

Introducción al análisis geoestadístico de datos en geociencias: teoría y aplicación

Introduction to geostatistical analysis of data in geosciences: theory and application

Maximiliano Garnier-Villarreal

Universidad de Costa Rica (UCR), Escuela Centroamericana de Geología (ECG), San José,
Costa Rica

maximiliano.garniervillarreal@ucr.ac.cr

(Recibido: 10/08/2021; aceptado: 12/01/2022)

RESUMEN: Kriging, el método de interpolación asociado a geoestadística, se ha usado y ha sido propuesto como el mejor método de interpolación, muchas veces sin realmente entender cómo es que se usa adecuadamente y dejando que el software que lo brinda decida cómo implementarlo. Esta aseveración tiene fundamento cuando se procede de la manera correcta, realizando los pasos necesarios durante el análisis y modelado geoestadístico, por lo que es necesario entender cómo aplicar Kriging correctamente para que los resultados obtenidos sean relevantes y confiables. Estos pasos se detallan en este trabajo, abordando la teoría, y mediante un ejemplo, con datos de la temperatura promedio de los últimos 10 para el 8 de Marzo para la provincia de San José, se pone en práctica el método usando el software estadístico libre R. Adicionalmente, se presenta una aplicación web de libre acceso para quienes no se sientan cómodos usando lenguajes de programación.

Palabras clave: geoestadística; kriging; R; variograma; interpolación; validación cruzada.

ABSTRACT: Kriging, the interpolation method linked to geostatistics, has been used and has been proposed as the best interpolation method, a lot of times without really understanding how to properly apply it and allowing the software that offers it to decide how to implement it. This statement is correct when used in the right way, performing all the necessary steps throughout the geostatistical analysis and modeling, for this reason is necessary to understand how to apply Kriging correctly so the results are relevant and reliable. These steps are detailed in this work, going over the theory, and through a practical example, using average temperature data over the last 10 years for March 8th for the San José province, the method is implemented using the free statistical software R. Additionally, a free access web application is shown for those who are not comfortable using programming languages.

Keywords: geostatistics; kriging; R; variogram; interpolation; cross-validation.



Introducción

En las ciencias que tienen una fuerte componente espacial (dentro de ellas geología) es común recolectar muestras, describirlas y tener la ubicación de dónde se recolectaron. En muchos casos el muestreo se hace con el fin de caracterizar una variable o proceso en el espacio, con lo que se tiene en mente pasar de puntos a una superficie (mapa).

Para poder generar estas superficies se pueden emplear diferentes Métodos de interpolación, donde comúnmente se ha dado a entender que Kriging es el método por excelencia a usar (casi que indiscriminadamente), pero muchos usuarios no saben cómo usar el método apropiadamente y cuándo es adecuado o no utilizarlo.

La facilidad que brindan programas de cómputo comerciales (Surfer, ArcGIS), con sus interfaces “*point y click*”, de implementar éste y otros métodos hace creer al usuario que es simplemente de escoger un método y decirle que lo ejecute, sin guiar al usuario de manera apropiada en el proceso necesario para obtener resultados significativos, confiables, y reproducibles. Cabe mencionar que el procedimiento y los pasos que se van a mostrar acá están disponibles en estos softwares pero no de manera frontal para el usuario.

El objetivo principal de este trabajo es de índole educativo/informativo y corresponde con introducir al lector en qué es la geoestadística y cómo realizar un análisis geoestadístico básico de manera apropiada. La idea es brindar una base y guía de cómo hacer una interpolación de los datos de interés y en español, ya que la mayoría de los textos (libros y artículos) están en inglés, y a veces se enfocan únicamente en los resultados (mapas) y no tanto en el proceso.

Para el procesamiento de los datos y la implementación de la geoestadística se va a utilizar el software estadístico libre multi-plataforma **R** (R Core Team, 2021) así como diferentes paquetes, que va a permitir el desarrollo de rutinas que se pueden reutilizar para análisis futuros. Adicionalmente se presentará una aplicación web, desarrollada en **R** y de libre acceso, que hace uso de lo expuesto aquí. Se recomienda al lector, si no está familiarizado con **R** o quiere profundizar más en su uso, consultar Garnier-Villarreal (2020).

Métodos de interpolación

De manera resumida y sin entrar en mucho detalle se mencionan diferentes métodos de interpolación comúnmente usados, para ellos se puede consultar Webster y Oliver (2007). De manera general se tienen: Polígonos de Thiessen, Triangulación, Vecinos naturales (*natural neighbours*), Inverso de la distancia (*inverse distance*), Superficies de tendencia (*trend surface*), Ajuste polinomial (*splines*), y **Kriging**.

El método de Kriging es lo que más se asocia con la geoestadística, y va a ser el énfasis de lo aquí presentado. El Kriging es considerado como el método más robusto y preciso, de ahí que en inglés es conocido como *blue* que quiere decir *best linear unbiased estimator*, y se puede traducir como **mejor estimador lineal no-sesgado** (Isaaks y Srivastava, 1989; Webster y Oliver, 2007).

Una ventaja de Kriging con respecto a otros métodos de interpolación más populares, es que a parte de estimar el valor de la variable de interés, estima además un error de la interpolación, lo que permite tener una idea de la calidad (incertidumbre) de los resultados (Isaaks y Srivastava, 1989; Webster y Oliver, 2007). El método ha sido utilizado para predecir la intensidad sísmica (Linkimer, 2008), el nivel de agua subterránea (Varouchakis, Hristopulos, y Karatzas, 2012), pérdida de suelo (G. Wang, Gertner, Fang, y Anderson, 2003), y temperatura del aire (Wang et al., 2017), entre otras.

Geoestadística

La geoestadística no es estadística (clásica) aplicada a datos geológicos, es un tipo de estadística que hace uso de la componente espacial de los datos y pretende caracterizar sistemas distribuidos en el espacio los cuales no se conocen por completo (Davis, 2002; Isaaks y Srivastava, 1989; Webster y Oliver, 2007). Hay que resaltar que Kriging es un método de interpolación (uno de los usos de la geoestadística) que corresponde con uno de los pasos en el análisis y modelado geoestadístico (Oliver y Webster, 2014), no hay que confundir o pensar que geoestadística es lo mismo que Kriging, que es un error común.

La geoestadística (Kriging) se ha utilizado más para la interpolación (estimación - Kriging) de variables en el espacio, pero también se puede utilizar para la simulación (Simulación Gaussiana Secuencial) de la variable de interés (otra forma de usar geoestadística que no es Kriging). El resultado de la interpolación es la distribución del valor promedio de la variable (cuál sería el valor más probable de encontrar), la simulación genera una cantidad definida de realizaciones (N) de la variable, que pueden estar condicionadas o no a datos observados, y presentan una distribución más heterogénea que la interpolación (Chilès y Delfiner, 1999; Goovaerts, 1997; Pebesma y Bivand, 2020; Pyrcz y Deutsch, 2014; Webster y Oliver, 2007). En este trabajo el enfoque va a ser en el uso más común y sencillo que es la interpolación (estimación) de una variable en el espacio.

La base de lo que se va a exponer corresponde con capítulos de Davis (2002), Swan y Sandilands (1995), Borradaile (2003), y McKillup y Darby Dyar (2010), y textos más detallados y exclusivos en la materia de Chilès y Delfiner (1999), Cressie (1993), Goovaerts (1997), Isaaks y Srivastava (1989), Pyrcz y Deutsch (2014), Webster y Oliver (2007), y Wackernagel (2003), los cuales corresponden con referencias clásicas y actualizadas. Para la implementación en **R** y más base teórica y práctica se puede consultar Nowosad (2019) y E. Pebesma y Bivand (2020).

A continuación se definen algunos conceptos fundamentales en geoestadística, que forman las bases teórica y práctica para el análisis geoestadístico.

Correlación espacial

El concepto fundamental en geoestadística y la estadística espacial en general, es que las observaciones son inter-dependientes en función de la distancia entre ellas, donde hay más similitud (relación) conforme más cercanas estén las observaciones y esa similitud o relación es más débil conforme la distancia incrementa (Chilès y Delfiner, 1999; Cressie, 1993; Goovaerts, 1997; Isaaks y Srivastava, 1989; Webster y Oliver, 2007).

Semivarianza

Esta es la medida que se usa para determinar la disimilitud (relación) entre observaciones que varían con la distancia, y se representa mediante la Ecuación (1), donde $Z(x_i)$ es el valor de la variable en la posición x_i , $Z(x_i+h)$ es el valor de la variable a una distancia h , N es el número total de puntos (observaciones), y $N(h)$ es el número de pares de puntos que se encuentran a una distancia h específica. **Se recomienda tener más de 30 pares de puntos por cada distancia h , y no calcular la semivarianza más allá de la mitad de la máxima distancia entre observaciones** (Chilès y Delfiner, 1999; Goovaerts, 1997; Isaaks y Srivastava, 1989; Webster y Oliver, 2007).

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i + h) - Z(x_i)]^2 \quad (1)$$

Si los datos se encuentran ordenados en una grilla regular se puede usar la separación entre puntos como las diferentes distancias h (Fig. 1 (a) y (b)). Si los datos se encuentran irregularmente espaciados es necesario agruparlos en franjas (Fig. 1 (c)), donde se requiere definir una tolerancia de la distancia (w , por lo general $h/2$), y una tolerancia angular ($\alpha/2$) (Oliver y Webster, 2014; Webster y Oliver, 2007).

Variograma experimental

Para visualizar la relación (o no) entre la semivarianza y la distancia (relación espacial de la variable) se usa el variograma experimental (Fig. 2).

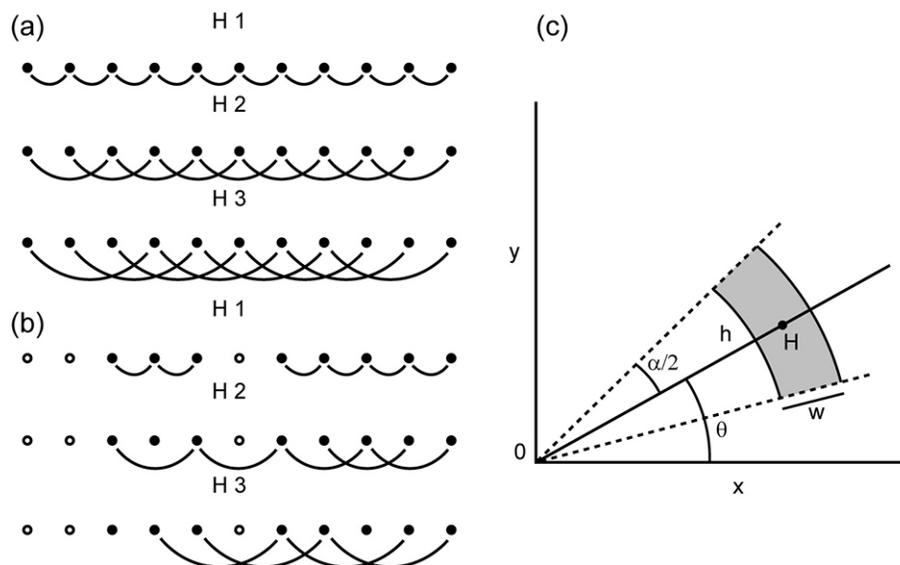


Fig. 1: Esquema del cálculo de la semivarianza para datos regularmente espaciados, donde los datos están completos (a) y donde hay datos faltantes (b); para datos irregularmente espaciados (c). Modificado de Webster y Oliver (2007).

