Diplomado en Análisis de Información Geoespacial

Análisis estructural

Autor:
M. en G. Alberto Porras Velázquez

Contoy 137 Esq. Chemax, Col. Lomas de Padierna, Delegación Tlalpan, CP 14240, México, CDMX.

3.2 Análisis estructural

Introducción

Una primera tarea a realizar como parte de cualquier análisis geoestadístico consiste en buscar la estructura de los datos que se estudian, esto implica analizar los atributos distintivos de la regionalización, de ahí el nombre de análisis estructural.

El semivariograma es la herramienta principal que permite describir el comportamiento de la variación espacial de un fenómeno. A través de un análisis exploratorio y con base en el modelado del variograma se busca estructura en el comportamiento del fenómeno. Posteriormente, el modelado del variograma se ajusta a un modelo teórico que será utilizado en el modelo de interpolación de la variable.

El semivariograma

La hipótesis intrínseca para las variables regionalizadas establece que la varianza de las diferencias entre los valores de una variable es una función que depende de la distancia de separación h entre las ubicaciones.

$$Var[Z(x+h) - Z(x)] = 2\gamma(h)$$

A la función $2\gamma(h)$ se le llama variograma y a $\gamma(h)$, semivariograma. Éste último queda definido mediante la expresión:

$$\gamma(h) = (1/2)\, E[Z(x+h) - Z(X)]^2$$

Es común utilizar indistintamente el término variograma o semivariograma para referirse a la función $\gamma(h)$, aunque estrictamente hablando la función $\gamma(h)$ es el semivariograma.

El semivariograma $\gamma(h)$ es una característica de disimilitud entre dos lugares dada una distancia de separación (h), y también es una medida que permite caracterizar la autocorrelación espacial en el proceso.

Para obtener un modelo del semivariograma a partir de los datos observados empíricamente, se realiza el semivariograma experimental definido como:

$$\hat{\gamma}(\boldsymbol{h}) = \frac{1}{2N_{\boldsymbol{h}}} \sum_{i=1}^{N_{\boldsymbol{h}}} \left(Z(x_i) - Z(x_i + h) \right)^2$$

El variograma experimental es un promedio de diferencias al cuadrado por clase de distancia h, en donde N(h) es el número de pares de puntos que se encuentran a una distancia h.

A continuación daremos un recorrido por distintos tipos de herramientas de apoyo en el análisis estructural.

El objetivo final es llegar a un modelo teórico del semivariograma que nos permita caracterizar la autocorrelación espacial en función de la distancia, así como utilizar ese mismo modelo para realizar estimaciones de los valores de la variable en sitios no muestreados.

Diagramas de dispersión h

En la ilustración 1 se muestran las ubicaciones de las mediciones tomadas para las concentraciones de diferentes metales en una pequeña región de Suiza, estos datos son conocidos como Jura (nombre del lugar). Esta región tiene una extensión aproximada de 14.5km^2.

El color de cada punto corresponde a la medición de la concentración de zinc en partes por millón (ppm). Las distancias para estos datos están expresadas en kilómetros, así que, por ejemplo, 100 metros se representan como 0.1 km en las gráficas que aparecerán sucesivamente.

Contoy 137 Esq. Chemax, Col. Lomas de Padierna, Delegación Tlalpan, CP 14240, México, CDMX.

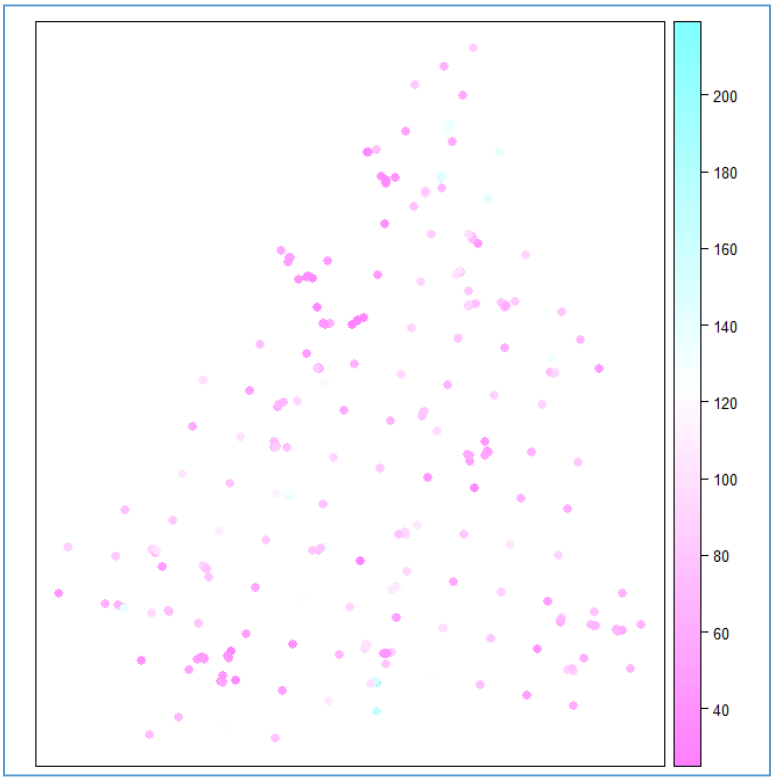

Ilustración 1. Distribución de las mediciones de zinc en la región de Jura. Porras, A. (2014)

Una primera herramienta para estudiar la correlación entre las mediciones de zinc es el diagrama de dispersión h. Para realizar este gráfico, primero hay que elegir una distancia de separación específica entre las observaciones.

Por ejemplo, nos podría interesar el estudio de la correlación entre las observaciones separadas a 100 metros; sin embargo, difícilmente encontraremos suficientes pares de observaciones (quizá ninguno) separadas justo a esa distancia. Es por eso que debemos especificar una tolerancia, por ejemplo, de 50 metros.

De este modo incluiremos en la clase de 100 metros todos los pares de observaciones que se encuentren entre los 50 (100m - 50m) y 150 (100m + 50m) metros de separación.

Contoy 137 Esq. Chemax, Col. Lomas de Padierna, Delegación Tlalpan, CP 14240, México, CDMX.

