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

Kriging 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 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.

# 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](#geostats-met-interp), donde comúnmente se ha dado a entender que Kriging es el método por excelencia a usar (casi que indiscriminadamente), pero no se ha profundizado en 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 & 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 artículo es de índole educativa/informativa y corresponde con introducir al lector en qué es la geoestadística y cómo realizar un análisis geoestadístico de manera apropiada. La idea es brindar una base y guía de cómo hacer una interpolación de los datos de interés 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 completo.

Para el procesamiento de los datos y la implementación de la geoestadística se va a utilizar el software libre multi-plataforma **R** (R Core Team, [2019](#ref-R-base)) 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](#ref-garnier-villarreal2020)).

# 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 & Oliver ([2007](#ref-webster2007)). 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)
* **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 **b**est **l**inear **u**nbiased **e**stimator, y se puede traducir como **mejor estimador lineal no sesgado** (Isaaks & Srivastava, [1989](#ref-isaaks1989); Webster & Oliver, [2007](#ref-webster2007)).

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 & Srivastava, [1989](#ref-isaaks1989); Webster & Oliver, [2007](#ref-webster2007)). El método ha sido utilizado para predecir la intensidad sísmica (Linkimer, [2008](#ref-linkimer2008rgac)), el nivel de agua subterránea (Varouchakis et al., [2012](#ref-varouchakis2012hsj)), pérdida de suelo (Wang et al., [2003](#ref-wang2003pers)), y temperatura del aire (Wang et al., [2017](#ref-wang2017rs)), 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](#ref-davis2002); Isaaks & Srivastava, [1989](#ref-isaaks1989); Webster & Oliver, [2007](#ref-webster2007)). 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 & Webster, [2014](#ref-oliver2014c)), 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 & Delfiner, [1999](#ref-chiles1999); Goovaerts, [1997](#ref-goovaerts1997); Pebesma & Bivand, [2020](#ref-pebesma2020); Pyrcz & Deutsch, [2014](#ref-pyrcz2014); Webster & Oliver, [2007](#ref-webster2007)). 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](#ref-davis2002)), Swan & Sandilands ([1995](#ref-swan1995)), Borradaile ([2003](#ref-borradaile2003)), y McKillup & Darby Dyar ([2010](#ref-mckillup2010)), y textos más detallados y exclusivos en la materia de Chilès & Delfiner ([1999](#ref-chiles1999)), Cressie ([1993](#ref-cressie1993)), Goovaerts ([1997](#ref-goovaerts1997)), Isaaks & Srivastava ([1989](#ref-isaaks1989)), Pyrcz & Deutsch ([2014](#ref-pyrcz2014)), Webster & Oliver ([2007](#ref-webster2007)), y Wackernagel ([2003](#ref-wackernagel2003)), los cuales corresponden con referencias clásicas y actualizadas. Para la implementación en **R** y más base teórica y práctica se puede consultar Nowosad ([2019](#ref-nowosad2019)) y Pebesma & Bivand ([2020](#ref-pebesma2020)).

## Conceptos básicos

En esta sección se definen algunos conceptos fundamentales en geoestadística, que forman las bases teóricas y prácticas para el análisis geoestadístico.

### Correlación espacial

El concepto fundamental en geoestadística y la estadística espacial en general, es que las observaciones son dependientes 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 & Delfiner, [1999](#ref-chiles1999); Cressie, [1993](#ref-cressie1993); Goovaerts, [1997](#ref-goovaerts1997); Isaaks & Srivastava, [1989](#ref-isaaks1989); Webster & Oliver, [2007](#ref-webster2007)).

### Semivarianza

Es la medida que se usa para determinar la disimilitud entre observaciones que varían con la distancia, y se representa mediante la Ecuación (1), donde es el valor de la variable en la posición , es el valor de la variable a una distancia , es el número total de puntos (observaciones), y es el número de pares de puntos que se encuentran a una distancia específica. **Se recomienda tener más de 30 pares de puntos por cada distancia , y no calcular la semivarianza más allá de la mitad de la máxima distancia entre observaciones** (Chilès & Delfiner, [1999](#ref-chiles1999); Goovaerts, [1997](#ref-goovaerts1997); Isaaks & Srivastava, [1989](#ref-isaaks1989); Webster & Oliver, [2007](#ref-webster2007)).