El cálculo de la semivarianza y su representación por medio del variograma experimental son los primeros pasos donde el usuario/analista tiene control sobre la construcción y representación de la relación espacial de la variable, y el resultado va a ser el insumo para pasos posteriores. Como decisiones fundamentales se tienen la escogencia de la distancia máxima y el intervalo de distancias (h). Conforme se varíen estos valores va a variar la semivarianza, cualquier ajuste que se le realice, y su posterior uso en la interpolación (Isaaks y Srivastava, 1989; Oliver y Webster, 2014; Webster y Oliver, 2007).

Modelo de variograma

El variograma experimental es una representación discreta de la relación espacial ya que se cuenta solo con puntos a las distancias definidas. Para poder interpolar valores a diferentes distancias es necesario tener un modelo continuo que se ajuste a los datos. Para ajustar un modelo hay que analizar el variograma experimental y realizar una estimación inicial de las partes o parámetros que lo van a definir (Goovaerts, 1997; Sarma, 2009; Webster y Oliver, 2007).

Partes

Las partes o parámetros que definen a un modelo de variograma se muestran en la figura 3, y son (Isaaks y Srivastava, 1989; Sarma, 2009; Webster y Oliver, 2007):

Meseta total (S , *sill* en inglés): Valor del variograma o semivarianza cuando la distancia h tiende a infinito (cuando la semivarianza se estabiliza), y por lo general es muy similar al valor de la varianza de la variable de interés.

Meseta parcial (C_1 , *partial sill* en inglés): La diferencia entre la meseta total y la pepita ($C_1 = S - C_0$). Si no hubiera pepita ($C_0 = 0$), entonces $C_1 = S$.

Pepita (C_0 , *nugget* en inglés): El intercepto, el valor de la semivarianza en el origen, y representa por lo general una discontinuidad del variograma en el origen, que se puede deber a la escala de muestreo o errores de medición.

Rango (a , *range* en inglés): El límite del área de influencia, es la distancia a partir del cual el variograma se estabiliza y se alcanza la meseta; a partir de esta distancia las observaciones se consideran independientes (sin relación).

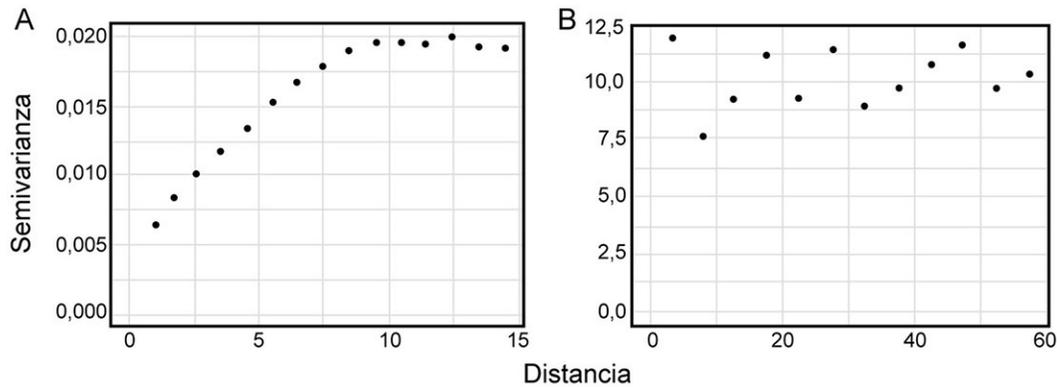


Fig. 2: Ejemplo de variogramas experimentales: A Mostrando la relación (dependencia) espacial de la variable, B Mostrando la ausencia de relación (dependencia) espacial de la variable.

Modelos

Aquí se exponen los principales tipos de modelos que se usan en geociencias (Goovaerts, 1997; Isaaks y Srivastava, 1989; Sarma, 2009; Webster y Oliver, 2007). La figura 4 muestra la forma de estos diferentes modelos, junto con sus partes.

Potencia: Es más usado cuando el variograma no se estabiliza o alcanza una meseta. Se calcula mediante la ecuación (2), donde α es la pendiente, $0 < \alpha < 2$ y controla la concavidad o convexidad del modelo. Un ejemplo se muestra en la figura 4 A.

$$\gamma(h) = C_0 + \alpha h^\lambda \tag{2}$$

Esférico: Es de los más usados en geociencias, presenta una meseta definida, y se caracteriza por presentar un comportamiento lineal cerca del origen. Se calcula mediante la ecuación (3), y un ejemplo se muestra en la figura 4 B.

$$\gamma(h) = \begin{cases} C_0 + C_1 \left[\frac{3}{2} \left(\frac{h}{a} \right) - \frac{1}{2} \left(\frac{h}{a} \right)^3 \right] & \text{para } h < a \\ C_0 + C_1 & \text{para } h > a \end{cases} \tag{3}$$

Exponencial: Este modelo tiene un comportamiento asintótico y no alcanza una meseta tan estable como el esférico, por esto lo que se usa en el modelo como rango es $r=a/3$, o sea, una tercera parte del rango esperado. Se calcula mediante la Ecuación (4) y un ejemplo se muestra en la figura 4 C.

$$\gamma(h) = C_0 + C_1 \left[1 - \exp\left(-\frac{h}{r}\right) \right] \tag{4}$$

Gaussiano: Este modelo es similar al exponencial en que no alcanza una meseta estable sino que tiene un comportamiento asintótico, y otra característica es que tiene un comportamiento suavizado cerca del origen. Como no alcanza una meseta el rango que se usa en el modelo es $r=a/\sqrt{3}$, o sea, el rango esperado entre la raíz de 3. Se calcula mediante la ecuación (5), y un ejemplo se muestra en la figura 4 D.

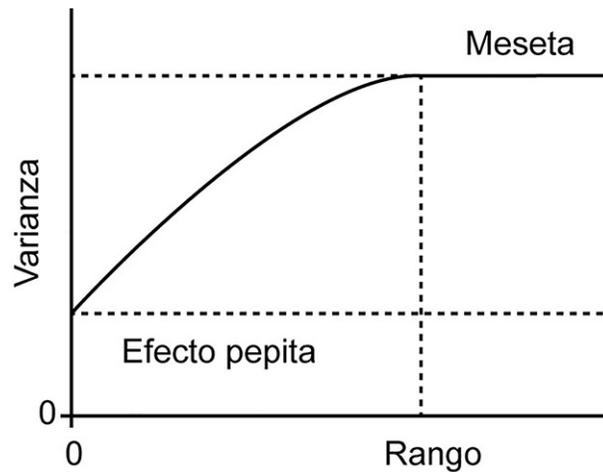


Figura 3: Modelo de variograma mostrando las partes: meseta, pepita, y rango. Modificado de Webster y Oliver (2007).

$$\gamma(h) = C_0 + C_1 \left[1 - \exp\left(-\frac{h^a}{r}\right) \right] \quad (5)$$

La figura 5 es una comparación de los tres modelos más comunes en geociencias, donde todos corresponden con una estructura que presenta los siguientes parámetros: $C_0=0$, $C_1=30$, y $a=210$. Hay que resaltar que el modelo esférico tiene un comportamiento lineal cerca del origen, el modelo exponencial un comportamiento más creciente (convexo), y el modelo gaussiano un comportamiento suavizado. Adicionalmente, los modelos exponencial y gaussiano no alcanzan la meseta de la estructura, contrario al esférico que sí la alcanza.

Anisotropía

La variable y su relación en el espacio puede no solo depender de la distancia sino también de la dirección en que se estima. Si hay una dependencia (comportamiento diferenciado) de la dirección se dice que existe una anisotropía y sino el comportamiento es isotrópico u omnidireccional. La anisotropía se representa como una elipse, donde el eje mayor va a alinearse con la dirección de mayor continuidad espacial y el eje menor con la dirección de menor continuidad espacial (perpendicular al eje mayor) (Chilès y Delfiner, 1999; Goovaerts, 1997; Isaaks y Srivastava, 1989; Oliver y Webster, 2014; Webster y Oliver, 2007). Lo acostumbrado es escoger direcciones en el rango de 0 a 180, ya que al tratarse de un elipse las direcciones mayores a 180 son simplemente el opuesto de direcciones menores a 180 (ejemplo: 220 es el opuesto de 40).

La anisotropía puede ser de dos tipos: geométrica o zonal. La geométrica es la más común y la más fácil de modelar. En la anisotropía geométrica se tiene, para las diferentes direcciones, la misma meseta pero diferente rango. En la anisotropía zonal se tiene el mismo rango pero mesetas diferentes (Chilès y Delfiner, 1999; Goovaerts, 1997; Isaaks y Srivastava, 1989; Webster y Oliver, 2007).

Para determinar la presencia o no de anisotropía se pueden usar el mapa de la superficie de variograma (Fig. 6 A) y/o variogramas direccionales (Fig. 6 B). La anisotropía geométrica va a presentar una dirección principal (eje mayor) que va a estar orientada en la dirección que presenta el mayor rango (mayor continuidad espacial), y una dirección menor (eje menor) orientada perpendicularmente a la principal (Chilès y Delfiner, 1999; Goovaerts, 1997; Isaaks y Srivastava, 1989; Webster y Oliver, 2007).