El método consistiría en identificar para cada observación todas las mediciones ubicadas a una distancia entre 50 y 150 metros.

En este escenario, supongamos que la observación i tiene una concentración de 100 ppm, y que las observaciones j y k están en el margen de distancia establecido, una con concentración de 115 ppm y otra con 125 ppm.

El valor de la observación i se conoce como cola, y a los valores de las concentraciones de los puntos j y k se les llama cabezas. Es decir, z(x) es la cola y z(x+h) será una cabeza.

En el diagrama de dispersión se producirán dos puntos (pares ordenados) con valores (100, 115) correspondientes a la relación entre i, j y el punto (100, 125) correspondiente a la relación i, k.

En el diagrama de dispersión h, al igual que en los diagramas de dispersión vistos para la descripción bivariada, la correlación se visualiza de acuerdo con el nivel de agrupamiento (o dispersión) de la nube de puntos en la gráfica. Un mayor agrupamiento implica una mayor correlación.

Se pueden plantear entonces varios diagramas de dispersión h para estudiar el decaimiento de la autocorrelación de la variable en términos de la distancia, de acuerdo con la ley de Tobler.

Por ejemplo, en la ilustración 2 se grafican ocho diagramas de dispersión h aplicables a la concentración de zinc para clases de distancia de 100 metros. Se observa que la correlación decae drásticamente para las observaciones que se encuentran a más de 500 metros de separación (r = 0.217).

En la práctica se recomienda realizar diferentes diagramas de dispersión, variando las clases de distancia cada 50 metros o 200 metros, por ejemplo, hasta obtener una descripción adecuada del comportamiento de la correlación.

El reto está en encontrar un equilibro entre el detalle y la generalidad. Si las clases de distancia y sus tolerancias asociadas son pequeñas, se puede tener mucha variabilidad en los diagramas, pero si las clases de distancia y sus tolerancias son demasiado grandes, se perderá el detalle del comportamiento.

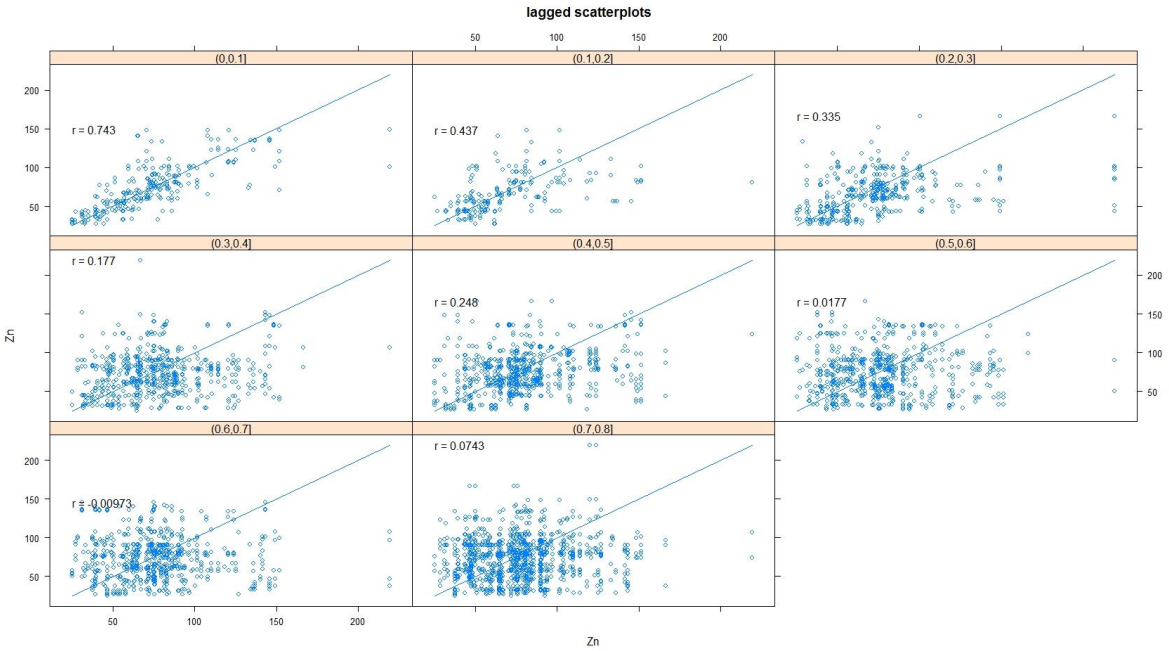

Ilustración 2. Diagramas de dispersión h para las concentraciones de zinc en los datos Jura. Porras, A. (2014)

Nube del variograma

Observa que la expresión matemática para el cálculo del semivariograma de una clase de distancia h es

$$\hat{\gamma}(h) = \frac{1}{2N_h}\sum_{i=1}^{N_h}\left(Z(x_i) - Z(x_i + h)\right)^2,$$

en donde se incluyen los términos $(Z(x_i) – Z(x_i +h))^2$, llamados semivarianzas.

La nube del variograma es una gráfica que muestra todos los términos de la semivarianza (para todas las distancias h). Esto significa que para cada punto observado se calcula la distancia hacia todos los puntos restantes y se obtiene el producto
$(Z(x_i) – Z(x_i +h))^2$.

En la gráfica (ilustración 3) el eje x corresponde a la distancia h de separación entre cada par de observaciones, y en el eje **y** al valor de la semivarianza correspondiente.

