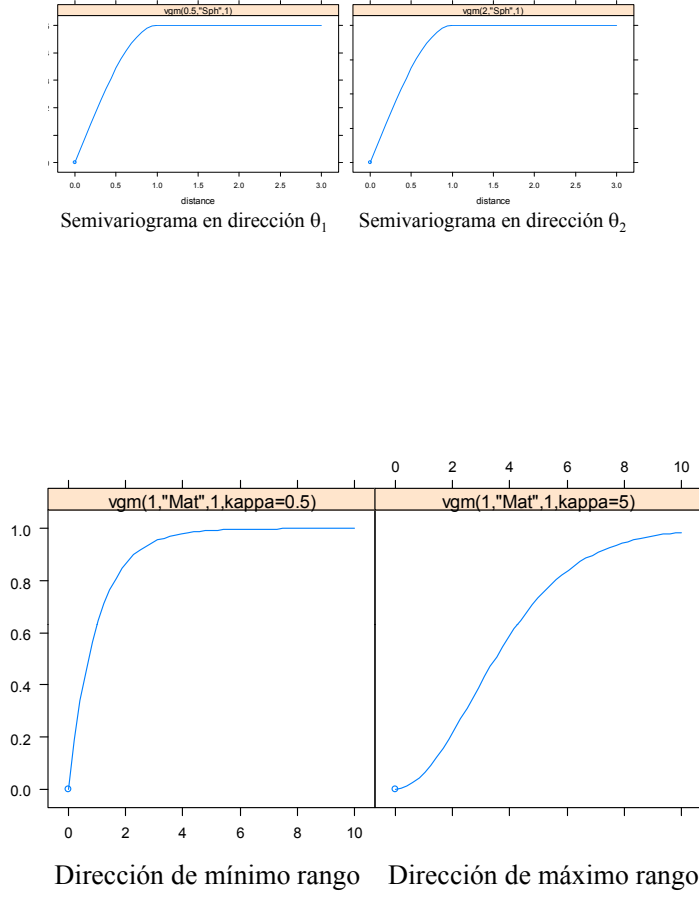


Figura 6: Caso de anisotropía geométrica.

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \lambda & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix},$$

donde θ es la dirección de máximo rango y λ es el ratio de anisotropía, que se calcula como el cociente del rango mínimo entre el rango máximo. Esta transformación se muestra en la figura 7.

Para determinar y corregir una anisotropía geométrica en \mathbb{R}^3 hay que caracterizar un elipsoide. En este caso hay que identificar tres ángulos θ_1, θ_2 y θ_3 y tres rangos. La transformación necesaria puede consultarse en el libro de Chilès y Delfiner.

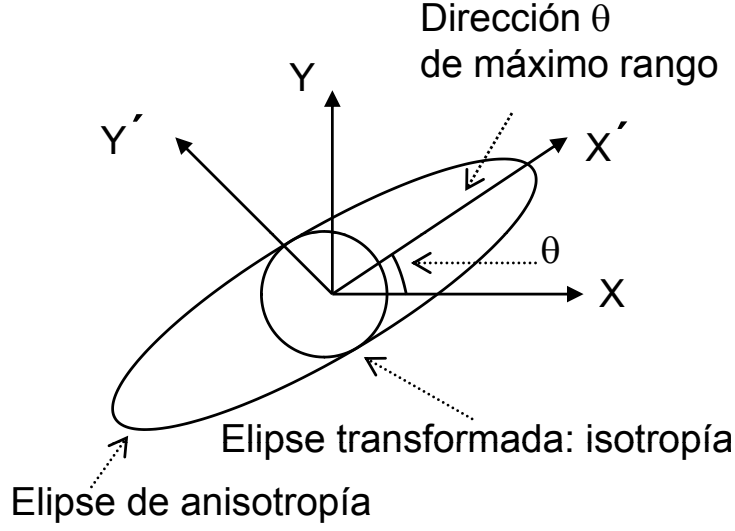


Figura 7: Corrección de una anisotropía geométrica ($d = 2$).

En la figura 8 se muestran un ejemplo de superficie de semivariograma en el cual puede apreciarse fácilmente una anisotropía geométrica. La figura 9 muestra la superficie de semivariograma de ese mismo modelo una vez que se ha realizado el cambio de coordenadas necesario para corregir la anisotropía.

