UNIVERSIDAD NACIONAL DE ASUNCIÓN

Facultad Politécnica

Diplomado en Computación Estadística y Análisis de Datos

Módulo 7: Métodos de Geoestadística

Clase 4: Introducción a la Geoestadística

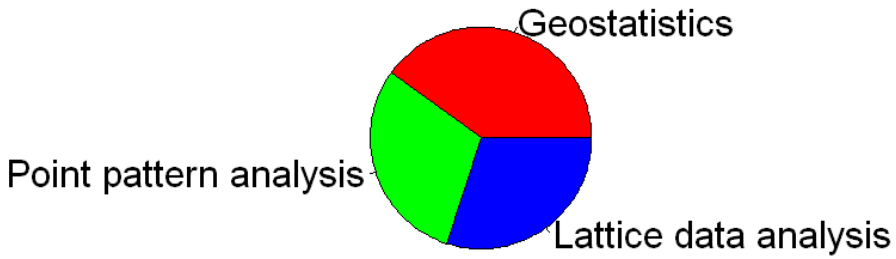
Profesor: Carlos Giménez

email: charlieswall@gmail.com

1. Estadística espacial

La Estadística espacial es la reunión de un conjunto de metodologías apropiadas para el análisis de datos que corresponden a la medición de variables aleatorias en diversos sitios (puntos del espacio o agregaciones espaciales) de una región de interés o área de estudio. De manera más formal se puede decir que la estadística espacial trata con el análisis de realizaciones de un proceso estocástico de interés.

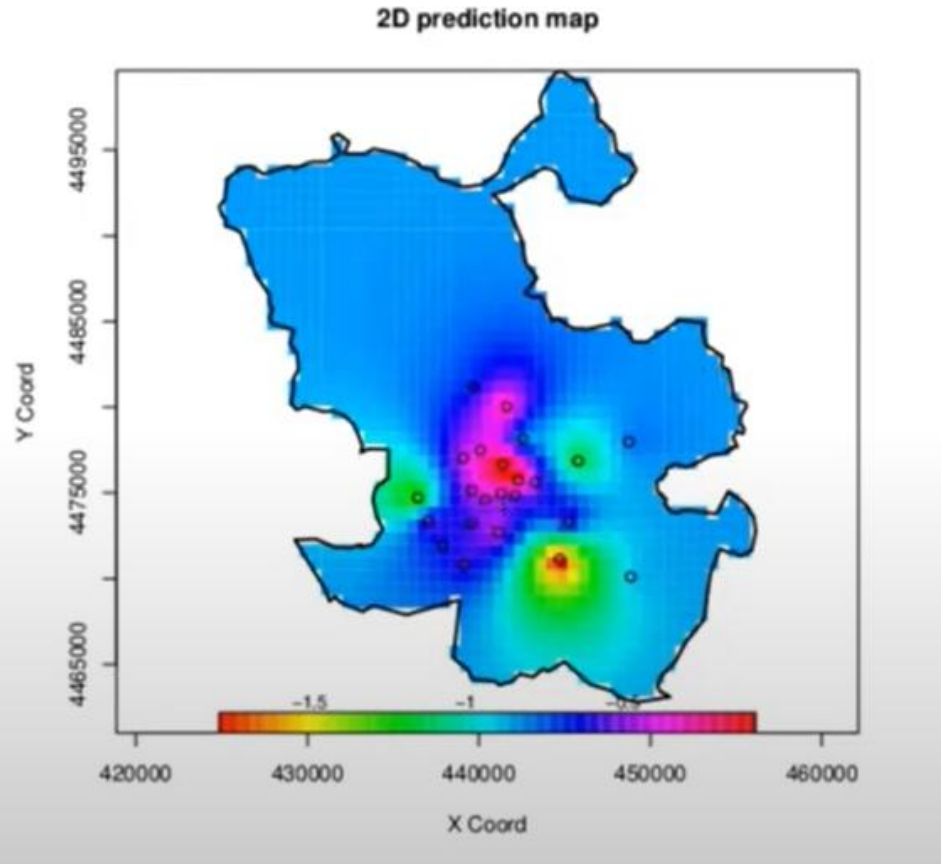
La estadística espacial se subdivide en tres grandes áreas. La pertinencia de cada una de ellas está asociada a las características del conjunto D de índices del proceso estocástico de interés.



1.1 Datos geoestadísticos

Las ubicaciones **s** provienen de un conjunto **D** continuo y son seleccionadas a juicio del investigador (D fijo). Algunos ejemplos de datos que pueden ser tratados con esta metodología son: Niveles de un contaminante en diferentes sitios de una parcela, contenidos auríferos de una mina, valores de precipitación. En los ejemplos anteriores es claro que hay continuidad espacial, puesto que en cualquier sitio de la parcela, de la mina, o del acuífero pueden ser medias las correspondientes variable.

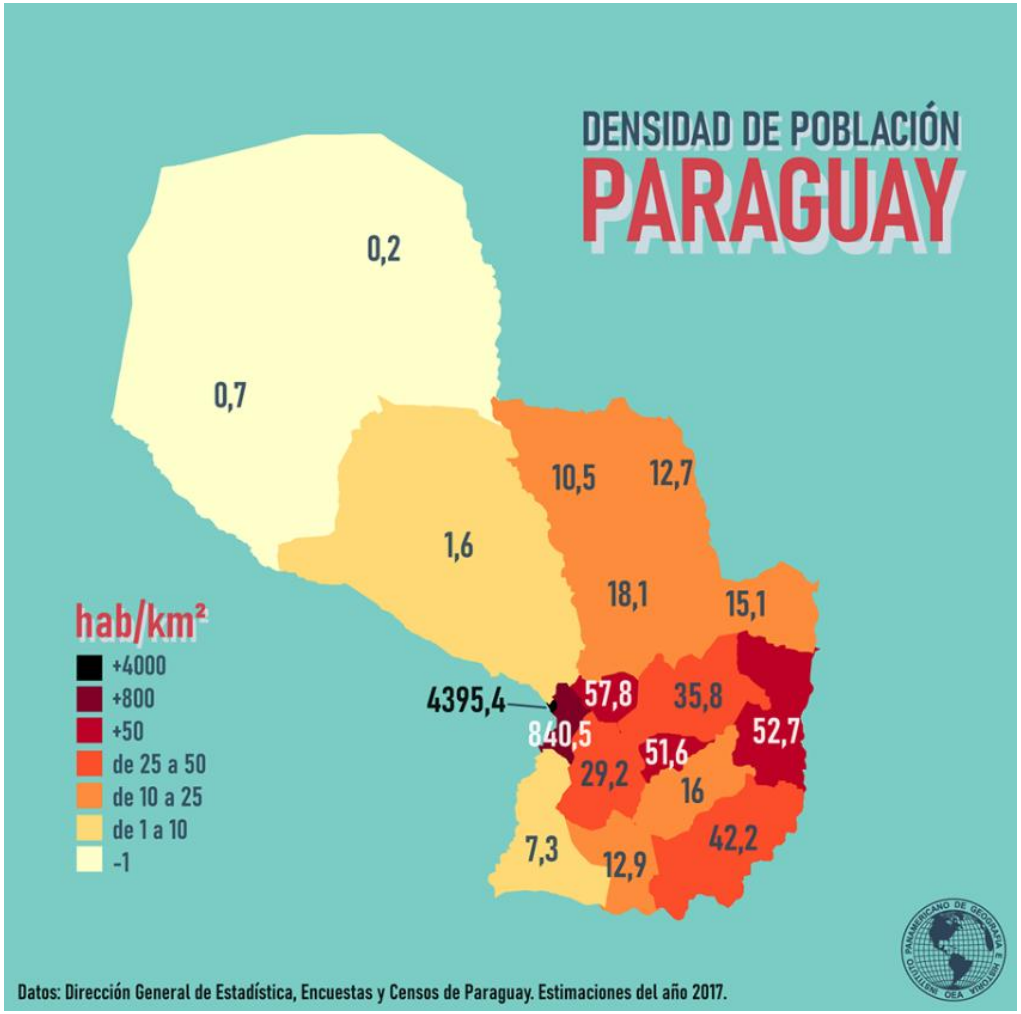
Es importante resaltar que en geoestadística el propósito esencial es la interpolación y si no hay continuidad espacial pueden hacerse predicciones carentes de sentido.