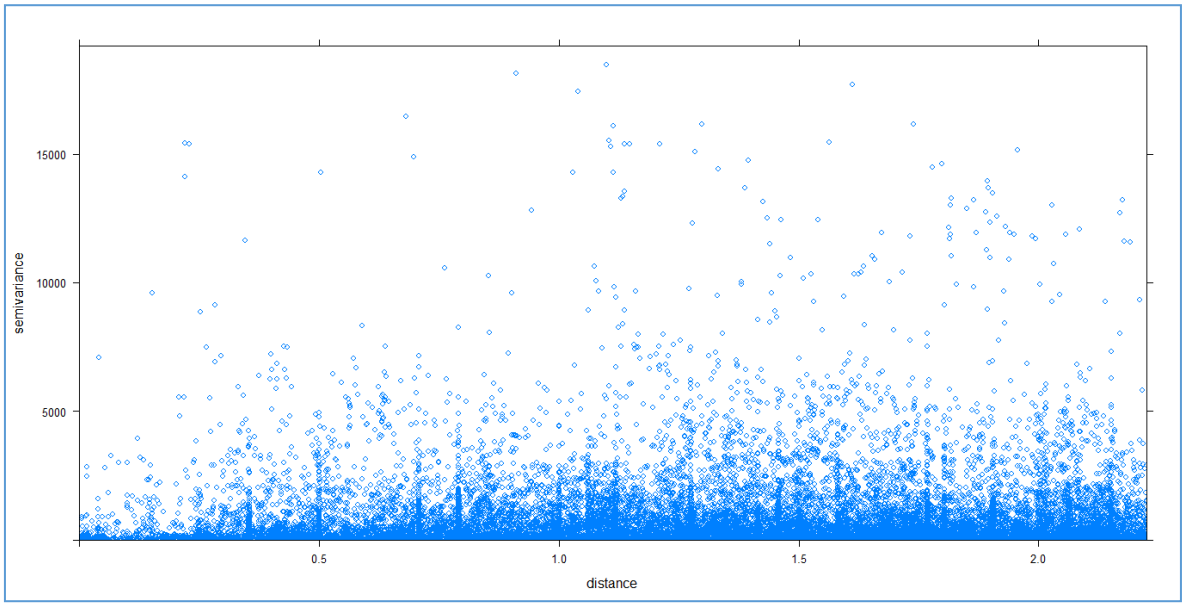

Ilustración 3. Nube del variograma para la concentración de zinc en los datos Jura. Porras, A. (2014)

El cálculo del variograma experimental implica promediar las semivarianzas que corresponden a una clase de distancia determinada. Por ejemplo, para la clase de distancia de100 metros con tolerancia de 100 metros, se incluirían las semivarianzas cuyos valores correspondientes en el eje **x** van de 0 a 200 metros (ilustración 4), así **γ** (100) sería un promedio de las semivarianzas incluidas entre 0 y 200 metros.

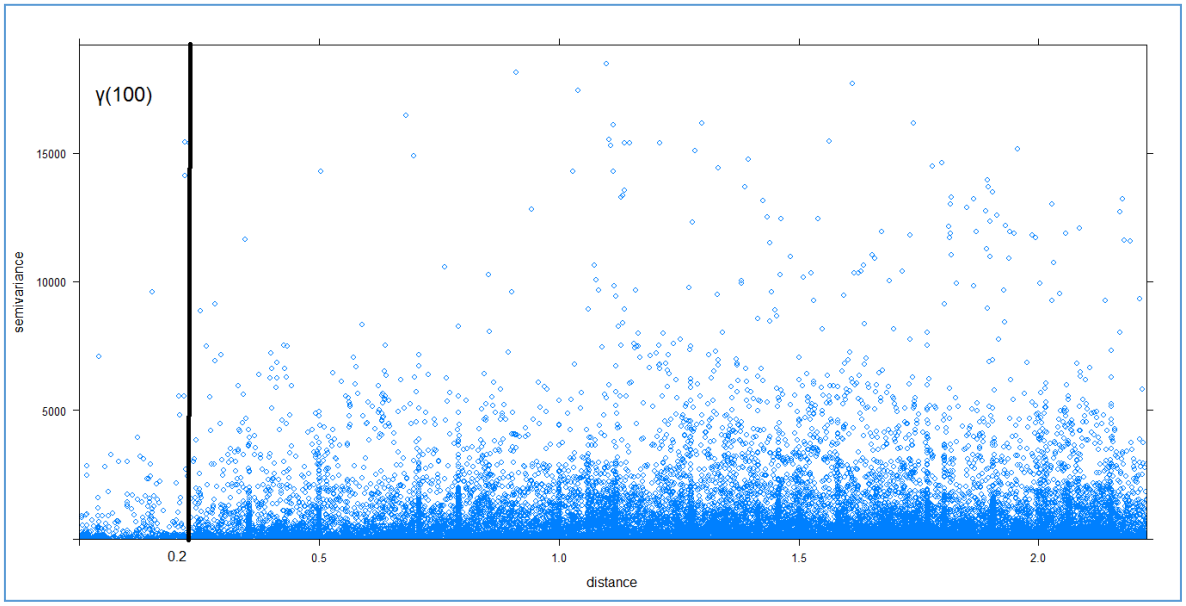

Ilustración 4. Cálculo de la función del semivariograma para γ(100). Porras, A. (2014)

La nube del semivariograma permite identificar los términos de la semivarianza capaces de sesgar el promedio de los datos. Dicho de otro modo, se pueden encontrar observaciones con valores extremos.

En la ilustración 5 se selecciona un conjunto de puntos correspondientes a las semivarianzas que se encuentran “arriba” del promedio, es decir, sobre la nube en donde se concentran los puntos.