4.1.4. Comportamiento en el origen

El orden del semivariograma en el origen se mide mediante el parámetro $\alpha \in \mathbb{R}$, donde

$$\lim_{|h| \rightarrow 0} \frac{\gamma(h)}{|h|^\alpha} = c.$$

Según esto podemos encontrar:

- Comportamiento parabólico: $\alpha > 1$, lo que quiere decir que $\lim_{|h| \rightarrow 0} \gamma'(h) = 0$. Indica un comportamiento regular de la variable regionalizada $Z(x)$.
- Comportamiento lineal: $\alpha = 1$, lo que quiere decir que $\lim_{|h| \rightarrow 0} \gamma'(h) \neq 0$. Indica un comportamiento irregular de la variable regionalizada.
- Comportamiento efecto pepita. Discontinuidad en el origen. Los datos son incorrelados.

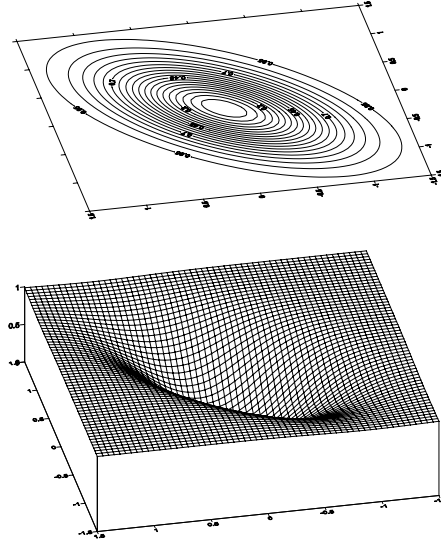


Figura 8: Modelo gaussiano anisotrópico con $\lambda = 0.5$, $\theta = 135$ y rango = meseta =1 ($d = 2$).

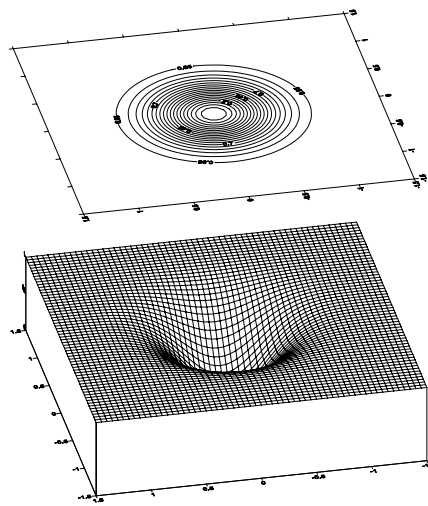


Figura 9: Modelo gaussiano isotrópico con rango =0.5 y meseta =1 ($d = 2$).

4.2. Modelos de semivariogramas

Tal como se comentó en el apartado 3.3, los semivariogramas son funciones que deben cumplir una serie de propiedades. Estas son condiciones que habitualmente $\hat{\gamma}(h)$ no cumple. Por esta razón se utilizan familias de semivariogramas teóricos válidos para ajustar los semivariogramas experimentales.

A continuación se muestran algunos semivariogramas teóricos válidos. Estos modelos se han representado en la figura 10 con valores $c = a = 1$.

- Modelo esférico, $\alpha = 1$.

$$\gamma(h) = \begin{cases} c \left(\frac{3}{2} \frac{|h|}{a} - \frac{1}{2} \frac{|h|^3}{a^3} \right) & , \text{ si } |h| \in [0, a] \\ c & , \text{ si } |h| > a \end{cases}.$$

- Modelo exponencial, $\alpha = 1$.

$$\gamma(h) = c \left[1 - \exp \left(-\frac{|h|}{a} \right) \right], \text{ si } |h| \geq 0.$$

- Modelo efecto pepita.

$$\gamma(h) = c, \text{ si } |h| > 0.$$

- Modelo gaussiano, $\alpha = 2$.

$$\gamma(h) = c \left[1 - \exp \left(-\frac{|h|^2}{a^2} \right) \right], \text{ si } |h| \geq 0.$$

- Modelo efecto agujero, $\alpha > 1$.