Location of the pollution MS in Madrid and map of predicted Nox leves using geoestatistical techniques

1.2 Datos Lattice

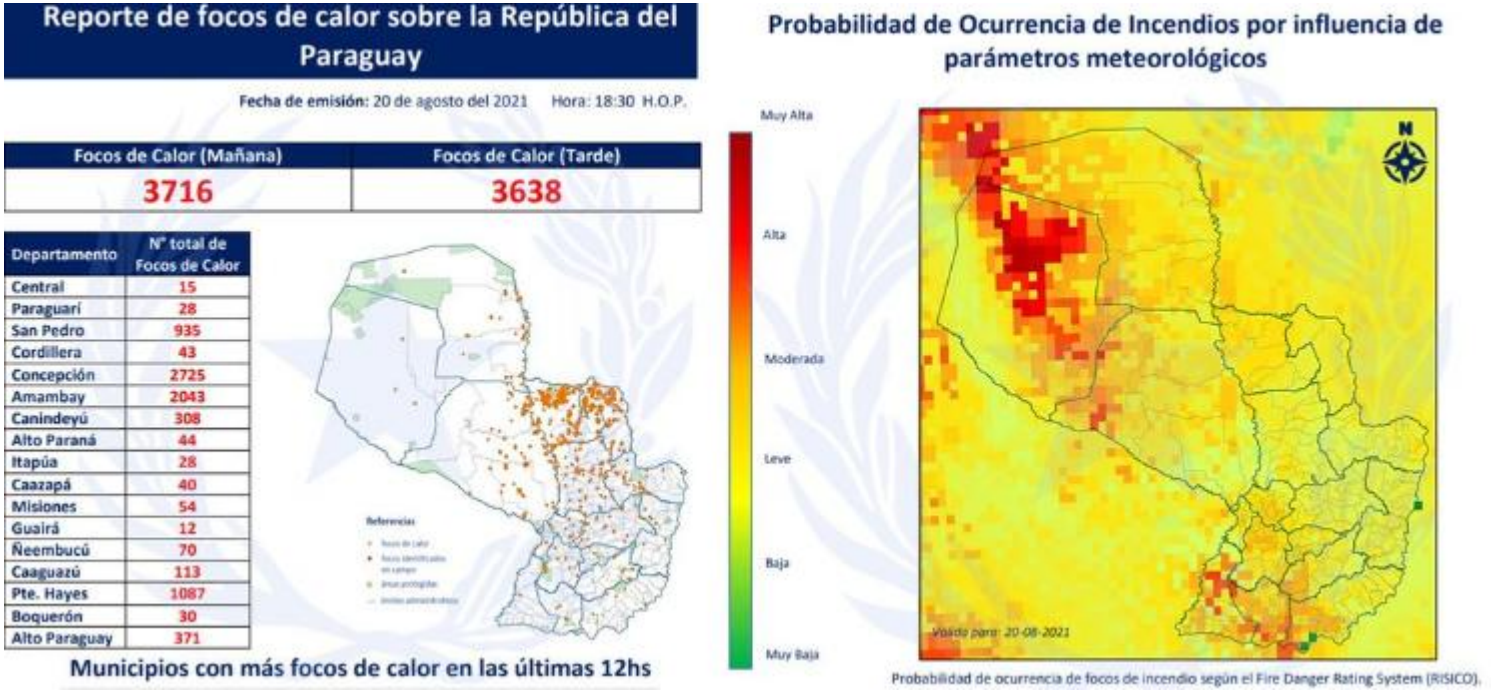
Las ubicaciones **s** pertenecen a un **conjunto D discreto** y son seleccionadas por el investigador (D fijo). Estas pueden estar regular o irregularmente espaciadas. Algunos ejemplos de datos en lattices son los siguientes: la tasa de morbilidad de enfermedades, tasa de accidentes, producción de algún cultivo (rango) de algún departamento o agregación espacial. En los ejemplos anteriores se observa que el conjunto de ubicaciones de interés es discreto y que estas corresponden a agregaciones espaciales más que a un conjunto de puntos del espacio.



Habitantes / km2 en Paraguay

1.3 Patrones espaciales de puntos o patrones espaciales

Patrones Espaciales: las ubicaciones pertenecen a un **conjunto D que puede ser discreto o continuo** y su selección no depende del investigador (**D aleatorio**). Ejemplos de datos dentro de esta área son: Localización de nidos de pájaros en una región dada, ubicación de los sitios de terremoto. Debe notarse que en los ejemplos anteriores hay aleatoriedad en la selección de los sitios. Una vez se ha hecho la selección de sitios es posible hacer medidas de variables aleatorias en cada uno de ellos. Por ejemplo si en primera instancia se establece la ubicación de árboles de pino dentro de un bosque, es posible que sea de interés medir en cada uno de los árboles el diámetro o la altura. En general el propósito de análisis en estos casos es el de determinar si la distribución de los individuos dentro de la región es aleatoria, agregada o uniforme.



Reporte de focos de calor sobre la República del Paraguay

Fecha de emisión: 20 de agosto del 2021 Hora: 18:30 H.O.P.

Focos de Calor (Mañana)	Focos de Calor (Tarde)
3716	3638

Departamento	N° total de Focos de Calor
Central	15
Paraguarí	28
San Pedro	935
Cordillera	43
Concepción	2725
Amambay	2043
Canindeyú	308
Alto Paraná	44
Itapúa	28
Caazapá	40
Misiones	54
Guairá	12
Ñeembucú	70
Caaguazú	113
Pte. Hayes	1087
Boquerón	30
Alto Paraguay	371

Municipios con más focos de calor en las últimas 12hs

Probabilidad de Ocurrencia de Incendios por influencia de parámetros meteorológicos

Muy Alta
Alta
Moderada
Leve
Baja
Muy Baja

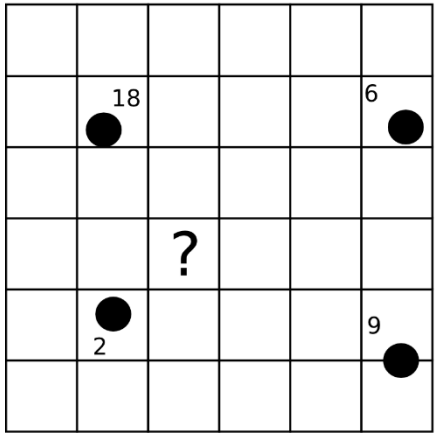
Probabilidad de ocurrencia de focos de incendio según el Fire Danger Rating System (RISICO).

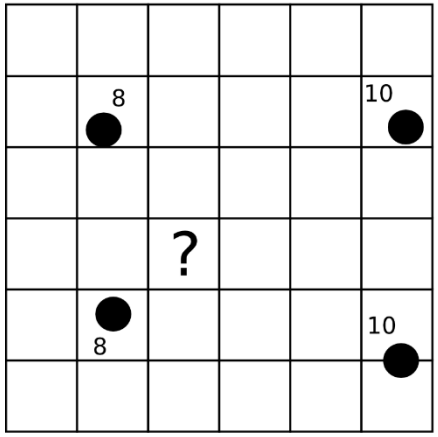
Interpolación

Cálculo de valores en puntos no muestreados, a partir de los valores recogidos en otra serie de puntos.