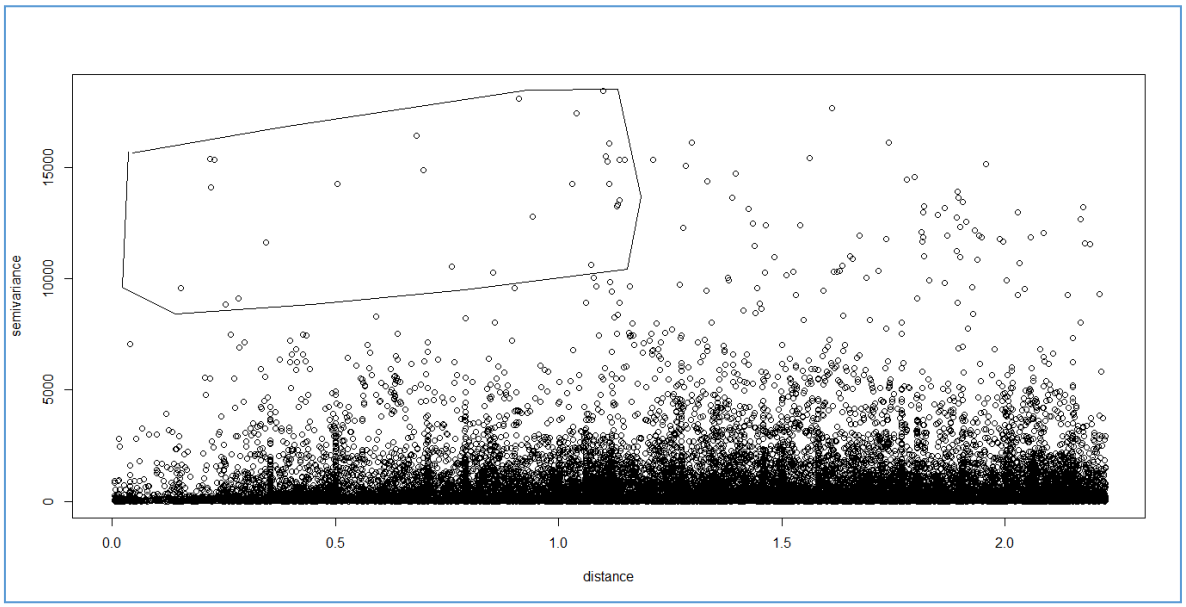

Ilustración 5. Selección de semivarianzas con valores “arriba” del promedio de la nube. Porras, A. (2014)

En el mapa se pueden identificar los pares de observaciones que implican el cálculo de cada una de las semivarianzas seleccionadas.

En la ilustración 6, cada línea une a un par de puntos que producen un resultado (un punto) en la nube del semivariograma.

Contoy 137 Esq. Chemax, Col. Lomas de Padierna, Delegación Tlalpan, CP 14240, México, CDMX.

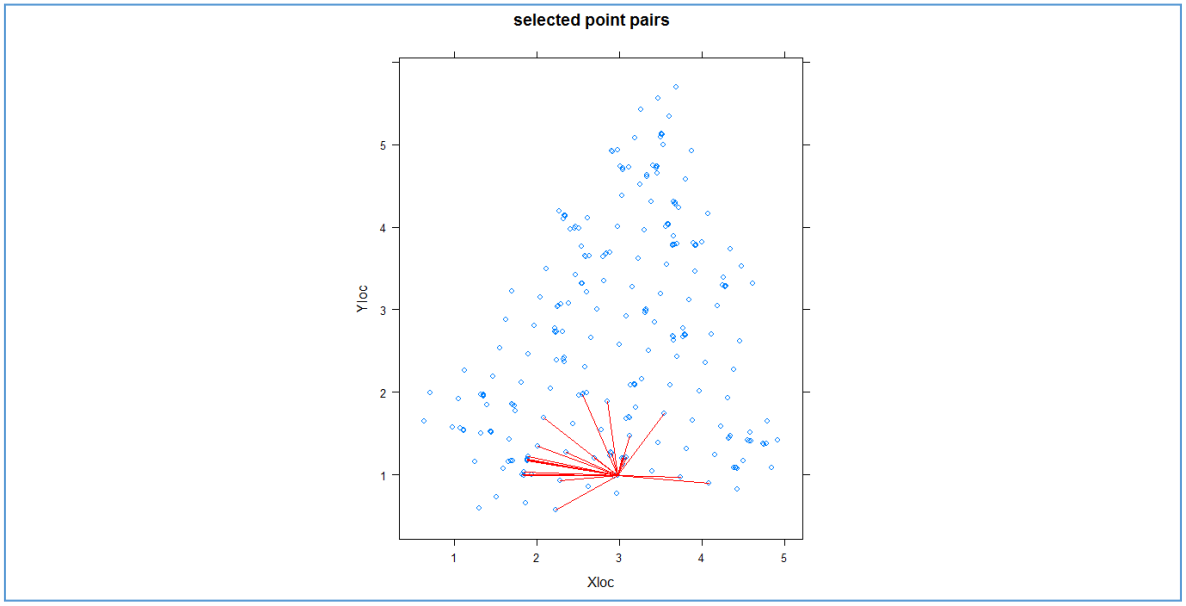

Ilustración 6. Ubicación de pares de observaciones que implican semivarianzas altas en la nube del variograma. Porras, A. (2014)

En el caso de esta selección todas las líneas convergen en un punto, lo cual impica que éste puede representar un valor extremo. Bajo criterios determinados (si hay errores de medición o características diferentes del fenómeno en una región determinada), las

observaciones se pueden eliminar de la muestra. En este caso, el dato es correcto y no debe borrarse.

El siguiente paso consiste en calcular el variograma experimental para distintas clases de distancia (por lo regular se dan a separaciones y tolerancias iguales).

Para que la gráfica defina una tendencia del semivariograma experimental se recomienda que la N(h), el número de términos de semivarianza en cada clase de distancia h, sea mínimo de 30 (ver ilustración 7).

También es recomendable que la h máxima sea menor que D/2, en donde D es la distancia máxima entre dos puntos. Esto debido a que para las distancias h muy grandes habrá pocos pares de puntos correspondientes a cada clase de distancia.

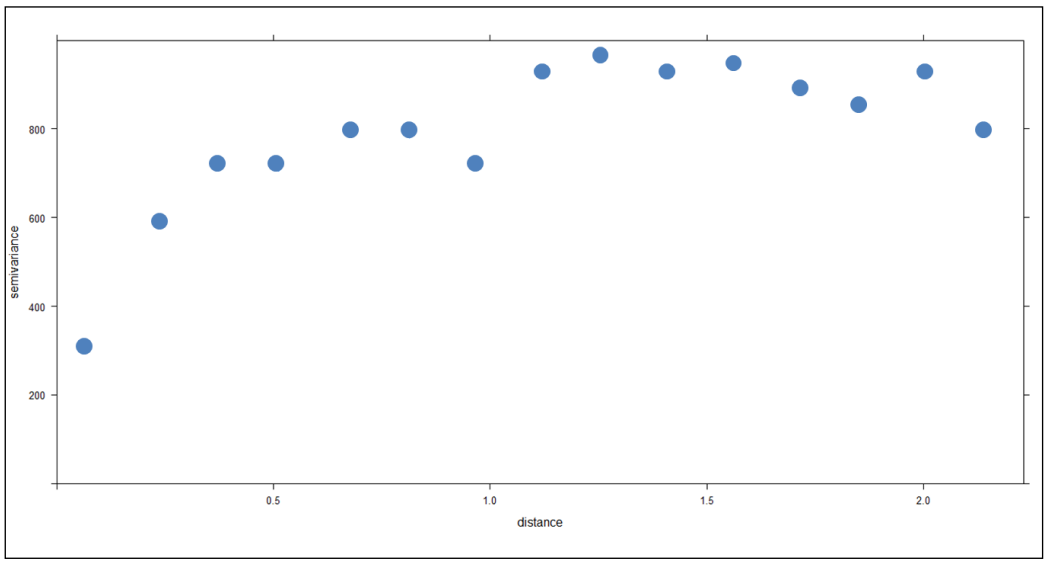

Ilustración 7. Semivariograma experimental para el zinc para los datos Jura. Porras, A. (2014)