Si los datos se encuentran ordenados en una grilla regular se puede usar la separación entre puntos como las diferentes distancias (Figura 1). Si los datos se encuentran irregularmente espaciados es necesario agruparlos en franjas (Figura 2), donde se requiere definir una tolerancia de la distancia (, por lo general ), y una tolerancia angular () (Oliver & Webster, [2014](#ref-oliver2014c); Webster & Oliver, [2007](#ref-webster2007)).

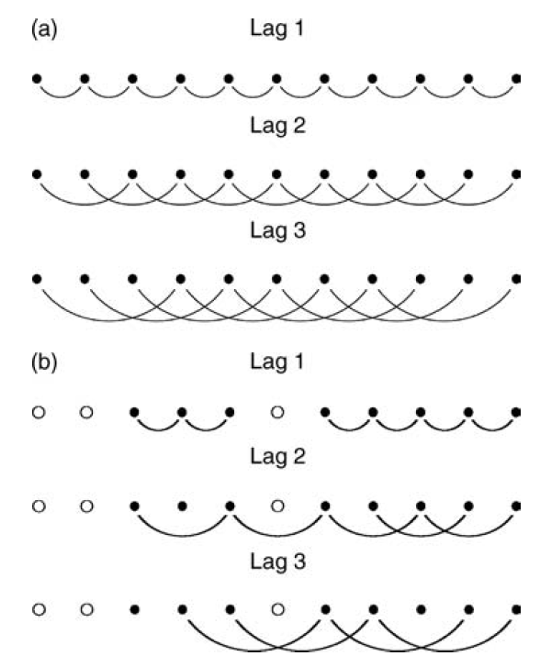


Figura 1: Esquema del calculo de la semivarianza para diferentes distancias donde los datos están completos (a) y donde hay vacíos de datos (b). Tomado de Webster & Oliver ([2007](#ref-webster2007)).

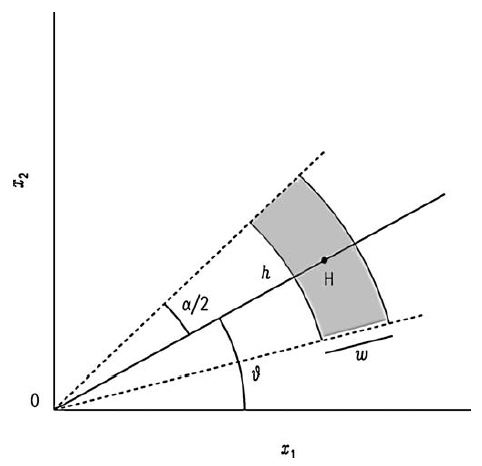


Figura 2: Esquema del calculo de la semivarianza para diferentes distancias donde los datos están irregularmente espaciados. Tomado de Webster & Oliver ([2007](#ref-webster2007)).

### Variograma experimental

Para visualizar la relación entre la semivarianza y la distancia (relación espacial de la variable) se usa el variograma experimental (Figura 3).

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 (). Conforme se varíen estos valores va a variar la semivarianza y cualquier ajuste que se le realice y su posterior uso en la interpolación* (Isaaks & Srivastava, [1989](#ref-isaaks1989); Oliver & Webster, [2014](#ref-oliver2014c); Webster & Oliver, [2007](#ref-webster2007)).