Clasificación de métodos de interpolación

- *Según los puntos considerados para el cálculo de valores.*
  - Globales: consideran que todos los puntos de los que disponemos tienen influencia sobre el valor a calcular en una celda.
  - Locales: sólo consideran un conjunto restringido de estos. Por umbral de distancia (todos los situados a una distancia menor que el umbral), o por conteo (los n puntos más cercanos), o bien ambos.
- *Según su valor en los puntos de partida.*
  - Exactos: los valores asignados a las coordenadas correspondientes a los puntos de origen son exactamente los recogidos en dichos puntos.
  - Aproximados, el valor en esas celdas es el que corresponde al mejor ajuste, y no ha de coincidir necesariamente con el valor original.
- *Según la inclusión o no de elementos probabilísticos.*
  - Estocásticos: elementos probabilísticos (GEOESTADÍSTICA - KRIGIN)
  - Determinísticos: aquellos que no los emplean (Interpolación Inverse Distance Weighted - IDW)

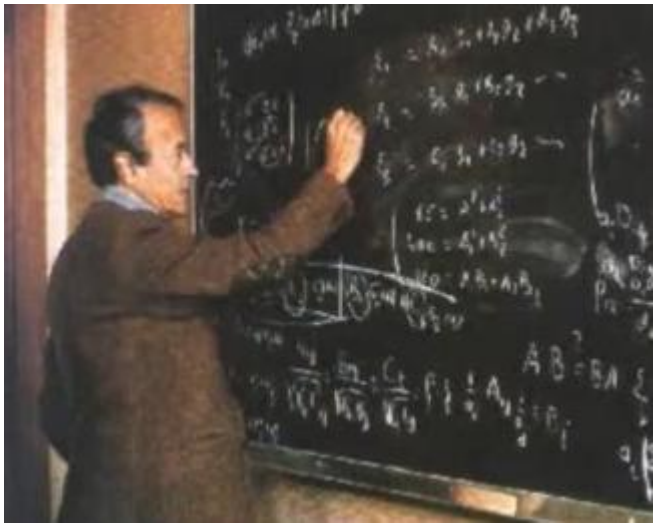




Geoestadística origen

Daniel Krige.
Minas de oro en Sudáfrica

En 1951, Krige expuso su trazador pionero de leyes de oro promedio ponderadas por distancia en el complejo de arrecifes de Witwatersrand.



George Matheron.
Minas de oro en Sudáfrica

La base teoríca fue elaborada por Matheron en los 60s, basado en la tesis de maestría de Danie Krige. Acuñando el termino **Kriging** en honor a Krige

Aplicaciones de la geoestadística

1-) Mapeo o estimación

- Reservas mineras
- Modelos digitales y superficiales del terreno
- Caracterización geoquímica
- Meteorología
- Yacimientos petroleros
- Geomarketing
- Contaminación ambiental
- Epidemiología

2-) Caracterización de la incertidumbre de la estimación

- Muestreo óptimo
- Riesgo ambiental
- Estimación de errores

Geoestadística

Petitgas (1996), la define como una **aplicación de la teoría de probabilidades a la estimación estadística de variables espaciales**. Su interés primordial es la estimación, predicción y simulación de dichos fenómenos. Esta herramienta ofrece una manera de describir la continuidad espacial, que es un rasgo distintivo esencial de muchos fenómenos naturales, y proporciona adaptaciones de las técnicas clásicas de regresión para tomar ventajas de esta continuidad.

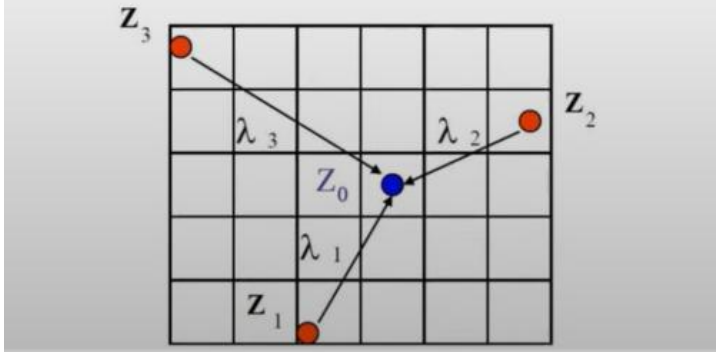
Este es un proceso que **calcula un promedio ponderado de las observaciones muestrales. Los pesos asignados a los valores muestrales son apropiadamente determinados por la estructura espacial de correlación establecida** en la primera etapa y por la configuración de muestreo.

Cuando el objetivo es hacer predicción, **la geoestadística opera básicamente en dos etapas.** La primera es **el análisis estructural**, en la cual se describe la correlación entre puntos en el espacio. En la segunda fase se hace **predicción en sitios de la región no muestreados** por medio de la técnica **kriging.**

Primera ley de la geografía:

Todas las cosas están relacionadas entre sí, pero las cosas más próximas en el espacio tienen una relación mayor que las distantes (Waldo Tobler, 1970)

$$Z_0^* = \sum_{i=1}^{n} \lambda_i Z_i$$



Tobler's First Law of Geography

Tobler's Law #1

Tobler 1

Tobler One

Toblerone

https://www.reddit.com/r/gis/comments/351ai0/frankly_spatial_statistics_lectures_can_get/

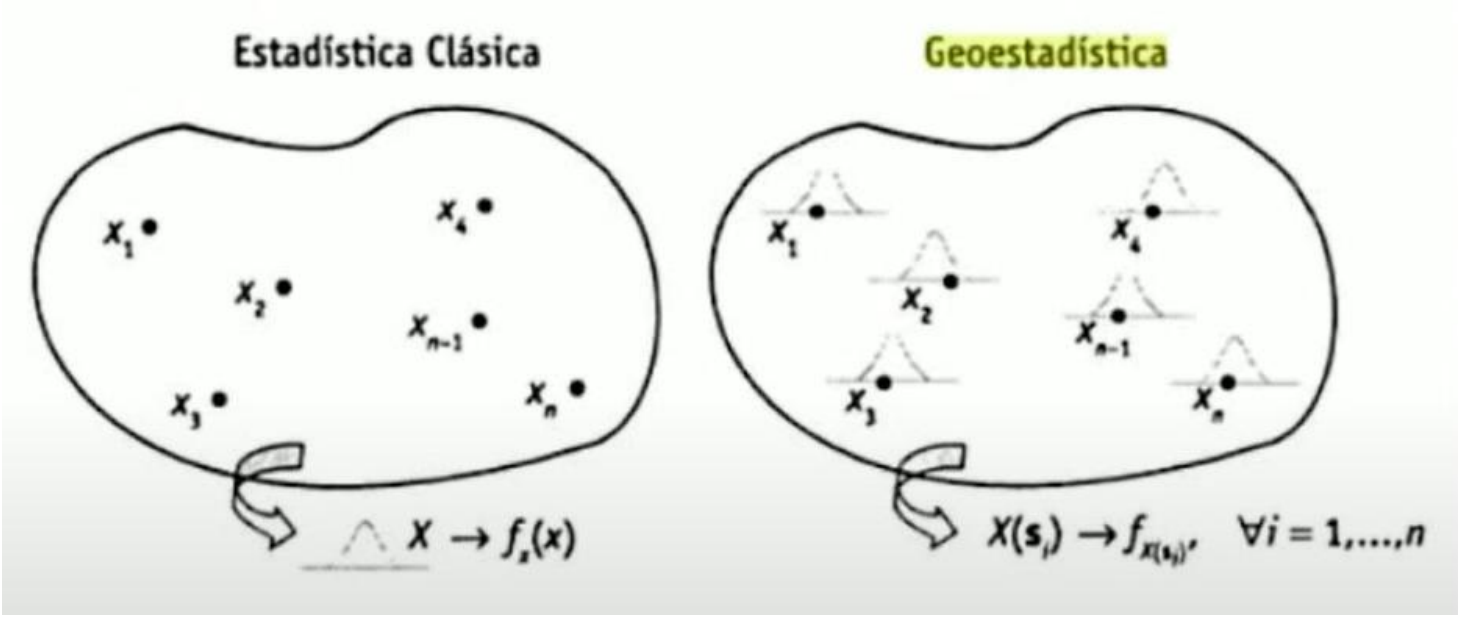
Geoestadística

- **Trabaja con datos continuos**
- **Variable regionalizada** (Una variable medida en el espacio de forma que presente una estructura de correlación, se dice que es una variable regionalizada.)
  - Variable aleatoria
  - Posición (espacio, tiempo)