Contoy 137 Esq. Chemax, Col. Lomas de Padierna, Delegación Tlalpan, CP 14240, México, CDMX.

El variograma experimental provee información relevante sobre el comportamiento estructural del fenómeno de estudio.

En la gráfica de la ilustración 7, cada punto representa el cálculo de una γ(h) distinta. Recordemos que el semivariograma es una medida de disimilitud entre las observaciones que se encuentran a una distancia h entre sí.

El comportamiento de la gráfica es típico de una variable regionalizada que cumple la hipótesis intrínseca. Se observa que a distancias pequeñas hay menor disimilitud entre los valores de la variable y que conforme aumenta la distancia esta disimilitud crece, hasta alcanzar un valor más o menos constante.

La distancia a partir de la cual esta gráfica se vuelve constante se llama rango y determina la distancia máxima en la que hay correlación entre las observaciones.

En el tema de las variables regionalizadas en la práctica no se requiere que el fenómeno sea débilmente estacionario en toda la región de estudio, sino que sea débilmente estacionario en una ventana, este rango determina el tamaño de la ventana de influencia.

Cada punto en la gráfica del variograma experimental puede considerarse como el resumen (promedio) de la información que aparece en un diagrama de dispersión h para una distancia determinada.

Características del variograma experimental

Para estimar los valores de la variable en sitios no muestreados es necesario un modelo teórico que describa el comportamiento espacial de la variable de estudio.

Contoy 137 Esq. Chemax, Col. Lomas de Padierna, Delegación Tlalpan, CP 14240, México, CDMX.

Recordemos que el variograma experimental es la gráfica de una serie de puntos y por eso se requiere de un modelo teórico (una función matemática) a la cual se ajusten los puntos del variograma experimental.

Estudiemos el comportamiento del variograma experimental para ajustarlo a un modelo teórico.

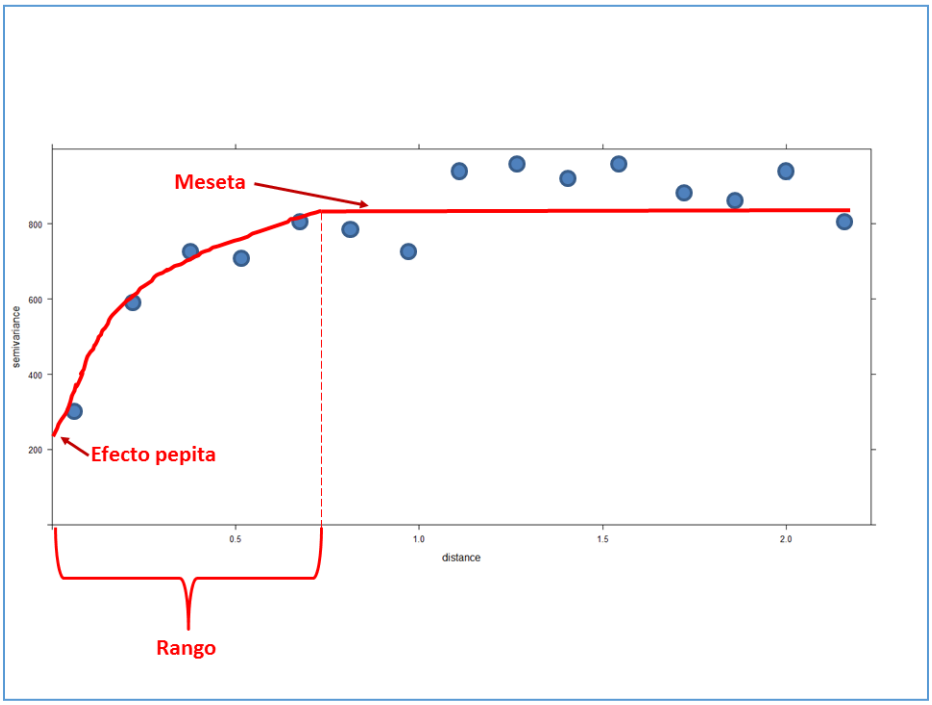

Ilustración 8. Comportamiento del variograma experimental. Porras, A. (2014)

En la ilustración 8 se muestran tres elementos característicos del variograma experimental que deben tomarse en cuenta para hacer un ajuste a un modelo teórico.

Efecto pepita (nugget). Se produce como consecuencia de las limitantes que impone la separación mínima entre dos observaciones y suele atribuirse a una variación en micro escala.

El nombre del efecto se debe a que éste se identificó cuando se estudiaba la concentración de oro en las minas.

Supongamos que en un terreno tomamos muestras muy cercanas una de la otra y descubrimos una pepita de oro. Si calculamos el semivariograma en un caso así, identificaremos que para distancias muy pequeñas (cercanas a cero) habrá grandes variaciones en los términos de la semivarianza.

Entre dos puntos muy cercanos tendremos concentraciones de oro muy distintas. En las cercanías de la pepita tendremos valores muy pequeños de concentración de oro y en la pepita hallaremos valores de concentración muy altos, por lo cual las diferencias al cuadrado $(Z(x_i) – Z(x_i +h))^2$ serán significativas.

En el ejemplo, el efecto nugget es de aproximadamente 220 unidades de semivarianza.