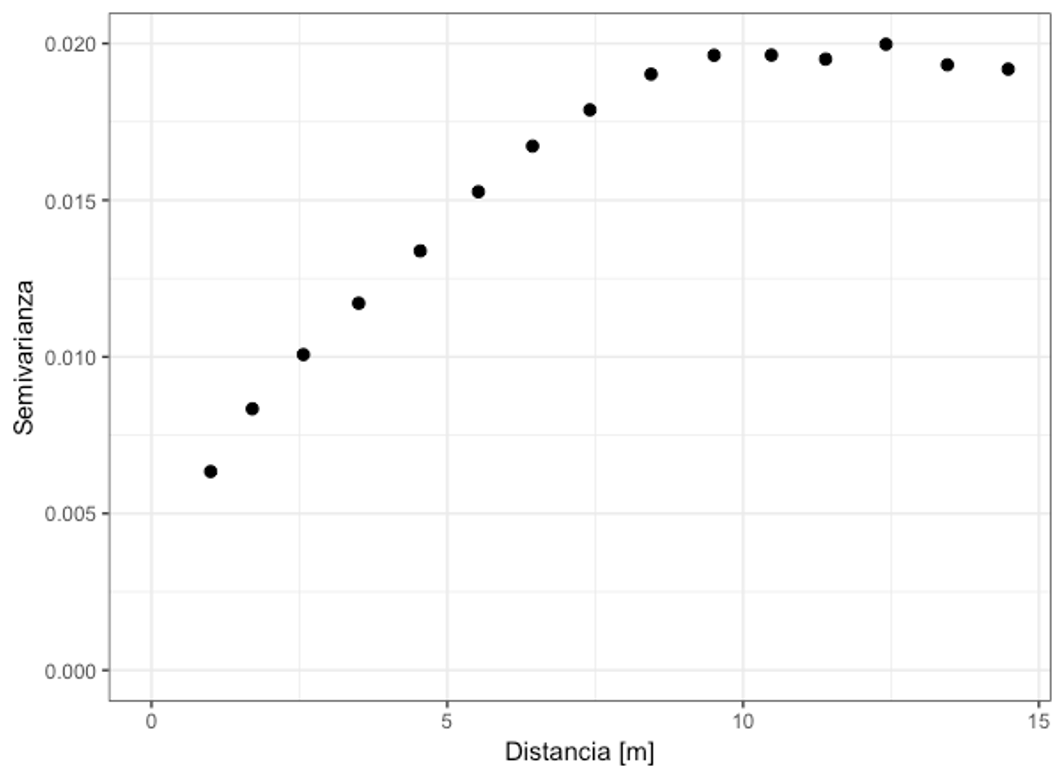


Figura 3: Ejemplo de variograma experimental, mostrando la relación espacial de la variable.

### 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 que lo van a definir, conforme se muestra en la Figura 4 (Goovaerts, [1997](#ref-goovaerts1997); Sarma, [2009](#ref-sarma2009); Webster & Oliver, [2007](#ref-webster2007)).

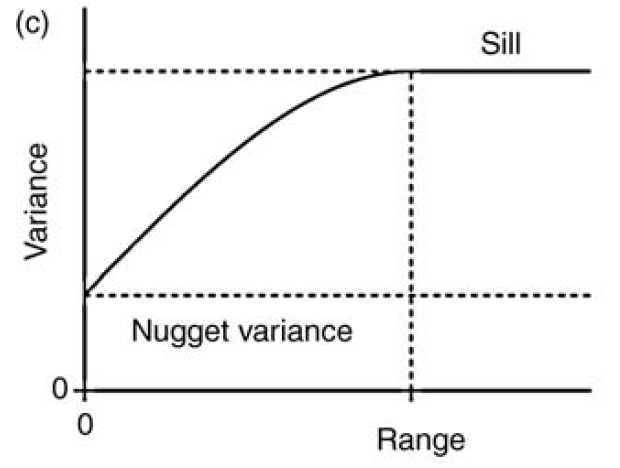


Figura 4: Modelo de variograma mostrando las partes: meseta, pepita, y rango. Tomado de Webster & Oliver ([2007](#ref-webster2007)).

Las partes del semivariograma son (Isaaks & Srivastava, [1989](#ref-isaaks1989); Sarma, [2009](#ref-sarma2009); Webster & Oliver, [2007](#ref-webster2007)):

* Meseta total (, sill en inglés): Valor del variograma o semivarianza cuando la distancia tiende a infinito, y por lo general es muy similar al valor de la varianza de la variable de interés.
* Meseta parcial (, partial sill en inglés): La diferencia entre la meseta total y la pepita (). Si no hubiera pepita (), entonces .
* Pepita (, 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 (, range en inglés): 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.

#### Tipos

Aquí se exponen los principales tipos de modelos que se usan en geociencias (Goovaerts, [1997](#ref-goovaerts1997); Isaaks & Srivastava, [1989](#ref-isaaks1989); Sarma, [2009](#ref-sarma2009); Webster & Oliver, [2007](#ref-webster2007)).

* Potencia

Es más usado cuando el variograma no se estabiliza o alcanza una meseta. Se calcula mediante la Ecuación (2), donde es la pendiente, y controla la concavidad o convexidad del modelo. Un ejemplo se muestra en la Figura 5.