Objetivos:

1.Caracterizar la variación espacial o estructura espacial

2. Interpolar: predecir valores de una variable en lugares no observados



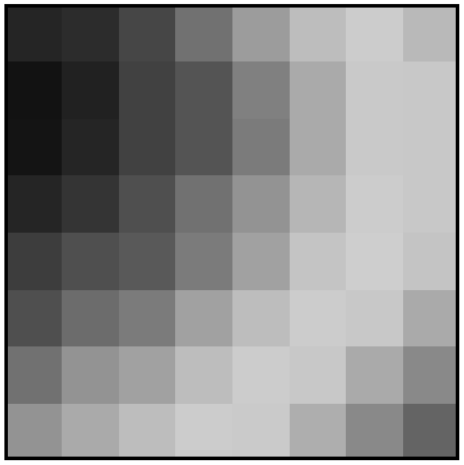
Autocorrelación espacial o dependencia espacial

Correlación de la variable consigo misma, de tal modo que los valores de esta variable en un punto guardan relación directa con los de esa misma variable en otros puntos cercanos (Tobler)

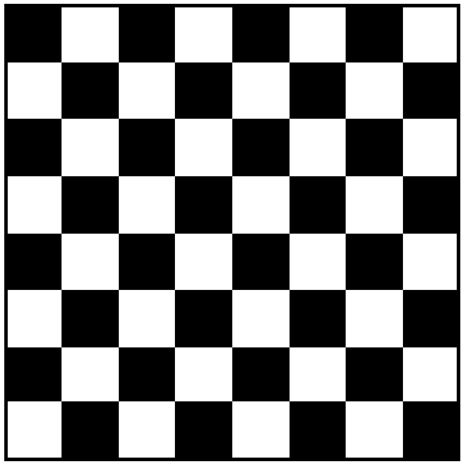
Tipos

a) Positiva: los valores altos suelen tener en su entorno valores también altos
b) Negativa: si los valores altos se rodean de valores bajos y viceversa.
c) Nula: sin autocorrelación espacial, independencia

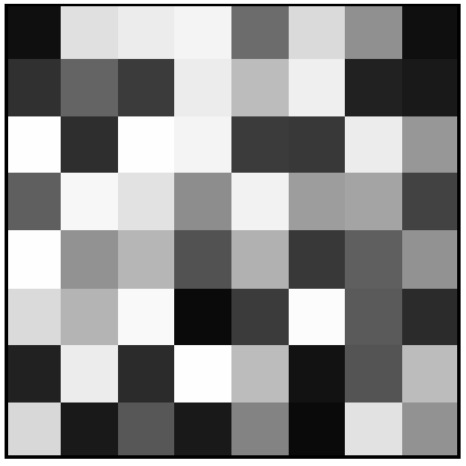
a)



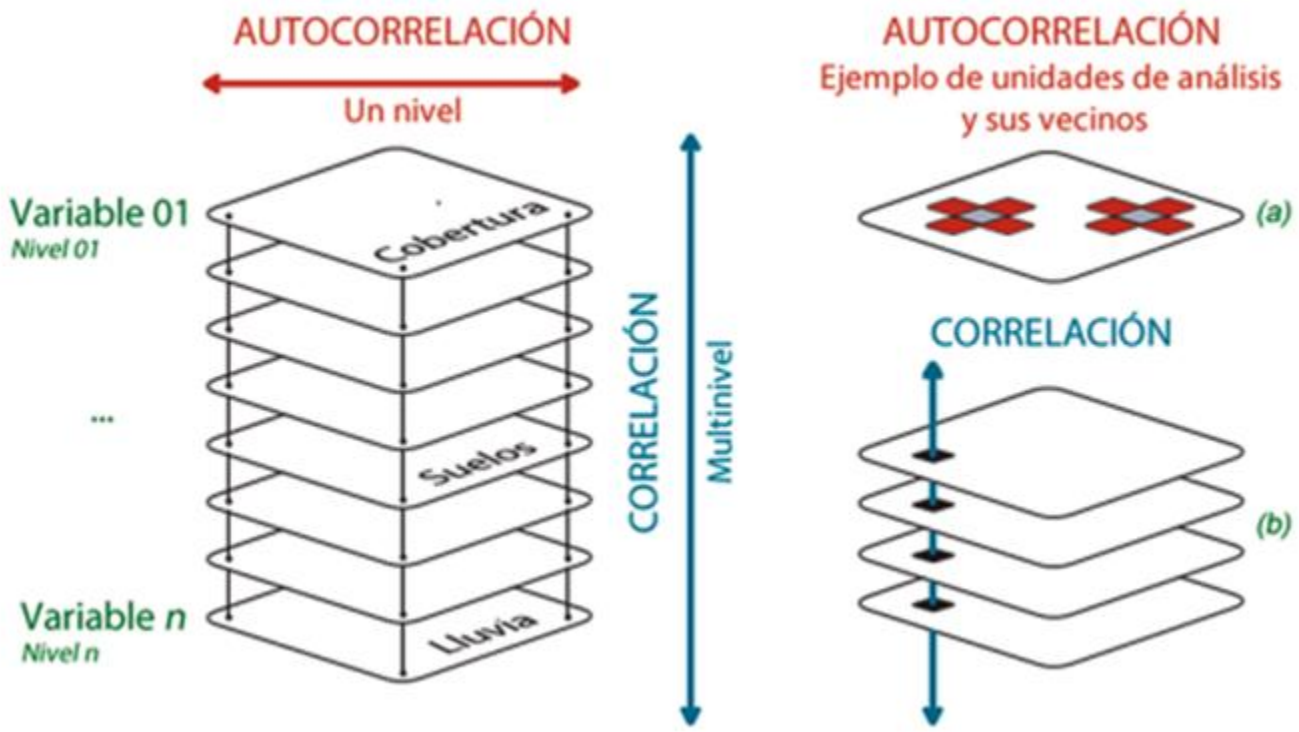
b)



c)



Autocorrelación espacial o dependencia espacial



Estacionariedad

- La estacionariedad significa que las características de una función aleatoria (nuestro proceso espacial que genera una variable) permanecen iguales cuando se desplaza un conjunto dado de n puntos de una parte de una región a otra. El proceso espacial se llama **estrictamente estacionario** si para cualquier conjunto de **n** puntos **s1,...,sn** la distribución multivariada no cambia. En otras palabras, todo se ve igual en todas partes y en cualquier momento. Bueno, esto puede no parecer una suposición razonable para los procesos geográficos.