La meseta (sill). Observa que en la ilustración 8 los valores para el variograma experimental parecen variar en torno a una constante, a partir de una distancia aproximada de 0.7 km. Este valor constante se identifica como la meseta, que en este caso es de aproximadamente 800 unidades de semivarianza.

El rango (range). La distancia a la cual se alcanza el valor de la meseta se llama rango (0.7km en el ejemplo) y constituye la máxima distancia para la que se considera que existe correlación entre los valores de la variable de estudio. En consecuencia, las observaciones separadas entre sí más allá de este rango no tendrán correlación entre sí.

El rango es un parámetro de importancia en el modelo, debido a que establece la ventana de influencia de manera que al momento de interpolar el valor de la variable en un punto desconocido, influirán en la interpolación las observaciones (valores medidos) que no estén a una distancia de separación con respecto al punto a interpolar mayor al rango.

Modelos teóricos para el semivariograma y ajuste del variograma experimental

A partir del variograma experimental se deben ajustar modelos matemáticos. El modelo teórico del variograma debe satisfacer ciertas condiciones para no terminar con una varianza negativa para el valor estimado, lo cual sería completamente inaceptable.

No es fácil reconocer funciones con esas propiedades, por lo que es mejor ajustar el variograma experimental a ciertos variogramas modelo que cumplan con estas propiedades.

Efecto pepita. Este modelo corresponde a un fenómeno puramente aleatorio con ninguna correlación entre los valores de la variable, sin importar su cercanía.

Efecto pepita (Nugget) : $\gamma(h) = 0$ si $h=0$

$\gamma(h) = C$ si $h>0$

Observa que en este modelo el valor de la meseta (C) se alcanza inmediatamente después de la distancia cero, lo cual indica la carencia de correlación entre las observaciones.

Modelo esférico. Este modelo corresponde a los fenómenos más frecuentemente observados y por ello es utilizado ampliamente. El comportamiento es de una tendencia casi lineal hasta que se alcanza el rango en donde se estabiliza el fenómeno (meseta). Las ecuaciones correspondientes son:

Contoy 137 Esq. Chemax, Col. Lomas de Padierna, Delegación Tlalpan, CP 14240, México, CDMX.

$$\gamma(h) = C[\frac{3}{2}\frac{h}{a} - \frac{1}{2}(\frac{h}{a})^3] \qquad \textbf{si} \quad \mathbf{h < a}$$

$$\gamma(h) = C \qquad \textbf{si}\ \mathbf{h \geq a}$$

en donde **C** es el valor de la meseta o sill, **h** la distancia y **a** el rango. En la ilustración 9 se muestra el comportamiento de esta función.

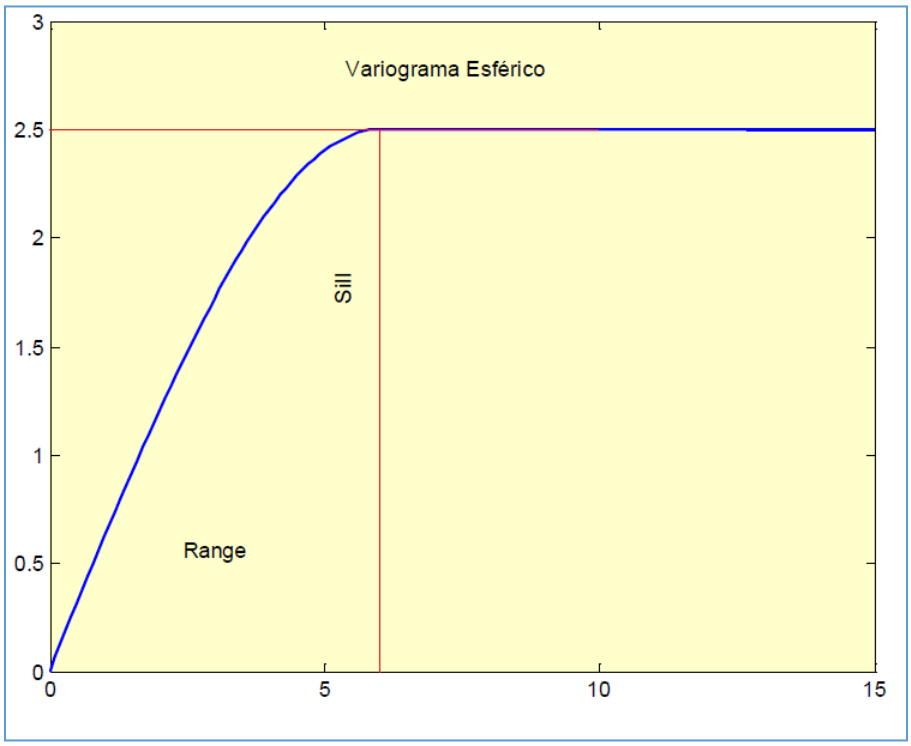

Ilustración 9. Modelo de variograma esférico. Porras, A. (2014)

Modelo exponencial. Tiene un comportamiento muy similar al esférico, la diferencia con aquel radica en que nunca se alcanza el valor de la meseta, pues su rango práctico **a**

llega sólo a 95 por ciento. En la ilustración 10 se muestra el comportamiento de la gráfica para este modelo.

$$\gamma(h) = C[1 - \exp(-\frac{3h}{a})]$$

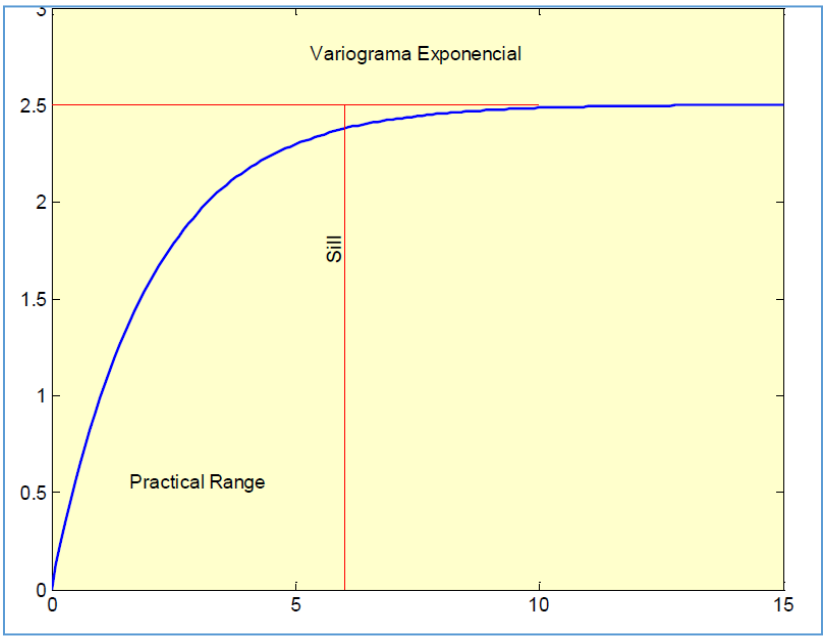

Ilustración 10. Modelo exponencial. Porras, A. (2014)