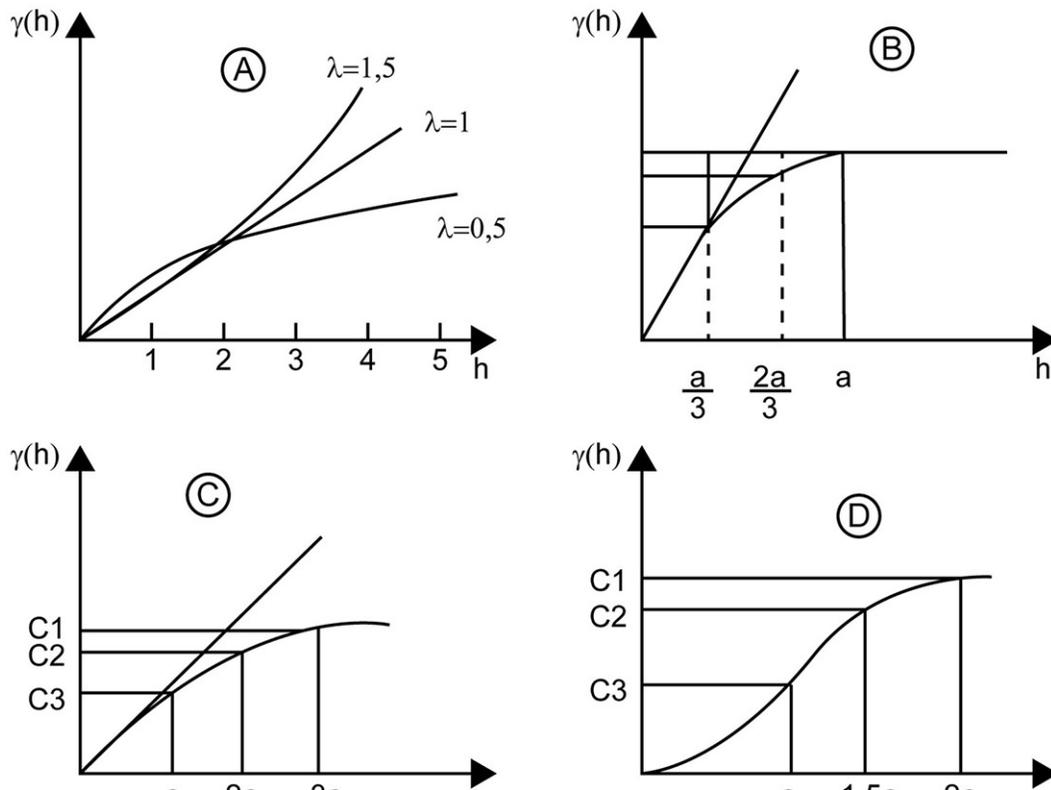


Fig. 4: Modelos más usados en geociencias: A Potencia, B Esférico, C Exponencial, D Gaussiano. Modificado de Sarma (2009).

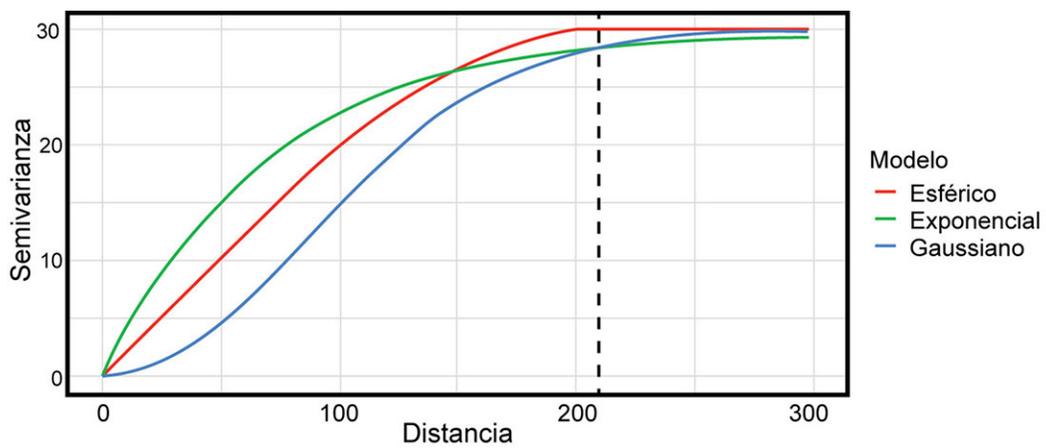


Fig. 5: Comparación visual de los tres modelos más usados en geociencias, todos representando la misma estructura ($C_0=0$, $C_1=30$, y $a=210$).

En la figura 6 la dirección principal coincide con los 35° y la menor con los 125°. En los diferentes softwares por lo general se expresa la anisotropía como una razón y va a depender del software cuál va en el numerador y cuál en el denominador. En el caso del paquete **gstat** la razón de anisotropía va a tener en el numerador la dirección menor y en el denominador la dirección mayor, por lo que la razón va a tener un rango de 0 a 1, donde mientras más cercano a 0 el valor mayor va a ser la anisotropía.

Validación cruzada

Dado que el objetivo de la interpolación es predecir valores en puntos donde no se tiene información, la mejor forma de evaluar el ajuste de un modelo específico sobre el variograma experimental es por medio de la validación cruzada. De manera general lo que se hace es dejar por fuera una o varias de las observaciones, se re-ajusta el modelo seleccionado, y se predice el valor de la variable para esas observaciones que se dejaron por fuera, repitiéndose el proceso hasta tener una predicción para todos los puntos (Chilès y Delfiner, 1999; Goovaerts, 1997; Hastie, Tibshirani, y Friedman, 2008; Isaaks y Srivastava, 1989; James, Witten, Hastie, y Tibshirani, 2013; Kuhn y Johnson, 2013; Oliver y Webster, 2014; Webster y Oliver, 2007; Witten, Frank, y Hall, 2011).

El tipo de validación cruzada más usado es LOO (*leave-one-out*), donde se deja por fuera una observación a la vez, se re-ajusta el modelo y se predice el valor de la variable para cada observación por separado (Goovaerts, 1997; Isaaks y Srivastava, 1989; Oliver y Webster, 2014; Webster y Oliver, 2007). El paquete **gstat** ofrece esta opción (por defecto) y la opción de *K-Fold*. En *K-Fold* se escoge una cantidad de grupos (K) en los que se dividen las observaciones (típicamente 5 o 10) y se deja un grupo de observaciones por fuera cada vez, se re-ajusta el modelo, se predice el valor de la variable para todas las observaciones del grupo que se dejó por fuera, y este proceso se repite K veces hasta tener predicciones para todos los puntos (Hastie et al., 2008; James et al., 2013; Kuhn y Johnson, 2013; Witten et al., 2011).

Una vez realizado el ajuste y la validación cruzada del modelo se obtienen valores predichos y observados para cada punto. Con esta información se pueden usar diferentes métricas, donde lo ideal sería comparar cada una de estas métricas para diferentes modelos ajustados, y se escogería el modelo que obtenga mejores métricas (Chilès y Delfiner, 1999; Goovaerts, 1997; Isaaks y Srivastava, 1989; Oliver y Webster, 2014; Webster y Oliver, 2007).

Dentro de las métricas más usadas están (Oliver y Webster, 2014; Webster y Oliver, 2007; Yao et al., 2013):

En estas métricas N es el total de observaciones (puntos), Y_i es el valor observado en el punto i, \hat{Y}_i es el valor predicho en el punto i, s_{ei}^2 es el error/varianza de la predicción, y \bar{Y} es la media (promedio) de la variable.

- Error medio (ME): El error corresponde con los residuales de lo observado menos lo predicho, una vez se tienen estos valores se les calcula la media e idealmente se esperaría obtener un valor cercano a 0. Se calcula mediante la ecuación (6) y al comparar modelos se escogería el modelo que presente un valor más cercano a 0.

$$ME = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}_i) \quad (6)$$

- Error cuadrático medio (RMSE): Este valor corresponde con la desviación promedio de los errores al cuadrado. Se encuentra en la escala de la variable e idealmente se prefieren valores pequeños. Se calcula mediante la ecuación (7) y comparando modelos se escogería el modelo que presente un RMSE menor.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2} \quad (7)$$

- Razón de desviación cuadrática media (MSDR): Esta valor compara la diferencia entre la predicción y valor actual con respecto a la varianza (error) obtenida de la interpolación (s_{ei}^2). Se esperaría que este valor ande cerca de 1. Se calcula mediante la ecuación (8) y comparando modelos se escogería el que presente un MSDR más cercano a 1.

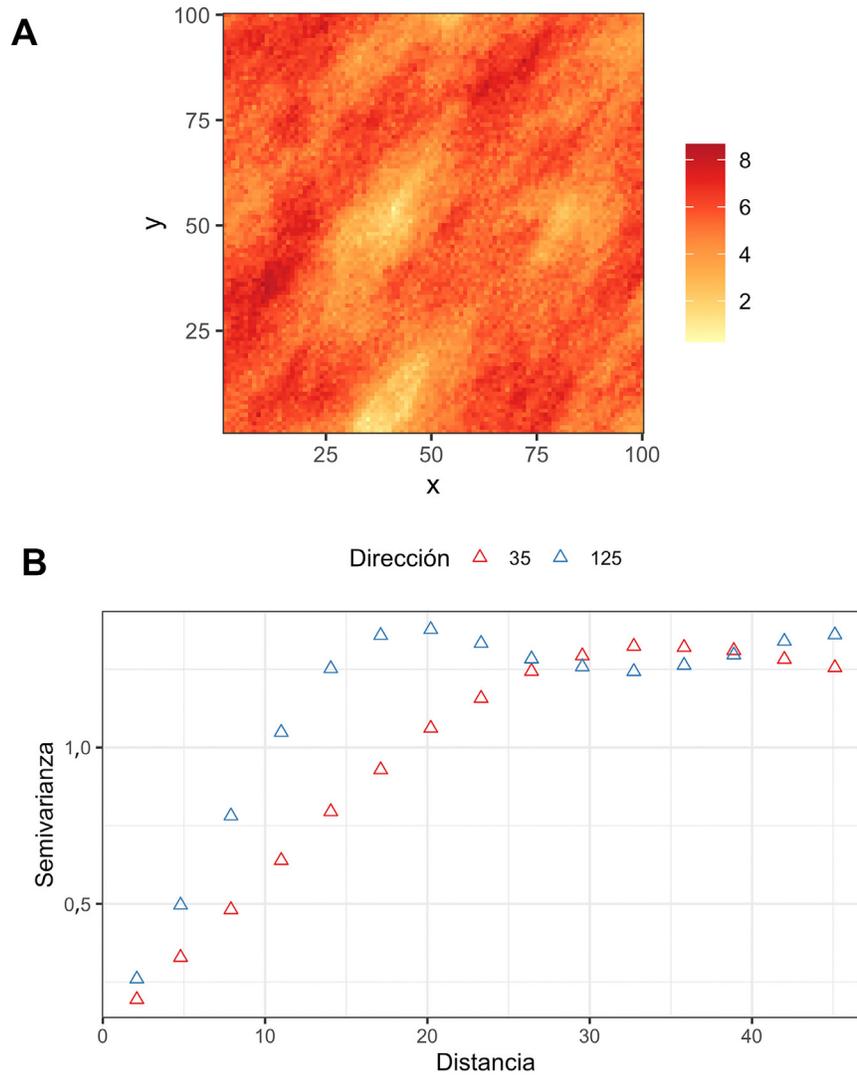


Fig. 6: A Ejemplo de mapa de la superficie de variograma, mostrando anisotropía donde el eje principal ocurre en la dirección 35° y el eje menor ocurre en la dirección 125°. B Variogramas direccionales donde se observa como en la dirección de 35 se alcanza un rango mayor (~25°), mientras que en la dirección perpendicular (125°) el rango es menor (~15°).

$$MSDR = \frac{1}{N} \sum_{i=1}^N \frac{(Y_i - \hat{Y}_i)^2}{s_{ei}^2} \tag{8}$$

- Error Porcentual Absoluto Medio (MAPE): Es una medida porcentual de la diferencia entre lo observado y lo predicho, con un rango de 0 a 1 o de 0 a 100 si se multiplica por 100. Se esperaría que este valor ande cerca de 0 o lo más bajo posible. Se calcula mediante la ecuación (9) y comparando modelos se escogería el que presente el MAPE más bajo.