- Una condición menos restrictiva viene dada por la **estacionariedad débil**, que también se denomina estacionariedad de segundo orden porque solo asume la estacionariedad de los dos primeros momentos de la distribución. Es decir, el proceso espacial tiene una media constante E[Z(x)] y una varianza Var[Z(x)] . Además, **la covarianza entre dos observaciones** separadas por una distancia h : cov(Z(x+h),Z(x)) **solo se basa en la distancia h entre las observaciones y no en la ubicación espacial x dentro de la región.**

- Un tipo específico de estacionariedad de segundo orden, y **más importante para el análisis de variogramas, se denomina estacionariedad intrínseca**. Aquí, asumimos la estacionariedad de segundo orden de las diferencias entre pares de valores en dos ubicaciones: Z(x+h)−Z(x) . **No nos interesa Z(x) sino las diferencias. Nuevamente, esto implica la suposición de que la varianza de estas diferencias no depende de la ubicación sino solo de la distancia de separación h:**

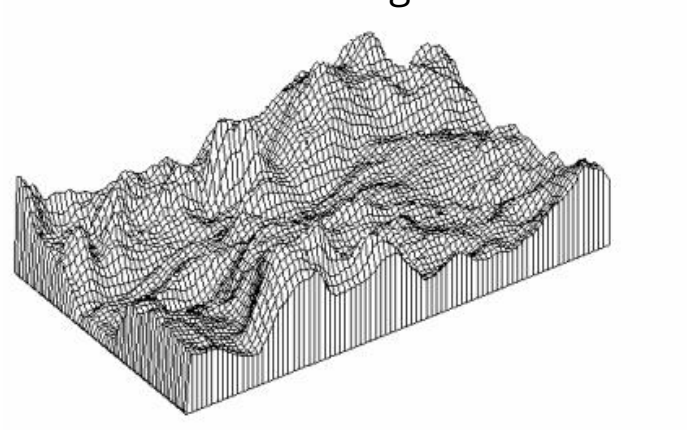
$$Var(Z(x+h) - Z(x)) = 2\gamma(h)$$

Estacionariedad

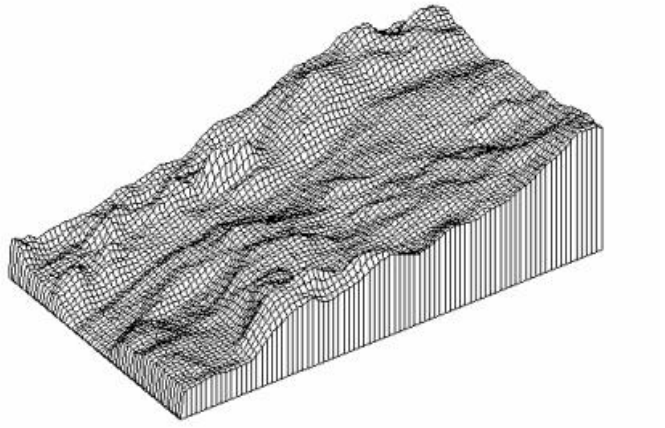
Un supuesto fundamental en el análisis geoestadístico es que el fenómeno es estacionario, para lo cual, entre otros aspectos, **el nivel promedio de la variable debe ser constante en todos los puntos del área de estudio**. La variable regionalizada es estacionaria si su función de distribución conjunta es invariante respecto a cualquier translación del vector h (distancia).

Una propiedad de un proceso espacial en el que la covarianza entre los valores de dos puntos depende únicamente de la distancia entre ellos. No existe tendencia espacial.

Una detección de tendencia en el gráfico de dispersión puede ser una muestra de que no se satisface dicho supuesto. El gráfico se construye tomando como eje de las abcisas la variable que representa la coordenada geográfica y en el eje de las ordenadas la variable cuantitativa de estudio. En la figura se muestra el grafico de una variable regionalizada estacionaria. Exceptuando fluctuaciones aleatorias, el valor promedio de la variable no muestra una tendencia definida en alguna dirección.

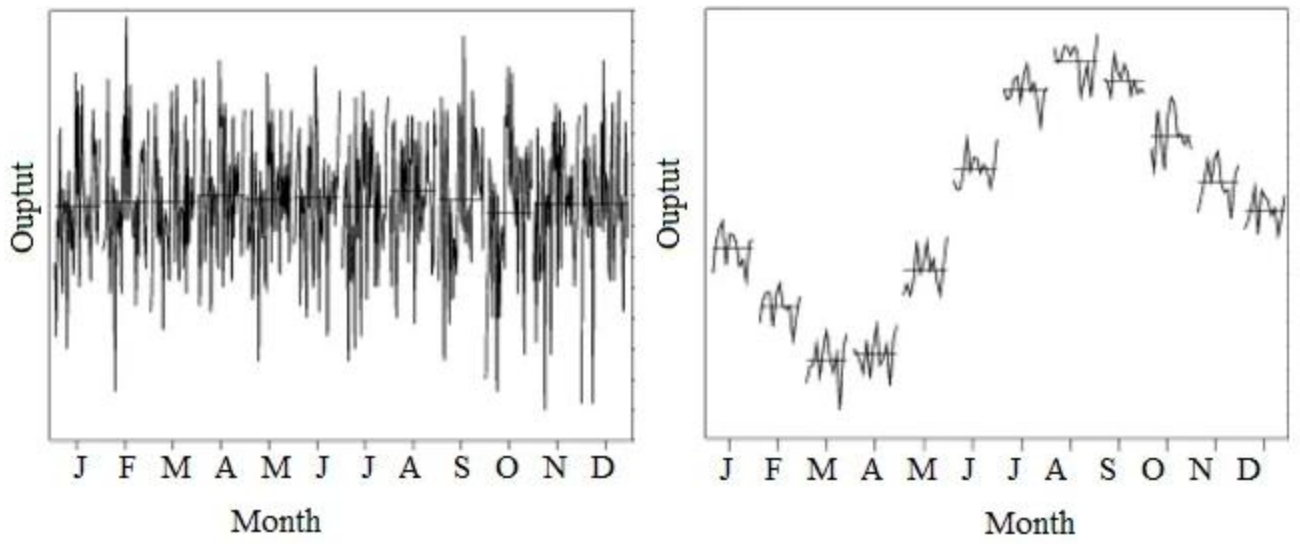


Variable estacionaria



Variable no estacionaria

Estacionariedad

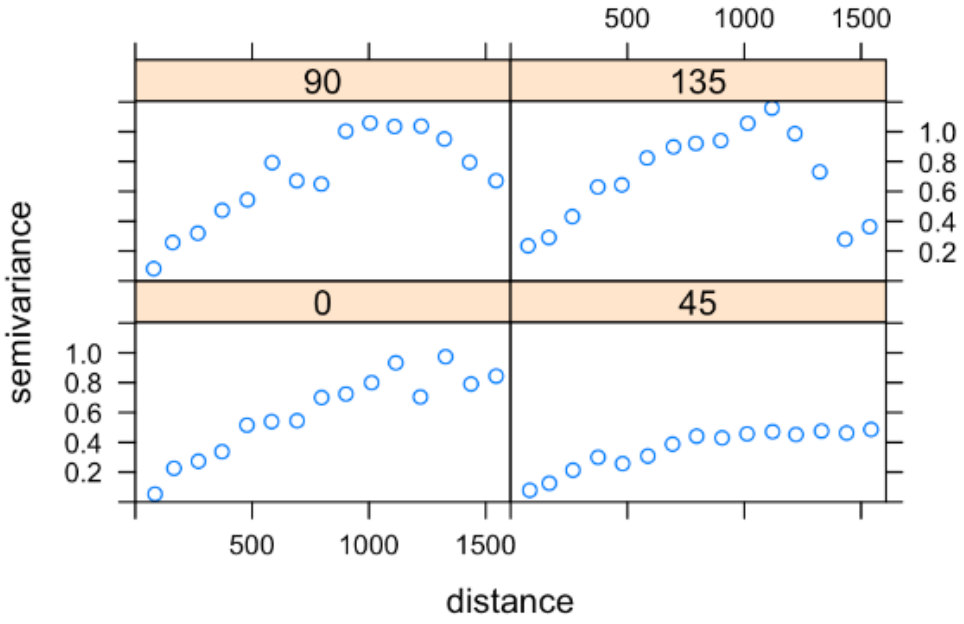


El gráfico de la izquierda es estacionario sin una tendencia evidente, mientras que el gráfico de la derecha muestra estacionalidad y no es estacionario.