Modelo gaussiano. Se utiliza cuando el fenómeno de estudio es extremadamente continuo, su rango práctico es **a** (ver ilustración 11).

$$\gamma(h) = C[1 - \exp(-\frac{3h^2}{a^2})]$$

Contoy 137 Esq. Chemax, Col. Lomas de Padierna, Delegación Tlalpan, CP 14240, México, CDMX.

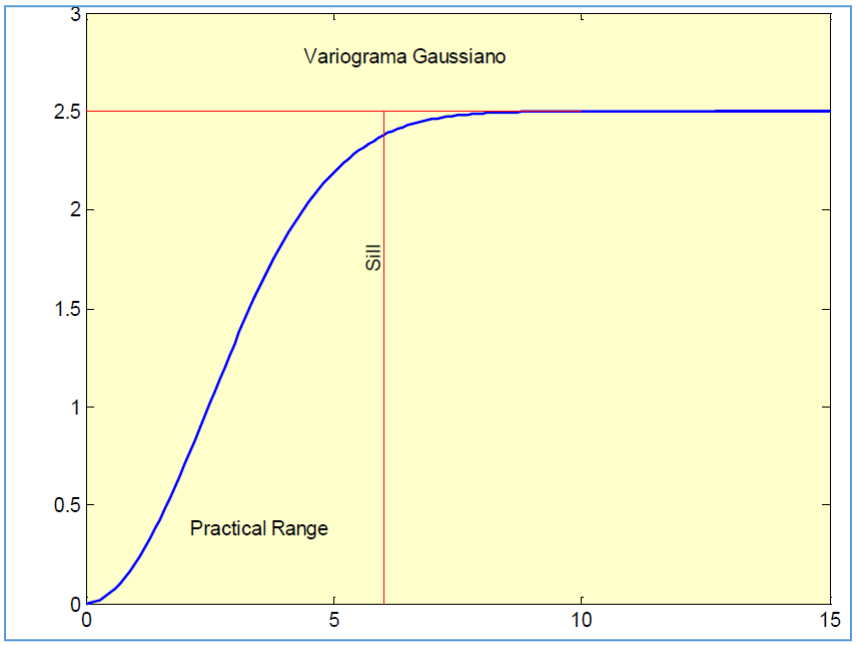

Ilustración 11. Modelo gaussiano. Porras, A. (2014)

Los modelos teóricos presentados son algunos de los más utilizados, sin embargo es pertinente mencionar que existe una gran cantidad de modelos teóricos que se pueden aplicar.

El ajuste de un variograma experimental a un modelo teórico puede darse con la combinación de varios modelos. Por ejemplo, el variograma de la ilustración 8 se puede modelar con una combinación de efecto pepita y un variograma esférico.

En este caso, el ajuste obtenido combina el modelo para el efecto pepita con C_n = 218 más un modelo esférico con una meseta parcial (meseta total – el efecto pepita) igual a 610 y un rango aproximado de 0.610km (ilustración 12).