$$\gamma(h) = c \left(1 - \frac{\sin(|h|)}{|h|} \right), \text{ con } |h| \text{ en radianes.}$$

- Modelos potenciales, $\alpha = \theta$.

$$\gamma(h) = c |h|^\theta, \theta \in (0, 2), \text{ si } |h| \geq 0.$$

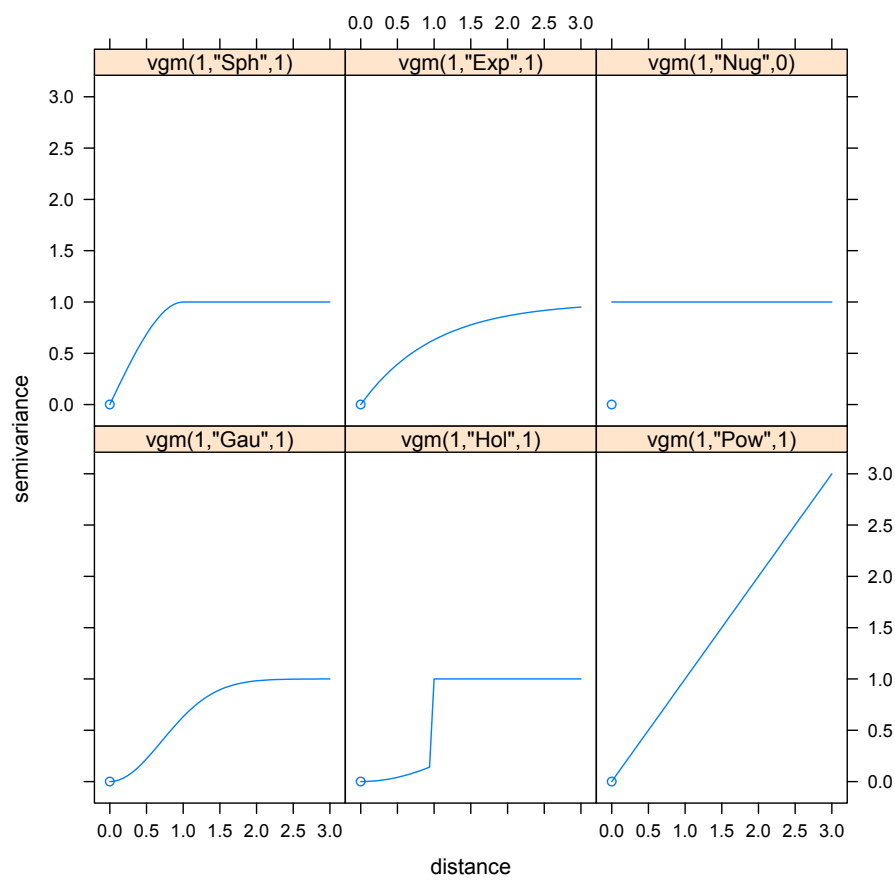


Figura 10: Modelos de semivariogramas teóricos.

4.2.1. Semivariogramas anidados

Aparecen cuando la variable presenta diversos comportamientos a distintas escalas:

$$Z(x) = m(x) + Y_1(x) + Y_2(x) + \cdots + Y_n(x),$$

donde $Y_1, Y_2, \dots, Y_n(x)$ son variables incorreladas. Entonces:

$$\gamma(h) = \gamma_1(h) + \gamma_2(h) + \cdots + \gamma_n(h).$$

En estas situaciones pueden ocurrir que $\gamma(h)$ presente pepita y/o anisotropía zonal.

4.3. Ajuste de modelos de semivariograma

Dado el semivariograma muestral $\hat{\gamma}(h)$, se selecciona un modelo $\gamma(h)$ válido y se ajustan las dos funciones haciendo variar los parámetros del modelo teórico.

- Ajuste manual

Casi todos los programas de Geoestadística, como el **Variowin** o el **R**, permiten el ajuste manual de las funciones.

- Ajuste automático

Hay varios métodos. Uno de los más utilizados es el de mínimos cuadrados ponderados (implementado en **R**).