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{Y_i - \widehat{Y}_i}{Y_i} \right| \quad (9)$$

- Estadístico de bondad de predicción (G): Este estadístico mide qué tan efectiva es la predicción a si se hubiera usado simplemente la media (promedio) de la variable. Valores de 1 indican una predicción perfecta, valores positivos indican que el modelo es más efectivo que usar la media, valores negativos indican que el modelo es menos efectivo que usar la media, y un valor de cero indica que sería mejor usar la media. Se calcula mediante la ecuación (10) y comparando modelos se escogería el que presente el G más cercano a 1 o más positivo.

$$G = 1 - \left[\frac{\sum_{i=1}^N (Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^N (Y_i - \bar{Y})^2} \right] \quad (10)$$

Todas estas métricas, excepto la MSDR, se pueden aplicar para cualquier modelo de cualquier método de interpolación. Para la MSDR se ocupa que el método brinde un error (varianza) de la predicción y ésta es una de las fortalezas de Kriging sobre la mayoría de métodos. Como recomendación, al comparar modelos si hay valores muy similares de la mayoría de las métricas se recomienda usar las métricas de MSDR y el estadístico G como las más importantes.

Kriging

Kriging es un método de interpolación (estimación), por lo que la idea es obtener valores de la variable en lugares donde no se pudo medir. El método hace uso del modelo ajustado para asignar pesos a los puntos a interpolar dependiendo de la distancia entre ellos. Los puntos más cercanos van a presentar valores menores de semivarianza (mayor peso) y los puntos más lejanos valores mayores de semivarianza (menor peso), y si hay puntos que caen fuera del rango estos van a tener una influencia mínima o nula (Chilès y Delfiner, 1999; Goovaerts, 1997; Isaaks y Srivastava, 1989; Webster y Oliver, 2007). Lo anterior se presenta de manera gráfica en la figura 7.

Dentro de las ventajas del Kriging están que compensa por efectos de agrupamiento (*clustering*) al dar menos peso individual a puntos dentro del agrupamiento que a puntos aislados, y da una estimación de la variable y del error (varianza de Kriging) (Chilès y Delfiner, 1999; Goovaerts, 1997; Isaaks y Srivastava, 1989; Trauth, 2015; Webster y Oliver, 2007). El resultado de la interpolación por medio Kriging, por lo general, suaviza los resultados, y sobre-estima valores pequeños y sub-estima valores grandes (Oliver y Webster, 2014; Webster y Oliver, 2007).

Kriging es un método general con diferentes variantes dependiendo de la información que se tenga, el tipo de variable, y la cantidad y tipos de variables a considerar. A manera más general también puede incorporar información temporal, por lo que se puede determinar y modelar la variación espacio-temporal de la variable o variables. **Es más recomendado usar Kriging cuando los datos están normalmente distribuidos, se tiene una buena cantidad de observaciones (depende pero 30, 40 o más es lo recomendado), son estacionarios (la media y varianza de la variable no varían significativamente, esto puede subsanarse con diferentes variantes), y hay una dependencia espacial de la variable** (variograma muestra un incremento de la semivarianza con la distancia) (Chilès y Delfiner, 1999; Goovaerts, 1997; Isaaks y Srivastava, 1989; Webster y Oliver, 2007).

Los tipos de Kriging más comunes son (Chilès y Delfiner, 1999; Goovaerts, 1997; Isaaks y Srivastava, 1989; Webster y Oliver, 2007):

- Simple (SK): Para esta variante se asume que se conoce la media de la variable (lo cual no es necesariamente cierto), y que la media es constante. En general no es práctico de usar.

- Ordinario (OK): Esta variante es la más usada, donde se asume una media constante pero desconocida, y adicionalmente los datos no deben presentar una tendencia.

- Lognormal (OK_{log}): Esta variante se usa cuando la variable tienen una fuerte asimetría positiva, donde se aplica el logaritmo a los datos, y sobre estos datos log-transformados se aplica el Kriging Ordinario; lo más común es usar el logaritmo natural.

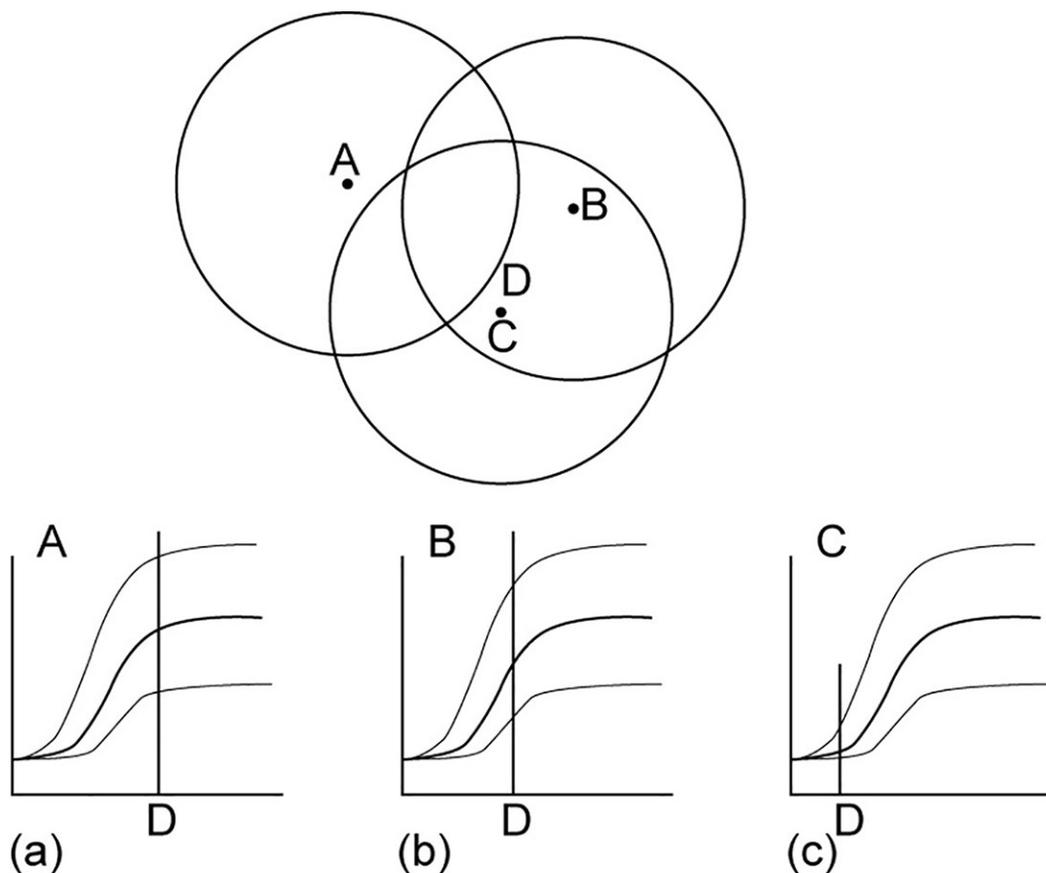


Fig. 7: Visualización del proceso de interpolación mediante Kriging, donde para el punto a interpolar (D), el punto que está más cercano (C) tiene más peso (influencia, baja semivarianza), y el punto más lejano (A) prácticamente no tiene peso ya que cae fuera del rango. Modificado de McKillup y Darby Dyar (2010).

Para obtener el resultado de la interpolación en la escala original de la variable NO es tan simple como exponenciar los resultados. Cressie (1993), Webster y Oliver (2007), Laurent (1963), y Yamamoto (2007) brindan más detalles de cómo realizar la transformación inversa de la manera más apropiada.

- Universal (UK): Esta variante aplica cuando la media no es constante y no se conoce; se le conoce también como Kriging con tendencia (*Kriging in the presence of a trend*). Esta es una forma de trabajar cuando los datos presentan una tendencia (típicamente en función de las coordenadas), como es el caso típico de niveles piezométricos. Lark, Cullis, y Welham (2006) brinda más detalles y técnicas más actualizadas de como lidiar con este tipo de situación.

- CoKriging (CK): Esta variante se usa cuando se quiere utilizar la información de 2 o más variables, y corresponde con la versión multivariable de Kriging. Es necesario que haya una relación entre las variables y su relación espacial, lo que se conoce como co-regionalización.

- Indicador (IK): Esta variante se usa cuando la variable es cualitativa (categórica) o se transforma una variable cuantitativa en cualitativa para determinar si la variable excede o no un umbral. El resultado es la probabilidad condicional de cada una de las categorías (niveles) de la variable.

Eldeiry y Garcia (2010), Kravchenko y Bullock (1999), Meng, Liu, y Borders (2013), Wang et al. (2017), y Yao et al. (2013) hacen uso de varios de los tipos de Kriging, así como de otros métodos de interpolación, describiendo brevemente los métodos y comparando los resultados entre ellos.

Análisis geoestadístico

Una vez presentada la teoría básica de la geoestadística se va a proceder a realizar un análisis geoestadístico típico (con el objetivo de estimar la distribución de la variable de interés en el espacio). Los datos corresponden con la temperatura promedio de los últimos 10 años para el 8 de Marzo para la provincia de San José. Los datos fueron tomados de Meteomatics (2021), de donde se pueden obtener diferentes parámetros meteorológicos/climáticos a nivel mundial. Se usan estos datos por tener amplia cobertura espacial, lo que los hacen muy didácticos para ejemplificar el uso de la geoestadística.

Se va a hacer uso de R que permite manipular y analizar datos (espaciales y no espaciales), y además tiene diversos paquetes (librerías) para realizar análisis geoestadísticos (Finley, Banerjee, y Gelfand, 2015; Gräler, Pebesma, y Heuvelink, 2016; Jing y Oliveira, 2015; Pebesma, 2004; Pebesma y Graeler, 2021; Ribeiro, Christensen, y Diggle, 2003). El código usado, así como su explicación, se pueden encontrar en el material extra disponible en el repositorio de GitHub del trabajo: https://github.com/maxgav13/intro_geostats.

Análisis exploratorio de datos

Antes de iniciar con el análisis geoestadístico es necesario estudiar la variable, ver su distribución (si se aproxima a una distribución normal) para determinar si es necesaria alguna transformación, y por medio de la varianza se puede tener una idea aproximada de la meseta total del variograma.

El resumen estadístico (Cuadro 1) y el histograma (Fig. 8) muestran que tiene una distribución aproximadamente normal, donde la media y mediana son similares y el histograma presenta una forma general de campana con una asimetría inferior a 1, por lo que no es necesaria ninguna transformación. La varianza de la variable es 19,258, lo que brinda una aproximación de la meseta total del variograma.