Isotropía y Anisotropía

Una propiedad de un proceso natural o datos donde la dependencia espacial (autocorrelación) cambia solo con la distancia entre dos ubicaciones:l a dirección no es importante. La anisotropía ocurre cuando la dependencia espacial (autocorrelación) cambia tanto con la distancia como con la dirección entre dos ubicaciones.

La isotropía es estudiada a través del cálculo de funciones de autocovarianza o de semivarianza muestrales en varias direcciones. Si estas tienen formas considerablemente distintas puede no ser válido el supuesto de isotropía.



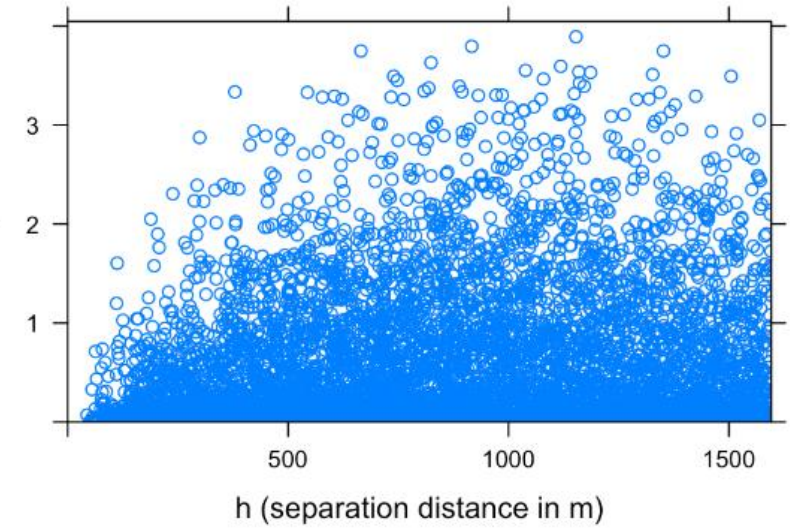
Análisis estructural: Variograma

La primera etapa en el desarrollo de un análisis geoestadístico es la determinación de la dependencia espacial entre los datos medidos de una variable. Esta fase es también conocida como análisis estructural. Para llevarla a cabo, con base en la información muestral, se usan tres funciones: **El semivariograma, el covariograma y el correlograma.**

Podemos trazar la diferencia entre pares de valores de muestra en función de la distancia de separación **h**. La llamada nube de variograma traza todas las posibles diferencias al cuadrado de los pares de observación $(Z(x) - Z(x+h))^2$ frente a su distancia de separación **h**. La diferencia se divide por 2 para tener en cuenta el hecho de que dos puntos comparten este valor. Por lo tanto, **y** a menudo se denomina **semivarianza** (la mitad de la varianza).

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

Dado que **y** es una diferencia, **es una medida de disimilitud, es decir, cuanto menor es la disimilitud, más similares son las observaciones**. Por lo tanto, l**a nube de variograma le permite observar si los pares de muestras más cercanos entre sí son más similares que los pares más separados**, lo que suele ser el caso. La nube de variograma también le permite observar la distribución de la disimilitud a distancias particulares. Por ejemplo, la nube de variograma a la derecha muestra que y tiene una distribución sesgada a cualquier distancia. Es decir, la mayoría de las parcelas son similares incluso a distancias mayores. Sin embargo, la disimilitud aumenta con distancias de hasta 500 mo más o menos.

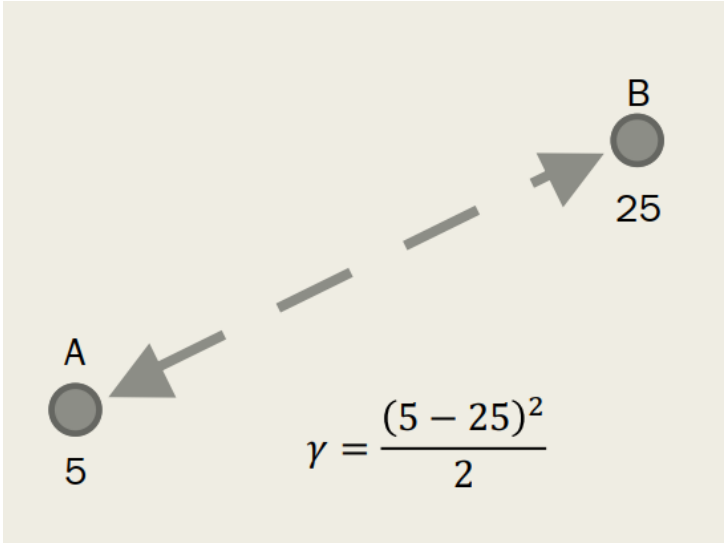


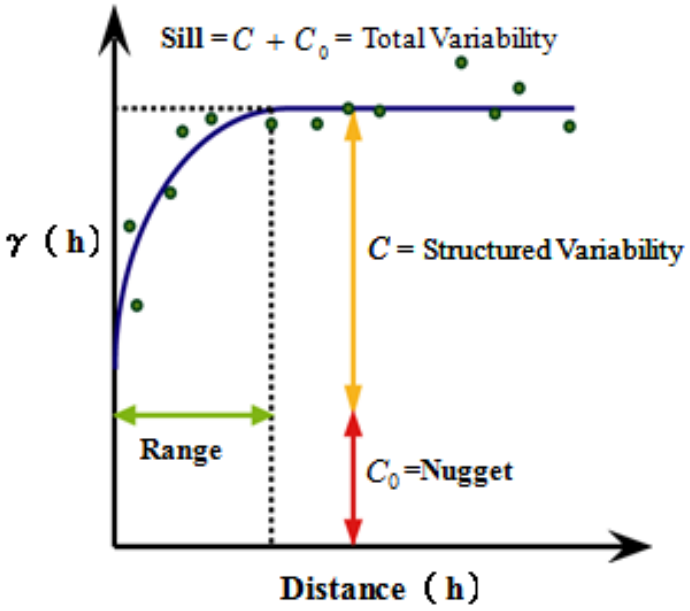
Análisis estructural: Semivariograma empírico o variograma muestreado

$$\bar{\gamma}(h) = \frac{\sum (Z(x+h) - Z(x))^2}{2n}$$

Una función de la distancia que separa dos ubicaciones que se usa para cuantificar la dependencia. Se calcula como la mitad de la diferencia cuadrada entre dos valores de datos en dos ubicaciones.

El semivariograma generalmente aumenta con la distancia antes de volverse plano y se describe mediante parámetros de nugget, umbral y rango.

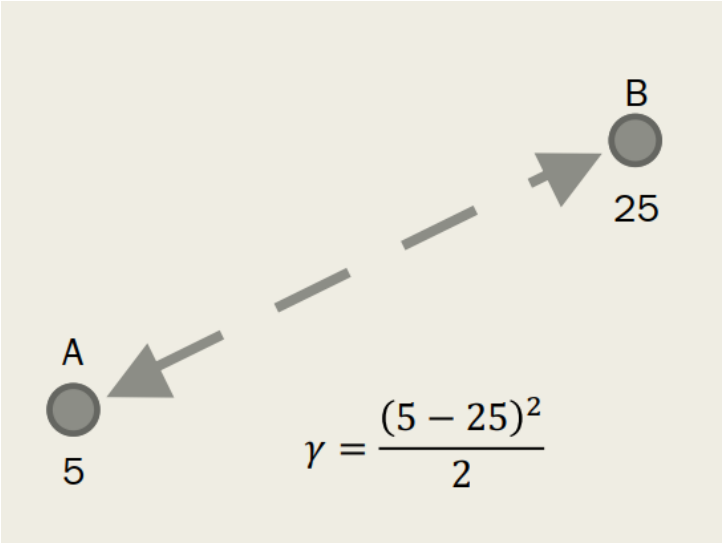




Análisis estructural: Semivariograma empírico o variograma muestreado

Se fundamenta en el concepto de semivarianza

La mitad del variograma γ(h), se conoce como la función de semivarianza y caracteriza las propiedades de dependencia espacial del proceso. Dada una realización del fenómeno, la función de semivarianza es estimada, por el método de momentos, a través del semivariograma experimental, que se calcula mediante