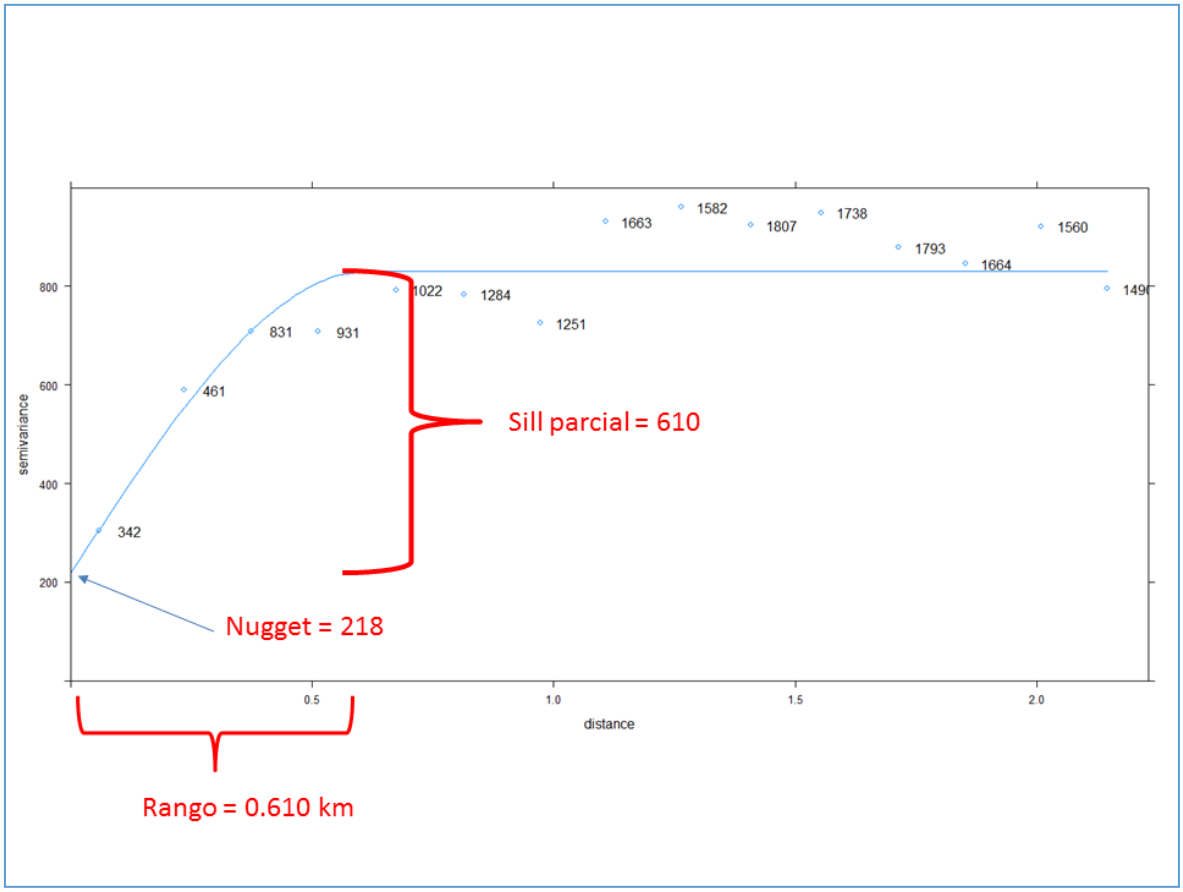

Ilustración 12. Ajuste del semivariograma experimental a un modelo teórico con dos componentes (efecto pepita y modelo esférico). Porras, A. (2014)

Te preguntarás cual es el mejor criterio o método para ajustar un variograma experimental a un modelo teórico. Pues bien, uno de los mejores ajustes que se puede hacer es a “ojo”, como lo hicimos en este caso.

Por lo regular, el método de mínimos cuadrados no ofrecerá un buen ajuste, ya que bajo este criterio todos los puntos de la gráfica tienen el mismo peso para minimizar el error y sabemos que los puntos en el variograma (correspondientes a distancias pequeñas) deben tener un ajuste mejor, pues determinan el comportamiento de la correlación.

Como veremos, inicialmente se puede ajustar un modelo a partir del criterio del investigador (a ojo) y posteriormente refinarlo con el software (como se verá en la práctica correspondiente a este tema).

En estas lecturas hemos utilizado el variograma como herramienta de modelaje, pero es pertinente mencionar que los mismos procesos se pueden aplicar a la covarianza que, a diferencia del semivariograma, es una medida de similitud.

Exite una equivalencia entre el semivariograma y la covarianza dada por

$\gamma(h) = C(0) - C(h)$

En la ilustración 13 se ve el comportamiento de ambas funciones.

Contoy 137 Esq. Chemax, Col. Lomas de Padierna, Delegación Tlalpan, CP 14240, México, CDMX.

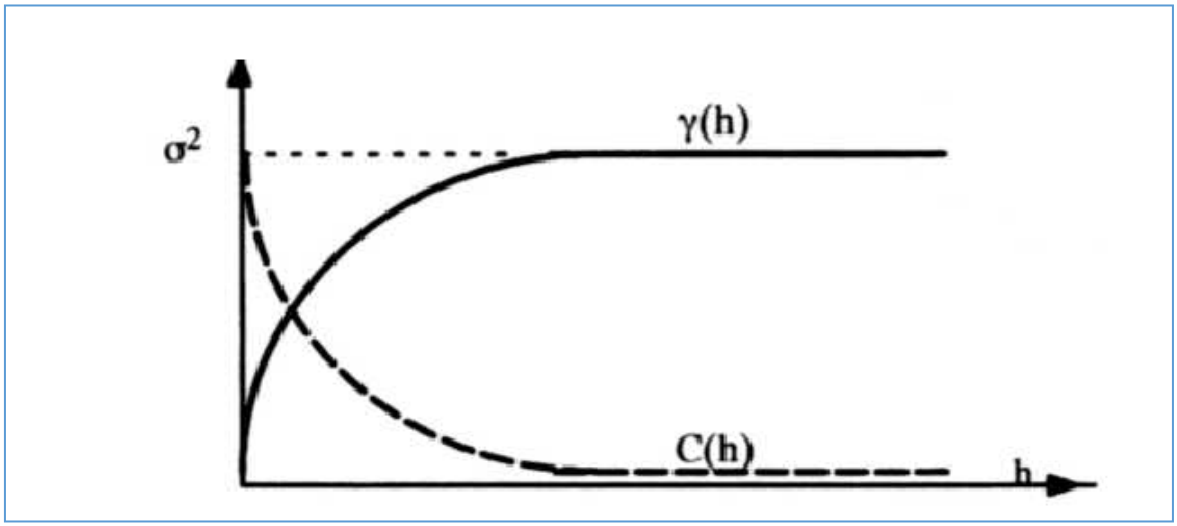

Ilustración 13. Comportamiento de las funciones del semivariograma y covarianza en función de la distancia (h). Armstrong, M. (1998)

Algunas consideraciones finales.

La Geoestadística es una disciplina práctica, por ende el conocimiento se da en buena medida a través de la experiencia. Los elementos vistos en estas lecturas constituyen una base mínima a partir de la cual puedes adentrarte en el tema.

Por ejemplo, el modelado que realizamos como parte del curso supone que los fenómenos son isotrópicos, es decir, que el comportamiento es el mismo en cualquier dirección. Sin embargo en la realidad no siempre sucede así, pues fenómenos como las precipitaciones pluviales no son isotrópicos.

Contoy 137 Esq. Chemax, Col. Lomas de Padierna, Delegación Tlalpan, CP 14240, México, CDMX.

En esos casos se requiere el modelado de variogramas direccionales, en donde, además de la distancia y la tolerancia de la distancia, se debe definir una tolerancia angular.

Si como estudiante deseas profundizar en este tema, te recomendamos revisar las referencias citadas en el curso. Los libros de Armstrong (1998) y Goovaerts (1997) son una buena base para los aspectos teóricos, mientras que en el libro de Bivand (2013) puedes consultar ejemplos de aplicación en el entorno **R**.

Referencias:

- Armstrong, M. (1998). Basic Linear Geostatistics. Berlin, Germany Springer-Verlag.
- Bivand R.S., Pebesma, E. y Gómez-Rubio, V. (2013). Applied Spatial Data Analysis with R (2ed). New York, United States of America: Springer.
- Goovaerts, P. (1997). Geostatistics for Natural Resources Evaluation. New York, United States of America: Oxford University Press.