En este caso los datos tienen coordenadas geográficas pero de manera general se recomienda trabajar los datos en sistemas de coordenadas planas (x,y) por lo que se recomienda convertirlas a estas conforme la zona de estudio, utilizando los códigos epsg respectivos. En este caso el código que corresponde es el 5367 para el sistema de coordenadas CRTM05. Para más información al respecto se puede consultar Garnier-Villarreal (2020), donde el capítulo 6 está dedicado al trato de datos espaciales en R.

Es buena práctica determinar las distancias entre los puntos, ya que como se explicó en la parte teórica, no es recomendado calcular el variograma experimental a más de la mitad de la distancia máxima entre puntos. Haciendo este paso se obtiene que la distancia máxima es de 142,00 km. Dada la zona de estudio tan grande se van a presentar las distancias en kilómetros para mayor facilidad y legibilidad, pero hay que tener en cuenta que los datos se encuentran en metros.

Un paso inicial antes de empezar con el análisis es visualizar la distribución de la variable en el espacio, para tener una idea preliminar de patrones que pueda presentar. La figura 9 muestra la ubicación de los datos, donde los puntos se encuentran rellenados de acuerdo al valor de la temperatura, y donde se puede observar que hay una predominancia de temperaturas mayores al SW y menores al NE.

Modelado geoestadístico

Habiendo estudiado la variable y hecho los pasos iniciales de manipulación, análisis y visualización se procede con el modelado geoestadístico, mostrando y explicando los distintos pasos. En este caso como la variable no requirió de ninguna transformación se va a usar el Kriging Ordinario.

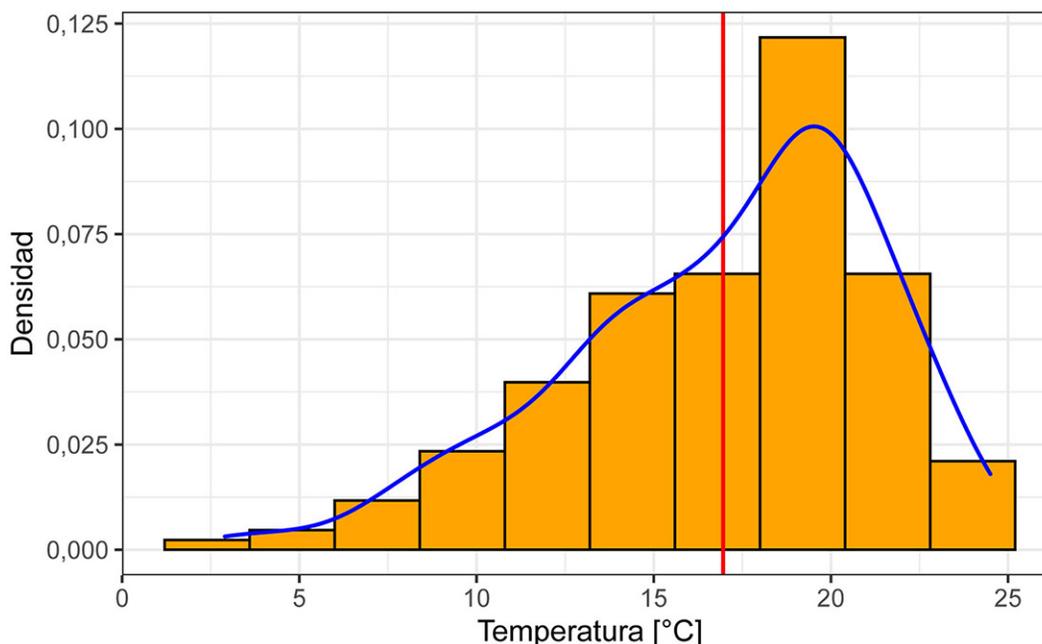


Fig. 8: Histograma de la temperatura promedio de los últimos 10 años para el 8 de Marzo para la provincia de San José. La línea roja corresponde con la media, y la curva azul con la curva de densidad empírica.

Variograma experimental

El primer paso es crear un variograma experimental omnidireccional. Se va a hacer uso del paquete **gstat** (Gräler et al., 2016; Pebesma, 2004; Pebesma y Graeler, 2021) para la geoestadística. En la construcción del variograma experimental se deben definir los argumentos de el intervalo de distancia deseado (h), y la distancia máxima a la cual calcular la semivarianza. Si recordamos la distancia máxima era 142,00 km, por lo que se escoge un valor ligeramente inferior a la mitad (70 km) para la distancia máxima y un valor de 4 km para el intervalo de distancias h . Es en este paso donde el usuario puede probar diferentes valores para obtener un variograma representativo, que muestre una estructura de dependencia espacial y que los puntos del variograma se hayan calculado con suficientes datos (recomendable 20 o más).

Una vez analizado el variograma omnidireccional se procede a determinar si existe la presencia o no de anisotropía. Para esto se usan tanto el mapa de la superficie de variograma (Fig. 11), como los variogramas direccionales (Fig. 12).

Para los variogramas direccionales hay que definir, adicionalmente, los argumentos de las direcciones y la tolerancia angular, donde lo más usado son direcciones cada 45° y la tolerancia angular es la mitad del intervalo entre direcciones ($22,5^\circ$). De nuevo, solo es necesario definir direcciones entre 0 y 180, y 180 se excluye por ser el opuesto de 0.

Analizando el mapa y los variogramas direccionales se concluye que hay una anisotropía con dirección principal de 135° y un rango aproximado de 50 km, y un rango aproximado de 25 km en la dirección de 45° , resultando en una razón de anisotropía de 0,5. Por lo anterior el modelado se realizará con los variogramas direccionales.

Cuadro 1

Resumen estadístico de la variable.

	N	Media	Desv. Est.	Min	Mediana	Max	MAD	CV	Asimetría
Temperatura	178	16,96	4,39	2,9	18,05	24,5	4,23	0,26	-0,73

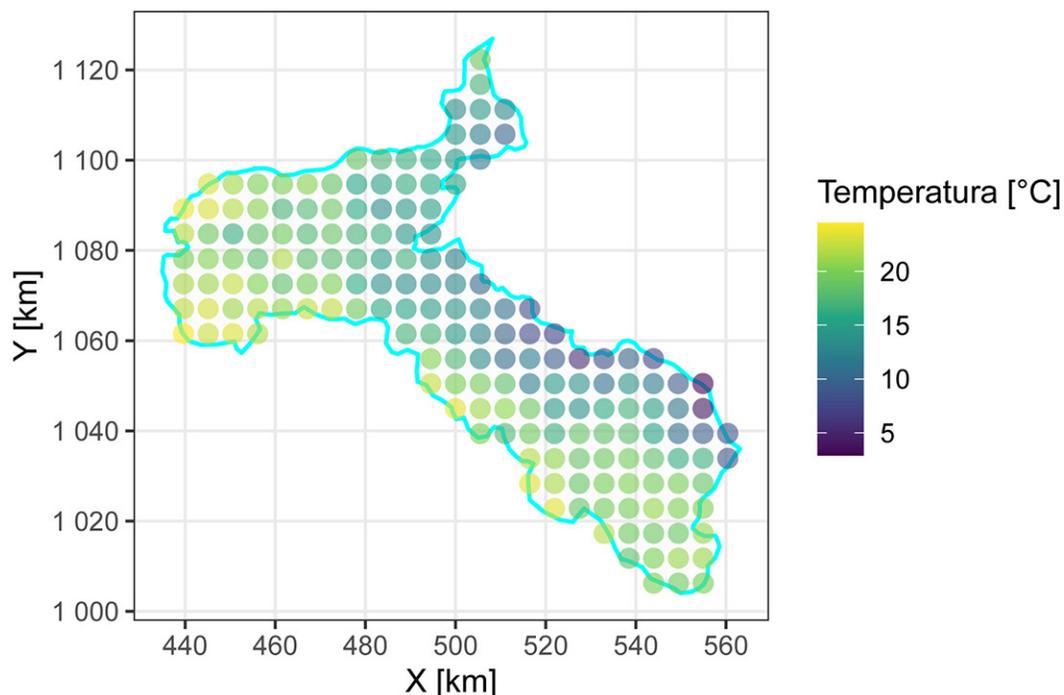


Fig. 9: Mapa de puntos mostrando la distribución espacial de la temperatura promedio de los últimos 10 años para el 8 de marzo para la provincia de San José.

Ajuste de modelo de variograma

Una vez creado el variograma experimental es necesario ajustarle un modelo para poder obtener valores a distancias no muestreadas. Antes de ajustar un modelo al variograma experimental es necesario estimar las partes del mismo (meseta, pepita, rango, anisotropía) y determinar valores iniciales, para posteriormente realizar el ajuste.

Usando los variogramas direccionales (Fig. 12), se puede estimar una pepita de aproximadamente 0, una meseta parcial de 25, un rango de 50000, una anisotropía a 135° con una razón de 0,5, y se puede usar un modelo tipo esférico ('Sph').

El modelo ajustado (Cuadro 2) se puede usar para calcular un error del ajuste inicial ($RMSE_{ajuste} = 4e - 04$), pero es más confiable el que se obtiene usando la validación cruzada, ya que el obtenido acá es un valor optimista. Lo anterior se da puesto que se calcula con respecto a los datos que se utilizaron para el ajuste (toda la información disponible) y esto siempre va a resultar en error menor que cuando se usa el modelo en datos no observados (Hastie et al., 2008; James et al., 2013; Kuhn y Johnson, 2013; Witten et al., 2011), que es el objetivo de la interpolación.

El modelo ajustado del cuadro 2 se puede interpretar así: el efecto pepita ('Nug') (que como es el intercepto solo aporta información a la semivarianza y no al rango) aporta 0,000 a la semivarianza ($C_0=0,000$); el modelo esférico ('Sph') aporta 23,990 a la semivarianza ($C_1=23,990$), con lo que la meseta total es $S=C_0+C_1=23,990$, y el rango mayor del modelo esférico es $a=84672,58$ en una dirección 135° , con una razón de anisotropía de 0,5.

Cuadro 2

Modelo ajustado.