Los parámetros del modelo teórico son escogidos de manera que minimizan una expresión del tipo:

$$\sum_{j=1}^J w_j^2 [\hat{\gamma}(h_j) - \gamma(h_j)]^2,$$

donde los pesos w_j determinan para qué valores de h_j se desea un ajuste más “fino”.

- Cuando $w_j=1$ el método es el de mínimos cuadrados ordinarios.
- Cuando los datos siguen una distribución normal el método máximo-verosímil coincide con el de mínimos cuadrados generalizados. Este último construye los

pesos w_j a partir de la matriz de varianzas-covarianzas de $\hat{\gamma}(h_j)$. Por todo esto, bajo normalidad (y dentro de mínimos cuadrados ponderados), una buena elección de los pesos es

$$w_j^2 = \{Var[\hat{\gamma}(h_j)]\}^{-1} \simeq \frac{N(h_j)}{2\gamma(h_j)^2}$$

El programa R utiliza por defecto los pesos $w_j^2 = \frac{N_{h_j}}{h_j^2}$, pero permite la opción de escoger otros.

Ejercicio 2. Análisis estructural.

1. Análisis de una variable de uno de los ficheros de datos de ejemplo.
2. Cálculo de semivariogramas experimentales para determinar el tipo de dependencia espacial.
3. Estudio del efecto de la media, la falta de normalidad y/o comportamiento anisotrópico.
4. Selección de modelos de semivariogramas teóricos.
5. Ajuste de las funciones mediante diversos métodos. Comparación de resultados.

Observaciones para el ejercicio 2.

Trabajaremos con la librería `gstat` del R. Ejecutando `data(package="gstat")` podéis ver una lista de ficheros de datos de esta librería. Quiero que hagáis un análisis estructural de una variable regionalizada.

He hecho una asignación aleatoria de trabajos a alumnos. Se recoge en la siguiente tabla:

| Nombre | Fichero | Variable |
|----------------------------------|----------------|----------|
| Cacheiro Martínez, Pilar | coash | coash |
| Castro Dacal, Silvia | fulmar | fulmar |
| Díaz Costoya, Iván | meuse | cadmium |
| Duarte Castro Rocha, Elisa María | meuse | lead |
| Leyenda Rodríguez, María | meuse | zinc |
| López Pallas, Ana Karina | pcb | PCB138 |
| López Veiga, David | prediction.dat | Cd |
| Martínez Silva, Isabel María | prediction.dat | Co |
| Oliveira Pérez, María | prediction.dat | Ni |
| Oviedo de la Fuente, Manuel | prediction.dat | Zn |
| Suárez Crespo, Silvia | sic.train | day01 |
| Touriño Sánchez, Ana | walker | V |

Una observación: El fichero `prediction.dat` se carga mediante la instrucción `data(jura)` y el fichero `sic.train` mediante la instrucción `data(sic2004)`. El resto de los ficheros se cargan directamente utilizando su nombre, tal como he hecho en el fichero de ejemplo.

Ejemplo del ejercicio 2. Para que os sirva de referencia, podéis echar un vistazo al estudio de la variable cobre del fichero `meuse`. Tenéis que daros cuenta de que al trabajar con otro conjunto de datos puede ocurrir que algunas de las opciones incluidas en este ejemplo carezcan de sentido, mientras que otras que quizá no se incluyan en el ejemplo sí lo tengan ahora. Cada uno debe saber adaptarse al tipo de datos que le haya tocado. Antes del 12 de marzo tenéis que enviarme un script de R similar al ejemplo.

Con el fichero `meuse` yo no he profundizado en las conclusiones sobre las características de los datos (normalidad, tendencia, anisotropía, modelo de dependencia espacial,...) porque no quiero dar demasiadas pistas en el ejemplo, pero es algo que vosotros sí debéis incluir. Como trabajo opcional sería bien recibido, además del script, un breve fichero de texto resumiendo y comentando las principales conclusiones a las que habéis llegado con vuestro fichero de datos.

Por último, soy consciente de que algunas variables serán más “bonitas” que otras a la hora de trabajar. Quiero que os deis cuenta de que eso es inevitable y que resulta totalmente imposible encontrar ficheros de datos con el mismo nivel de dificultad. A

la hora de corregir tendré en cuenta qué datos teníais asignados y qué habéis podido obtener de ellos.