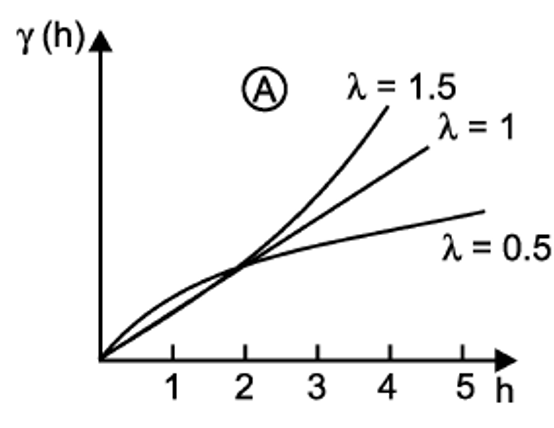


Figura 5: Modelo de potencia. Tomado de Sarma ([2009](#ref-sarma2009)).

* 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 6.

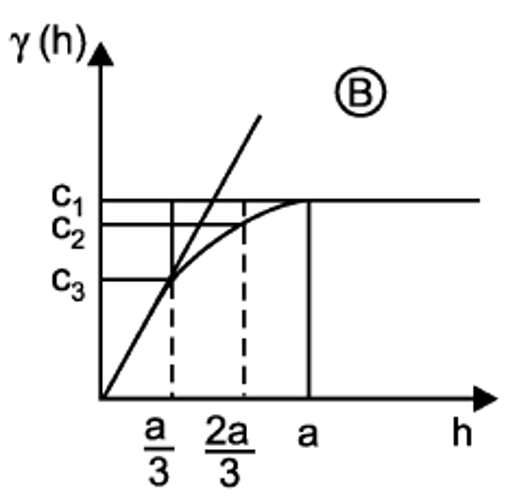


Figura 6: Modelo esférico. Tomado de Sarma ([2009](#ref-sarma2009)).

* 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 , o sea, una tercera parte del rango esperado. Se calcula mediante la Ecuación (4) y un ejemplo se muestra en la Figura 7.

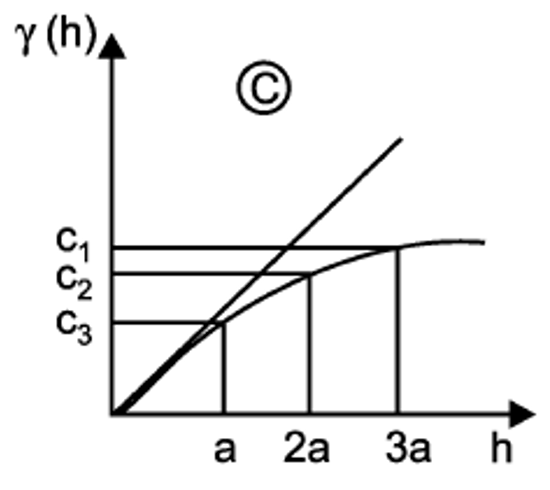


Figura 7: Modelo exponencial. Tomado de Sarma ([2009](#ref-sarma2009)).

* 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 , o sea, el rango esperado entre la raíz de 3. Se calcula mediante la Ecuación (5), y un ejemplo se muestra en la Figura 8.

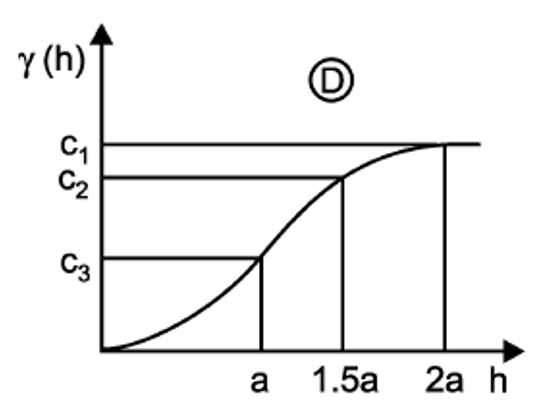


Figura 8: Modelo gaussiano. Tomado de Sarma ([2009](#ref-sarma2009)).

La Figura 9 es una comparación de los tres modelos más comunes en geociencias, donde todos corresponden con una estructura que presenta los siguientes parámetros: , , y . 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.