Modelo	Meseta	Rango	Razón	Dirección Principal
Nug	0,00	0,00	0,0	0
Sph	23,99	84672,58	0,5	135

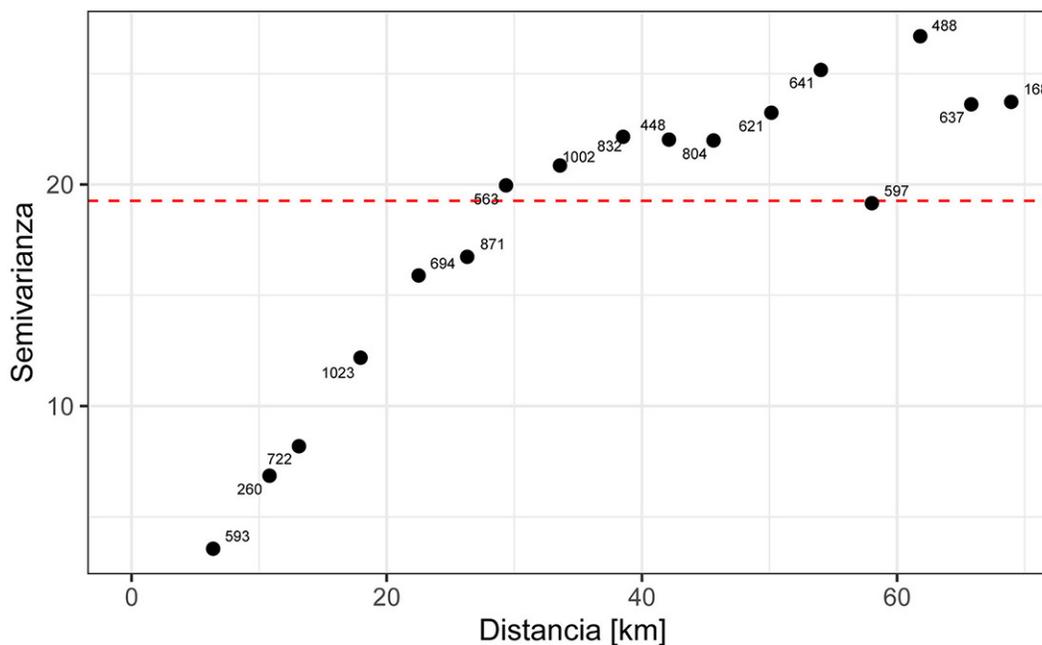


Fig. 10: Variograma experimental omnidireccional de la temperatura promedio de los últimos 10 años para el 8 de Marzo para la provincia de San José. Las etiquetas de los puntos muestran con el número de pares de puntos usados para el cálculo de la semivarianza. La línea roja punteada corresponde con la varianza de la variable, que es una aproximación a la meseta total.

Con el modelo ajustado se puede visualizar este sobre el variograma omnidireccional (Fig. 13) y los variogramas direccionales (Fig. 14). En general, para todos los casos se observa que el modelo ajustado es válido y representativo para todos los casos.

Validación cruzada

Para evaluar de manera más realista el ajuste de cualquier modelo es mejor usar la validación cruzada. Es en este paso que se podrían probar diferentes modelos, donde se obtienen las métricas de ajuste de un modelo, se ajusta un nuevo modelo y se obtienen sus métricas de ajuste, y así iterativamente. Una vez ajustados diferentes modelos y con sus diferentes métricas, se puede tener un criterio más robusto de cuál modelo se ajusta mejor a los datos.

Las métricas usadas aquí son las que se introdujeron anteriormente: el error cuadrático medio (RMSE), la razón de desviación cuadrática media (MSDR), el error porcentual absoluto medio (MAPE), y el estadístico de bondad de predicción (G). Adicionalmente se estima la correlación (r) entre los valores observados y predichos, donde lo que se busca es determinar qué tan similares son los valores entre sí (Fig. 15 A).

Como se mencionó arriba las métricas son más útiles cuando se comparan modelos, pero para este caso, usando solo el modelo esférico, se puede decir que presentan valores aceptables (Cuadro 3): la MSDR está cerca de 1, la correlación (r) es alta, el RMSE es menor a la desviación estándar de los datos (4,388), el MAPE es bajo y cercano a 0, y el estadístico G es positivo y cercano a 1. Conforme APA -*American Psychological Association* (2010), valores que no pueden por definición ser superiores a 1 o inferiores a -1 (r , R^2 , MAPE, y G) se reportan sin el '0' inicial.

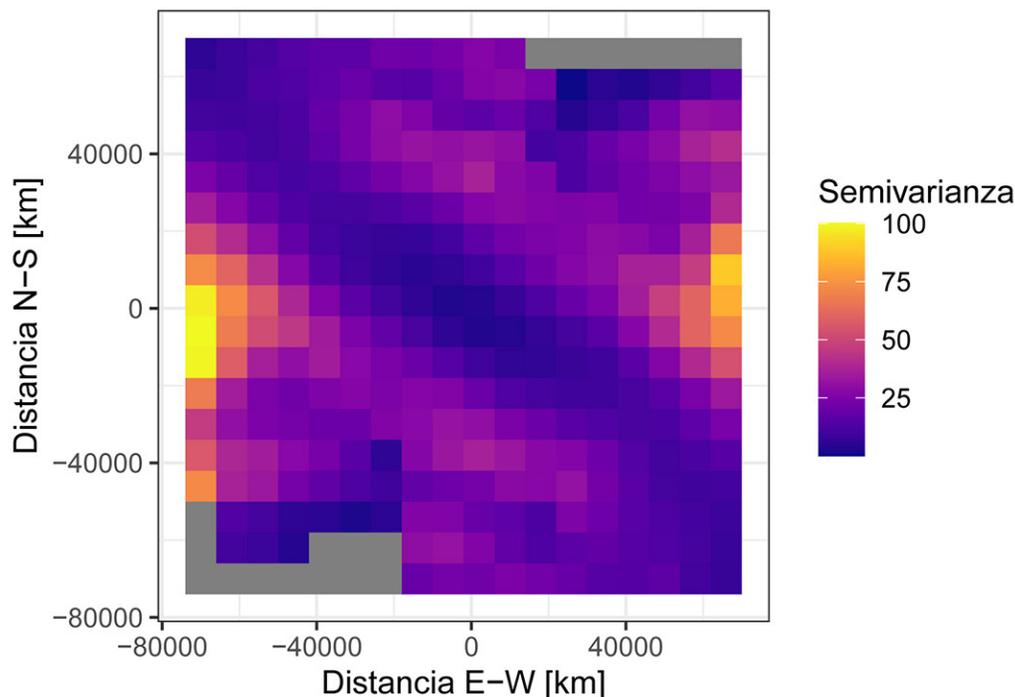


Figura 11: Mapa de la superficie de variograma. Se observa una dirección preferencial a aproximadamente 135°.

Si recordamos el error del ajuste inicial sobre los datos que se realizó el ajuste fue $RMSE_{ajuste} = 4e-04$, que como se puede observar es mucho menor al error de la validación cruzada $RMSE_{xval} = 1,475$, de ahí que se le definiera como optimista, y sea el error de la validación cruzada un mejor indicador de la capacidad predictiva del modelo seleccionado.

Adicionalmente se pueden explorar los residuales ya que idealmente se esperaría que presenten una distribución normal centrada en 0. Lo anterior se puede apreciar en la figura 15 B, donde el histograma es aproximadamente normal, y no presenta una asimetría importante (menor a 1: -0,304) y se encuentra centrado alrededor de 0.

Las métricas tanto como los residuales indican que el modelo ajustado es un modelo apropiado para proceder con la interpolación.

Cuadro 3

Métricas de ajuste para la validación cruzada.

Métrica	Valor
RMSE	1,475
MSDR	0,758
r	.942
R ²	.887
MAPE	.085
G	.886

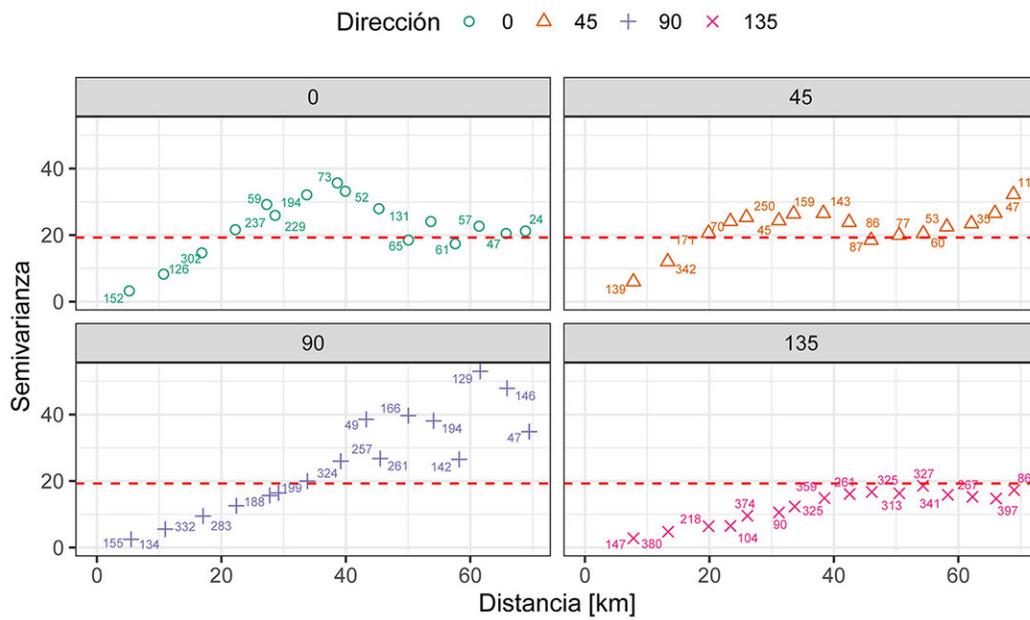


Fig. 12: Variogramas experimentales direccionales cada 45°. La línea roja punteada representa la varianza de la variable, lo que se aproxima a la meseta total. Se observa un mayor rango en la dirección 135° y un menor rango en la dirección 45°.

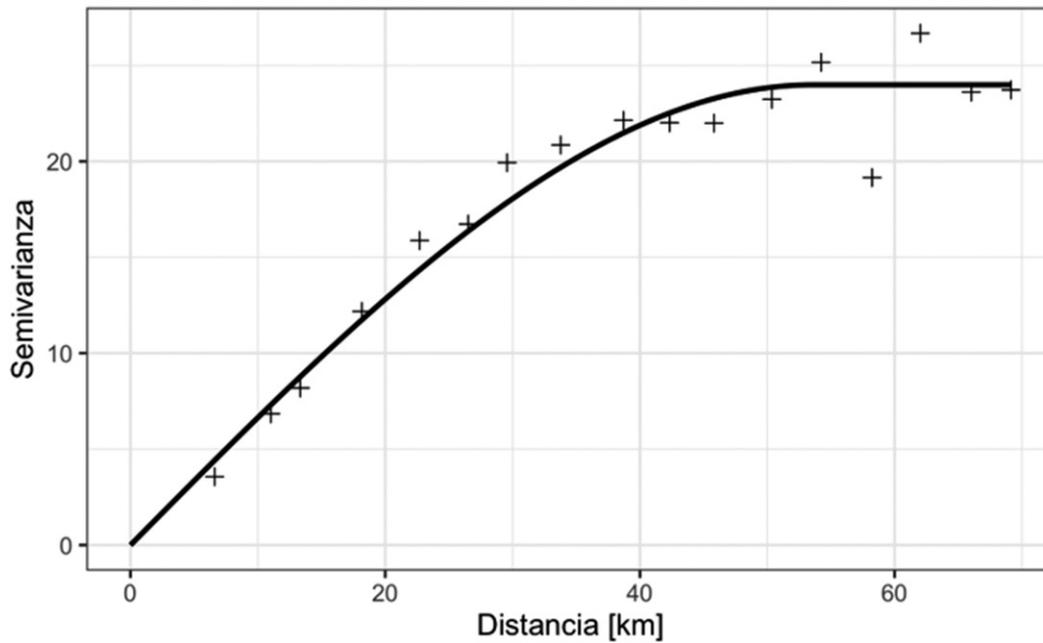


Fig. 13: Variograma experimental omnidireccional con el modelo esférico ajustado sobrepuesto.