$$\bar{\gamma}(h) = \frac{\sum (Z(x+h) - Z(x))^2}{2n}$$

donde $Z(x)$ es el valor de la variable en un sitio x, $Z(x+h)$ es otro valor muestral separado del anterior por una distancia h y n es el número de parejas que se encuentran separadas por dicha distancia. La función de semivarianza se calcula para varias distancia h. En la práctica, debido a irregularidad en el muestreo y por ende en las distancias entre los sitios, se toman intervalos de distancia $\{[0, h], (h, 2h], (2h, 3h], \cdots\}$ y el semivariograma experimental corresponde a una distancia promedio entre parejas de sitios dentro de cada intervalo y no a una distancia h específica. Obviamente el número de parejas de puntos n dentro de los intervalos no es constante.

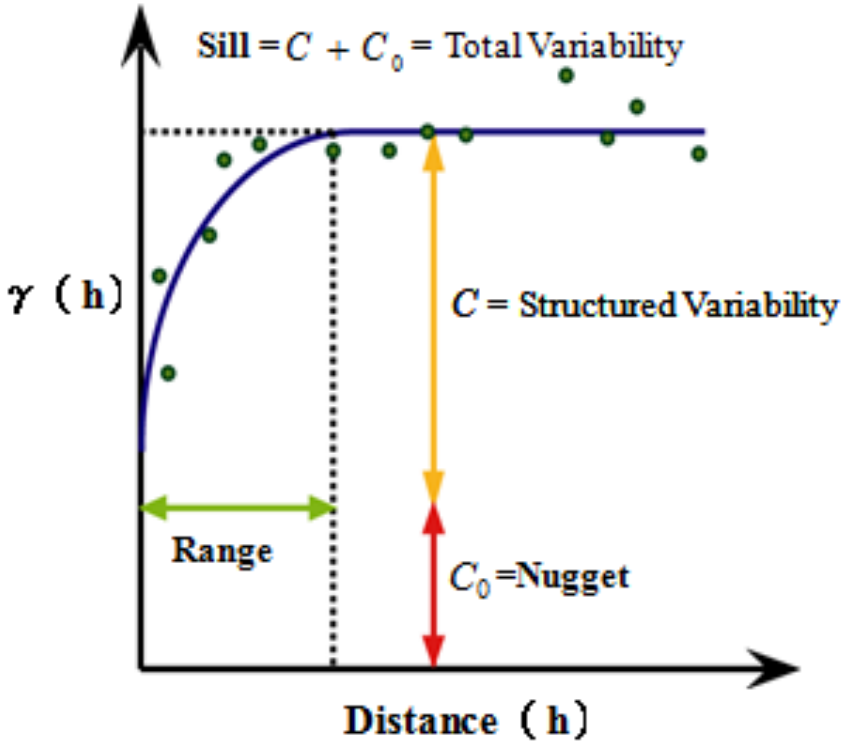
Para interpretar el semivariograma experimental se parte del criterio de que a menor distancia entre los sitios mayor similitud o correlación espacial entre las observaciones. Por ello en presencia de autocorrelación se espera que para valores de h pequeños el

Análisis estructural: Semivariograma teoríco

Rango. máxima distancia hasta la cual existe dependencia espacial. Es el valor en el que se alcanza la máxima varianza, o a partir del cual ya presenta una tendencia asintótica.
Sill o meseta. El máximo valor del variograma. Representa la máxima variabilidad en ausencia de dependencia espacial.
Nugget o pepita. Conforme la distancia tiende a cero, el valor de la semivarianza tiende a este valor. Representa una variabilidad que no puede explicarse mediante la estructura espacial (ruido o error de medición).

La parte más importante del trabajo de la Kriging, es la de ajustar el semivariograma empírico a un semivariograma teórico que es utilizado dentro de la fórmula de Kriging para modelar espacialmente la variable.

Se modela dentro del rango.