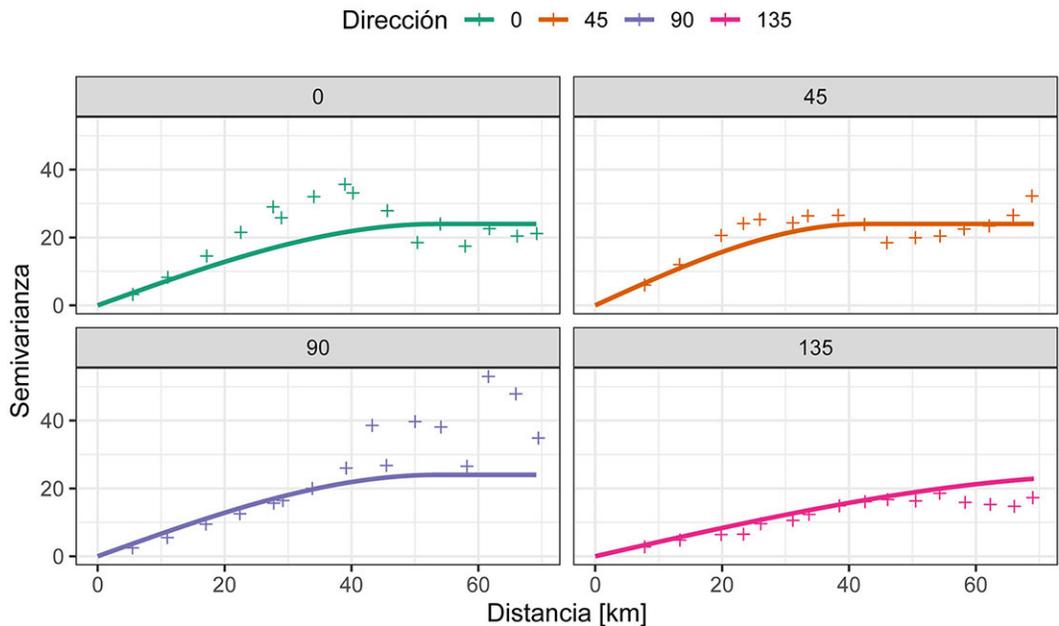


Figura 14: Variogramas experimentales direccionales con el modelo esférico ajustado sobrepuesto, mostrando que el modelo es válido.

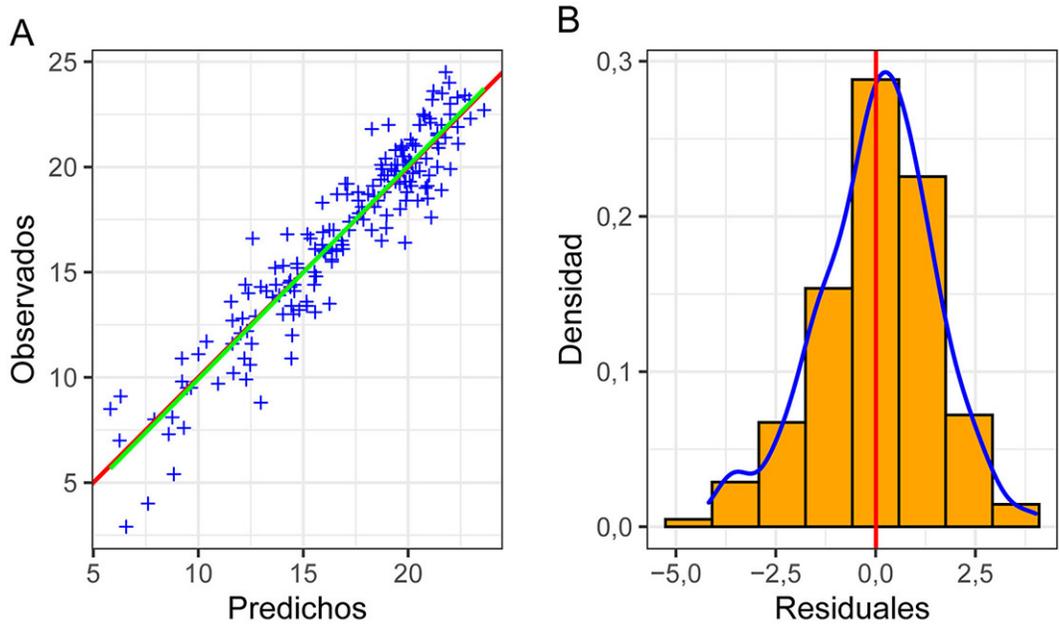


Fig. 15: Análisis de los resultados de validación cruzada. A Relación entre los valores observados y predichos por la validación cruzada. La línea roja es la línea 1:1 y la línea verde es la regresión entre los datos. B Histograma de los residuales de la validación cruzada. La línea roja es la media de los residuales y la curva azul la curva de densidad.

Interpolación (Kriging)

Para recalcar nuevamente, el análisis geoestadístico es un proceso que conlleva el cálculo del variograma, el ajuste de un modelo, la validación del modelo a usar, y por último la interpolación mediante Kriging. Si no se realizan con cuidado los pasos el resultado de la interpolación puede no tener validez o sentido.

Los mapas finales tanto de la predicción (estimación) como de la varianza (error de estimación) se presentan en la figura 16. En el mapa de la predicción se observa una tendencia de valores altos de la temperatura hacia el SW del área y de valores bajos de temperatura hacia el NE, con una zonación con dirección de rumbo paralela al eje mayor de la anisotropía. Esta distribución espacial de la temperatura tiene sentido puesto que la costa (clima más cálido) se encuentra en dirección SW, y las partes montañosas (clima más fresco) de la meseta central en la dirección NE. El mapa de la varianza va a presentar los valores más bajos en los puntos de muestreo y valores mayores en puntos más distantes de los muestreados, que es un comportamiento típico de Kriging.

Aplicación web

Lo demostrado acá se encuentra implementado en una aplicación web (Garnier-Villarreal, 2019), la cual puede ser usada accediendo a la siguiente dirección <https://maximiliano-01.shinyapps.io/geostatistics/>. La idea de la aplicación es llevar de la mano al usuario por los mismos pasos presentados acá, usando una interfaz más familiar, sin necesidad de que sepa usar **R** o lenguajes de programación, pero sí es necesario que se entienda y tenga conciencia de lo que conlleva un análisis geoestadístico de principio a fin.

La aplicación puede leer (cargar) archivos '.txt' o '.csv', donde el archivo tiene que contener por lo menos tres columnas: coordenada-x, coordenada-y, variable de interés. La figura 17 muestra la interfaz de la aplicación.

Conclusiones

Kriging es un método de interpolación, de varios disponibles, para obtener predicciones (estimaciones) en puntos donde no se tienen observaciones, y adicionalmente presenta diferentes variantes, por lo que no es único y depende del objetivo de investigación el cómo se implementa. Cuando es posible y adecuado usarlo, típicamente, brinda los mejores resultados, además de proporcionar una estimación del error sobre los valores estimados.

Kriging como tal es uno de los posibles usos de la geoestadística, ya que es un paso (el último típicamente) durante un análisis geoestadístico donde el objetivo es la predicción (estimación) de una o varias variables en el espacio. Se menciona, brevemente, que otro posible producto de la geoestadística es la simulación, la cual puede ser más representativa en casos donde la heterogeneidad, y no el comportamiento promedio, de la variable es el interés principal.

La geoestadística, como rama de la estadística espacial, se enfoca en la caracterización de procesos y variables que tienen una fuerte componente espacial, por lo que existe una inter-dependencia entre las observaciones, a diferencia de la estadística clásica.

Este trabajo muestra los pasos, cuidados, y decisiones que hay que tomar durante un análisis geoestadístico típico, haciendo énfasis en que para obtener resultados válidos y confiables es necesario desarrollar estos pasos con criterio y no dejarlos a decisión de un programa de cómputo. Se recomienda que cuando se hace uso de Kriging se detalle el tipo, así como el modelo que se ajustó y sus parámetros.

En el ejemplo de la temperatura promedio de los últimos 10 años para el 8 de Marzo para la provincia de San José, se determinó la existencia de anisotropía con un eje mayor en dirección 135°, y una distribución de la temperatura donde valores altos se encuentran hacia el SW y valores bajos hacia el NE.

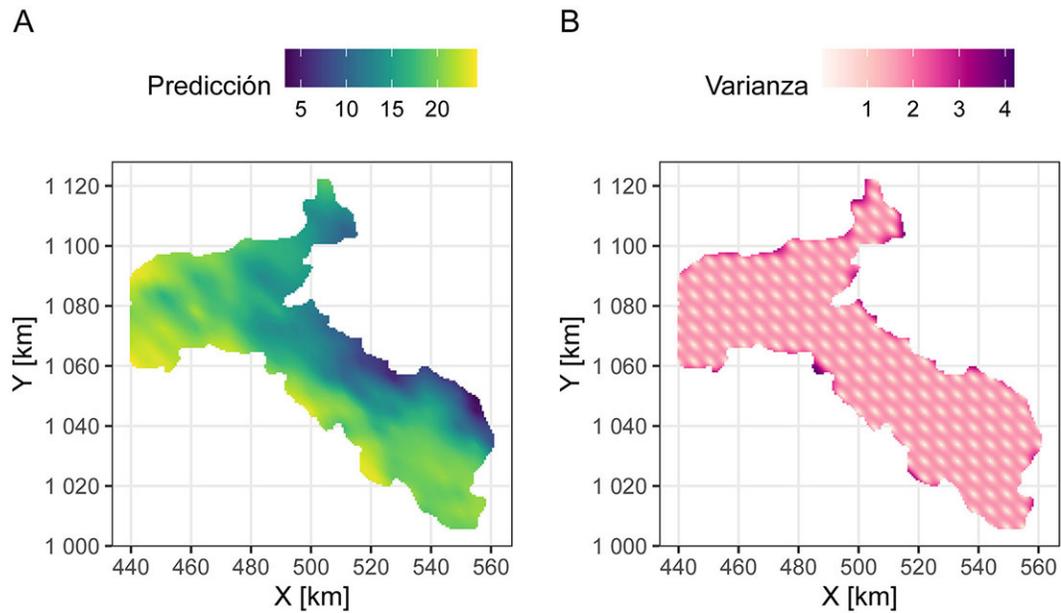


Fig. 16: Mapas de predicción (A) y de la varianza/error (B) de la temperatura para la provincia de San José, para la fecha del 8 de Marzo.

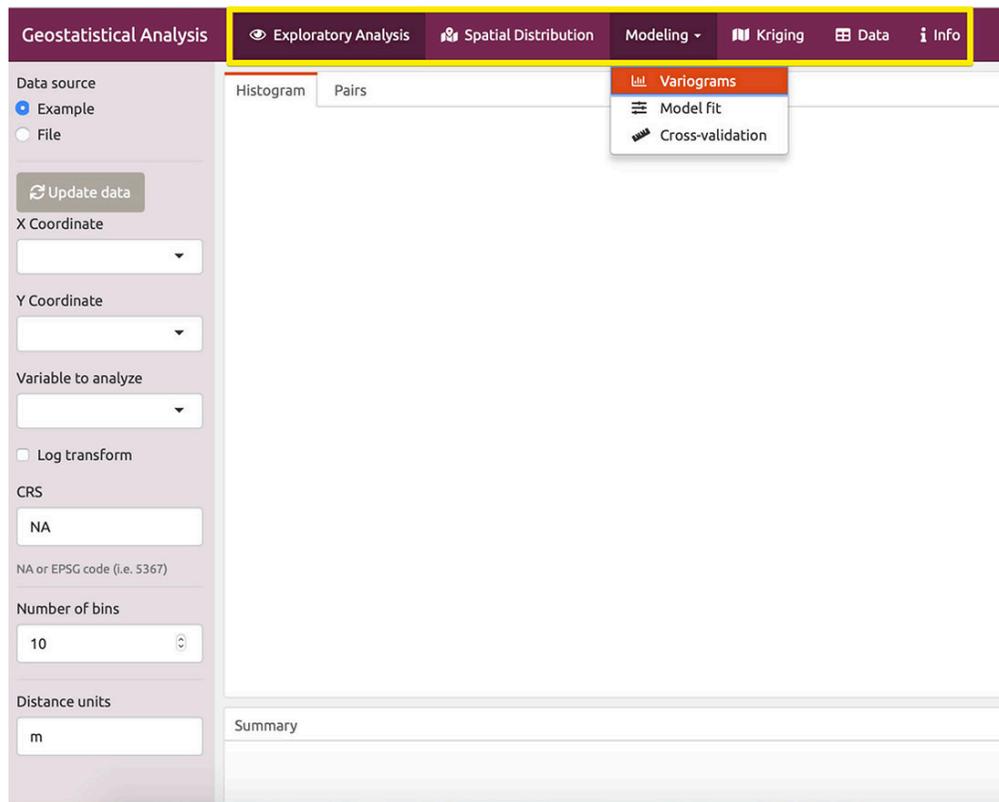


Fig. 17: Interfaz de la aplicación web para realizar análisis geoestadístico. El rectángulo amarillo encierra las viñetas (pasos) a seguir durante el análisis, incluyendo además la viñeta para desplegar los datos y la de información adicional de la aplicación.

R es un lenguaje de programación muy flexible que permite crear rutinas para reusar posteriormente en análisis futuros similares. El material de este trabajo se encuentra disponible en un repositorio en GitHub (https://github.com/maxgav13/intro_geostats), que puede ser descargado para su uso. En el repositorio se puede consultar un material extra que presenta y detalla el código usado durante el análisis geoestadístico. Además se presenta de manera muy rápida una aplicación web que hace uso del mismo código, pero de una manera más amigable para quienes no se sienten cómodos con lenguajes de programación.

Referencias bibliográficas

- American Psychological Association. (2010). *Publication Manual of the American Psychological Association* (6.^a ed.). Washington, D.C.: American Psychological Association.
- Borradaile, G. J. (2003). *Statistics of earth science data: their distribution in time, space and orientation*. New York: Springer-Verlag Berlin Heidelberg.
- Chilès, J.-P., y Delfiner, P. (1999). *Geostatistics: modeling spatial uncertainty* (2.^a ed.). New York: John Wiley & Sons.
- Cressie, N. A. C. (1993). *Statistics for spatial data*. Hoboken, NJ: John Wiley & Sons.
- Davis, J. C. (2002). *Statistics and data analysis in Geology* (3.^a ed.). New Delhi: John Wiley & Sons.
- Eldeiry, A. A., y Garcia, L. A. (2010). Comparison of ordinary kriging, regression kriging, and cokriging techniques to estimate soil salinity using LANDSAT images. *Journal of Irrigation and Drainage Engineering*, 136(6), 355-364. doi: 10.1061/(ASCE)IR.1943-4774.0000208
- Finley, A. O., Banerjee, S., y Gelfand, A. (2015). spBayes for large univariate and multivariate point-referenced spatio-temporal data models. *Journal of Statistical Software*, 63(13), 1-28. doi: 10.18637/jss.v063.i13
- Garnier-Villarreal, M. (2019). *Geostatistical analysis (Versión 1)*. Recuperado de <https://maximiliano-01.shinyapps.io/geostatistics/>
- Garnier-Villarreal, M. (2020). *Geología numérica: ciencia de datos para geociencias*. Recuperado de <https://geologia-numerica.netlify.app>
- Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. New York: Oxford University Press.
- Gräler, B., Pebesma, E., y Heuvelink, G. (2016). Spatio-temporal interpolation using gstat. *The R Journal*, 8, 204-218. Recuperado de <https://journal.r-project.org/archive/2016/RJ-2016-014/index.html>
- Hastie, T., Tibshirani, R., y Friedman, J. (2008). *The elements of statistical learning: data mining, inference, and prediction* (2.^a ed.). New York: Springer.
- Isaaks, E. H., y Srivastava, R. M. (1989). *Applied geostatistics*. New York: Oxford University Press.
- James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*. New York: Springer.
- Jing, L., y Oliveira, V. D. (2015). geoCount: An R package for the analysis of geostatistical count data. *Journal of Statistical Software*, 63(11), 1-33. doi: 10.18637/jss.v063.i11
- Kravchenko, A., y Bullock, D. G. (1999). A comparative study of interpolation methods for mapping soil properties. *Agronomy Journal*, 91(3), 393-400. doi: 10.2134/agronj1999.00021962009100030007x
- Kuhn, M., y Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.
- Lark, R. M., Cullis, B. R., y Welham, S. J. (2006). On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. *European Journal of Soil Science*, 57(6), 787-799. doi: 10.1111/j.1365-2389.2005.00768.x
- Laurent, A. G. (1963). The lognormal distribution and the translation method: description and estimation problems. *Journal of the American Statistical Association*, 58(301), 231-235.
- Linkimer, L. (2008). Application of the kriging method to draw isoseismal maps of the significant 2002-2003 Costa Rican earthquakes. *Revista Geológica de América Central*, 38, 119-134. doi: 10.15517/rgac.v0i38.4220
- McKillop, S., y Darby Dyar, M. (2010). *Geostatistics explained: an introductory guide for earth scientists*. Recuperado de www.cambridge.org/9780521763226

- Meng, Q., Liu, Z., y Borders, B. E. (2013). Assessment of regression kriging for spatial interpolation – comparisons of seven GIS interpolation methods. *Cartography and Geographic Information Science*, 40(1), 28-39. doi: 10.1080/15230406.2013.762138
- Meteomatics. (2021). Meteomatics weather API. Recuperado de www.meteomatics.com
- Nowosad, J. (2019). *Geostatystyka w R*. Recuperado de <https://bookdown.org/nowosad/geostatystyka/>
- Oliver, M. A., y Webster, R. (2014). A tutorial guide to geostatistics: computing and modelling variograms and kriging. *CATENA*, 113, 56-69. doi: 10.1016/j.catena.2013.09.006
- Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, 30, 683-691.
- Pebesma, E., y Bivand, R. (2020). *Spatial data science*. Recuperado de <https://keen-swartz-3146c4.netlify.app>
- Pebesma, E., y Graeler, B. (2021). *gstat: Spatial and Spatio-Temporal Geostatistical Modelling, Prediction and Simulation (Versión paquete de R 2.0-7)*. Recuperado de <https://github.com/r-spatial/gstat/>
- Pyrz, M. J., y Deutsch, C. V. (2014). *Geostatistical reservoir modeling (2.ª ed.)*. New York: Oxford University Press.
- R Core Team. (2021). *R: a language and environment for statistical computing*. Recuperado de <https://www.R-project.org/>
- Ribeiro, P. J., Christensen, O. F., y Diggle, P. J. (2003). geoR and geoRglm: Software for Model-Based Geostatistics. En K. Hornik, F. Leisch y A. Zeileis (eds), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (pp. 1-16). Vienna, Austria. Recuperado de <https://www.r-project.org/conferences/DSC-2003/Proceedings/RibeiroEtAl.pdf>
- Sarma, D. D. (2009). *Geostatistics with application in Earth Sciences (2.ª ed.)*. New Delhi: Springer.
- Swan, A. R. H., y Sandilands, M. (1995). *Introduction to geological data analysis*. Oxford: Blackwell Science.
- Trauth, M. H. (2015). *MATLAB® Recipes for Earth Sciences (4.ª ed.)*. Berlin: Springer-Verlag Berlin Heidelberg.
- Varouchakis, E. A., Hristopulos, D. T., y Karatzas, G. P. (2012). Improving kriging of groundwater level data using non-linear normalizing transformations—a field application. *Hydrological Sciences Journal*, 57(7), 1404-1419. doi: 10.1080/02626667.2012.717174
- Wackernagel, H. (2003). *Multivariate Geostatistics (3.ª ed.)*. Springer-Verlag Berlin Heidelberg.
- Wang, G., Gertner, G., Fang, S., y Anderson, A. B. (2003). Mapping Multiple Variables for Predicting Soil Loss by Geostatistical Methods with TM Images and a Slope Map. *Photogrammetric Engineering & Remote Sensing*, 69(8), 889-898. <https://doi.org/10.14358/PERS.69.8.889>
- Wang, M., He, G., Zhang, Z., Wang, G., Zhang, Z., Cao, X., ... Liu, X. (2017). Comparison of spatial interpolation and regression analysis models for an estimation of monthly near surface air temperature in China. *Remote Sensing*, 9(12), 1278. doi: 10.3390/rs9121278
- Webster, R., y Oliver, M. A. (2007). *Geostatistics for Environmental Scientists (2.ª ed.)*. West Sussex, Inglaterra: John Wiley & Sons.
- Witten, I. H., Frank, E., y Hall, M. A. (2011). *Data mining: practical machine learning tools and techniques (3.ª ed.)*. Burlington, MA: Elsevier.
- Yamamoto, J. K. (2007). On unbiased backtransform of lognormal kriging estimates. *Computational Geosciences*, 11(3), 219-234. doi: 10.1007/s10596-007-9046-x
- Yao, X., Fu, B., Lü, Y., Sun, F., Wang, S., y Liu, M. (2013). Comparison of four spatial interpolation methods for estimating soil moisture in a complex terrain catchment. *PLoS ONE*, 8(1), e54660. doi: 10.1371/journal.pone.0054660