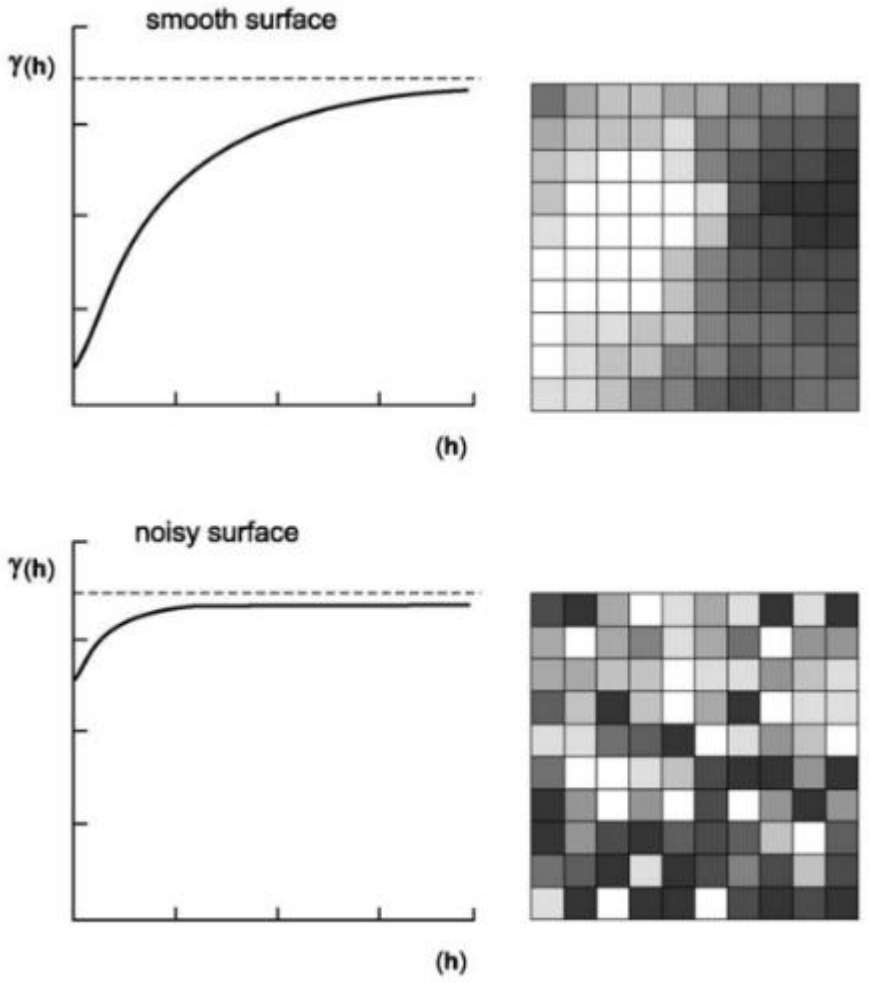
X= grupos de distancias

Y= semivarianza media

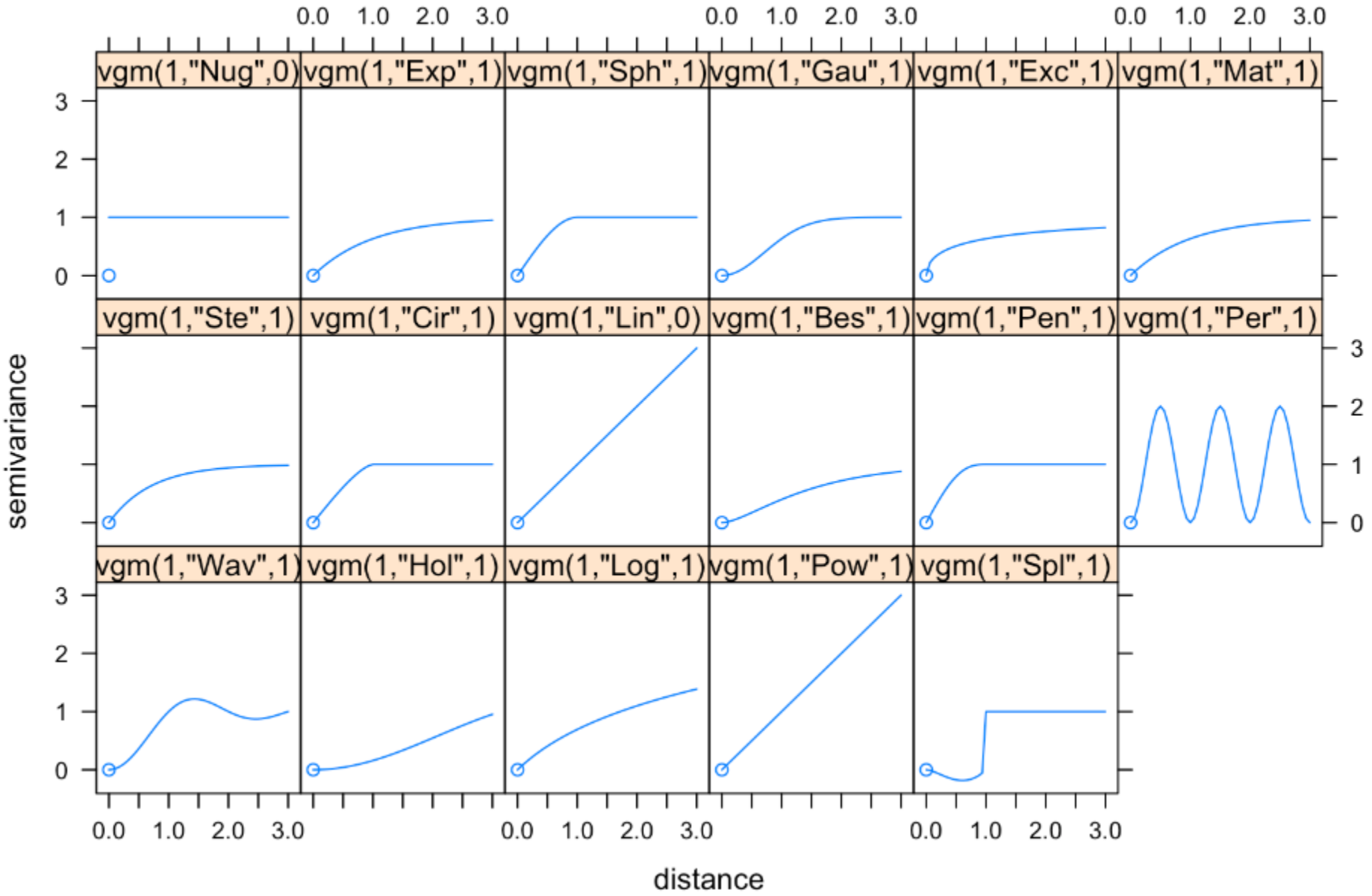
Análisis estructural: Semivariograma teoríco

En general dichos modelos pueden dividirse en no acotados (lineal, logarítmico, potencial) y acotados (esférico, exponencial, gaussiano) (Warrick *et al*., 1986).

Los del segundo grupo garantizan que la covarianza de los incrementos es finita, por lo cual son ampliamente usados cuando hay evidencia de que presentan buen ajuste.



Tipos de semivariograma teórico



Tipos Kriging

Media constante

simple kriging = media conocida
ordinary kriging = media desconocida
indicator kriging = datos categoricos / above - under threshold

Tendencia en la media

- **co-kriging** = covariables, solo disponible en ciertos puntos
- **universal kriging** = tendencia en función de las coordenadas, covariables en todos los puntos de predicción. deterministic part is modeled as a function of coordinates
- **kriging with external drift=** tendencia en función de covariables, covariables en todos los puntos de predicción. both components are predicted simultaneously.
- **Regresión kriging =** se maneja por separado la componente determinística y estocástica. the deterministic (regression) and stochastic (kriging) predictions are done separately

Elección del método

Las características de la variable a interpolar.

1. valores de precipitación máxima anual, no es adecuado utilizar aquellos métodos que suavicen excesivamente la superficie resultante, ya que se estarían perdiendo los valores extremos que, por la naturaleza del valor interpolado, son de gran interés.

Las características de la superficie a interpolar.

1. variaciones bruscas en puntos de discontinuidad tales como acantilados en el caso de interpolar elevaciones, son aplicables mediante la imposición de barreras con métodos como el de distancia inversa, pero no con otros como el kriging.

La calidad de los datos de partida.

1. los datos de partida son de gran precisión → los métodos exactos porque preservan la información original.
2. datos de partida contienen mucho ruido → Kriging que suavizan el resultado y el efecto de ruido es preferible.

El rendimiento de los algoritmos.

1. basados en distancia son rápidos y requieren un tiempo de proceso aceptable
2. kriging, más complejos y el tiempo de proceso es elevado.

El conocimiento de los métodos.

2 Bibliografía y materiales de consulta

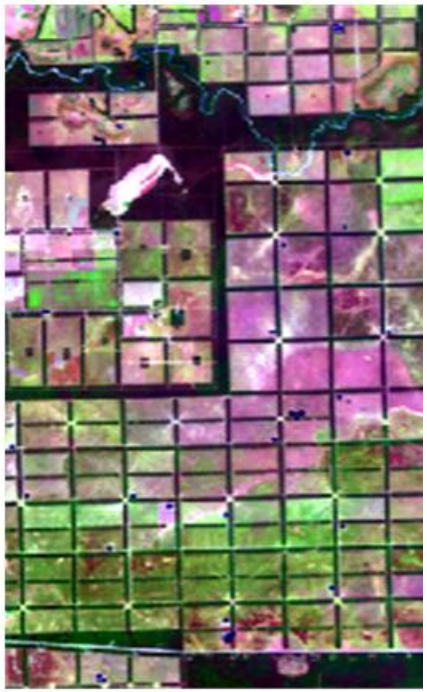
- Humboldt-Universität zu Berlin. Department of Geography. Spatial interpolation in R. https://pages.cms.hu-berlin.de/EOL/gcg_quantitative-methods/Lab14_Kriging.html#Overview
- Ballari Daniela, Conceptos de interpolación y geoestadística (presentación). https://rstudio-pubs-static.s3.amazonaws.com/416462_1463d00750c54fce89955a925eaa4957.html
- Giraldo R, 2021. Introducción a la geoestadística. https://geoinnova.org/wp-content/uploads/2021/08/LIBRO_-DE-_GEOESTADISTICA-R-Giraldo.pdf
- Fernández, R. 2003. Geoestadística espacio-temporal. https://rubenfcasal.github.io/files/Geoestadistica_espacio-temporal.pdf
- Pebesma, E. The meuse data set: a brief tutorial for the gstat R package. https://cran.r-project.org/web/packages/gstat/vignettes/gstat.pdf
- Fernández-Avilés, G. 2021. Taller introducción a la geoestadística con R. UMUR. https://www.youtube.com/watch?v=iAQD-qaEy5Y&list=PLcKa2WC0f2Cq36vOSlJARv1uvoq9Xdrcd&index=2&t=2726s

Muchas gracias!!

Carlos Giménez Larrosa
Correo: charlieswall@gmail.com

UNIVERSIDAD NACIONAL DE ASUNCIÓN

Facultad Politécnica



Diplomado en Computación Estadística y Análisis de Datos

Estructura y manejo de datos espaciales en R