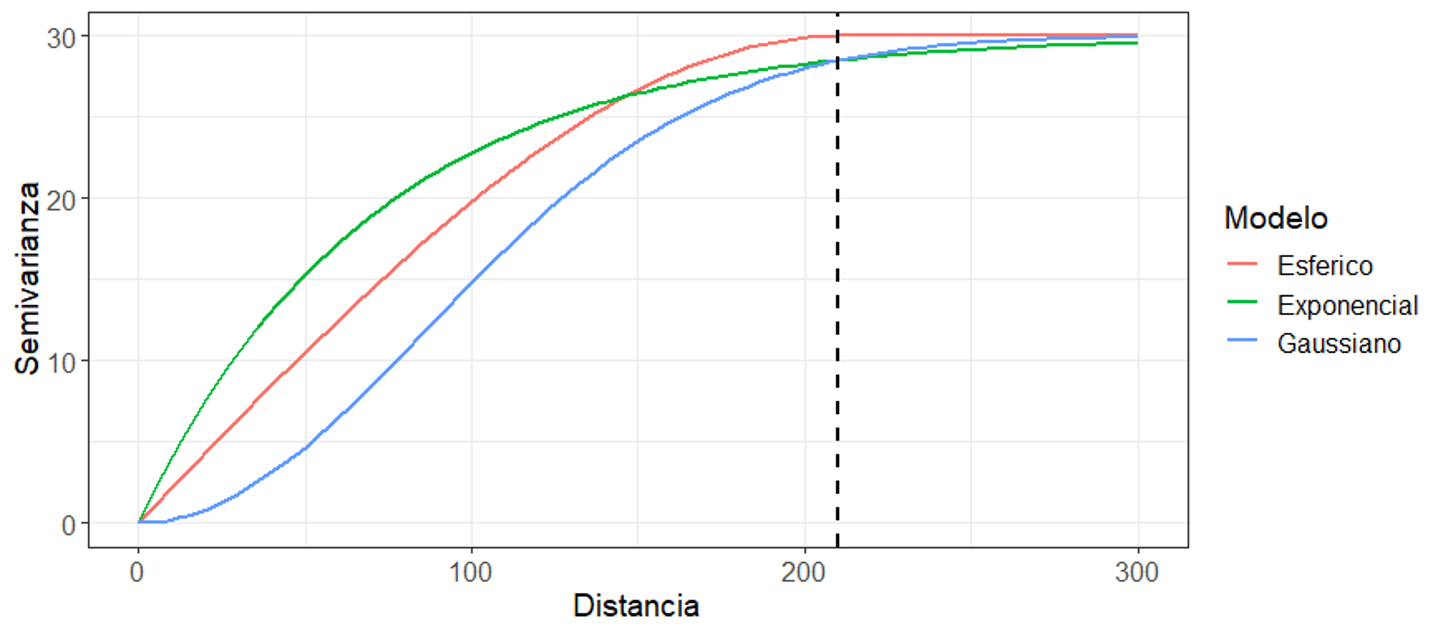


Figura 9: Comparación visual de los tres modelos más usados en geociencias, todos respresentando la misma estructura (, , y ).

### 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 (Chilès & Delfiner, [1999](#ref-chiles1999); Goovaerts, [1997](#ref-goovaerts1997); Isaaks & Srivastava, [1989](#ref-isaaks1989); Oliver & Webster, [2014](#ref-oliver2014c); Webster & Oliver, [2007](#ref-webster2007)).

La anisotropía puede ser de dos tipos: geométrica o zonal. La geométrica es al 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 & Delfiner, [1999](#ref-chiles1999); Goovaerts, [1997](#ref-goovaerts1997); Isaaks & Srivastava, [1989](#ref-isaaks1989); Webster & Oliver, [2007](#ref-webster2007)).

Para determinar la presencia o no de anisotropía se pueden usar el mapa de la superficie de variograma (Figura 10 **A**) y/o variogramas direccionales (Figura 10 **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 & Delfiner, [1999](#ref-chiles1999); Goovaerts, [1997](#ref-goovaerts1997); Isaaks & Srivastava, [1989](#ref-isaaks1989); Webster & Oliver, [2007](#ref-webster2007)).

En la Figura 10 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.

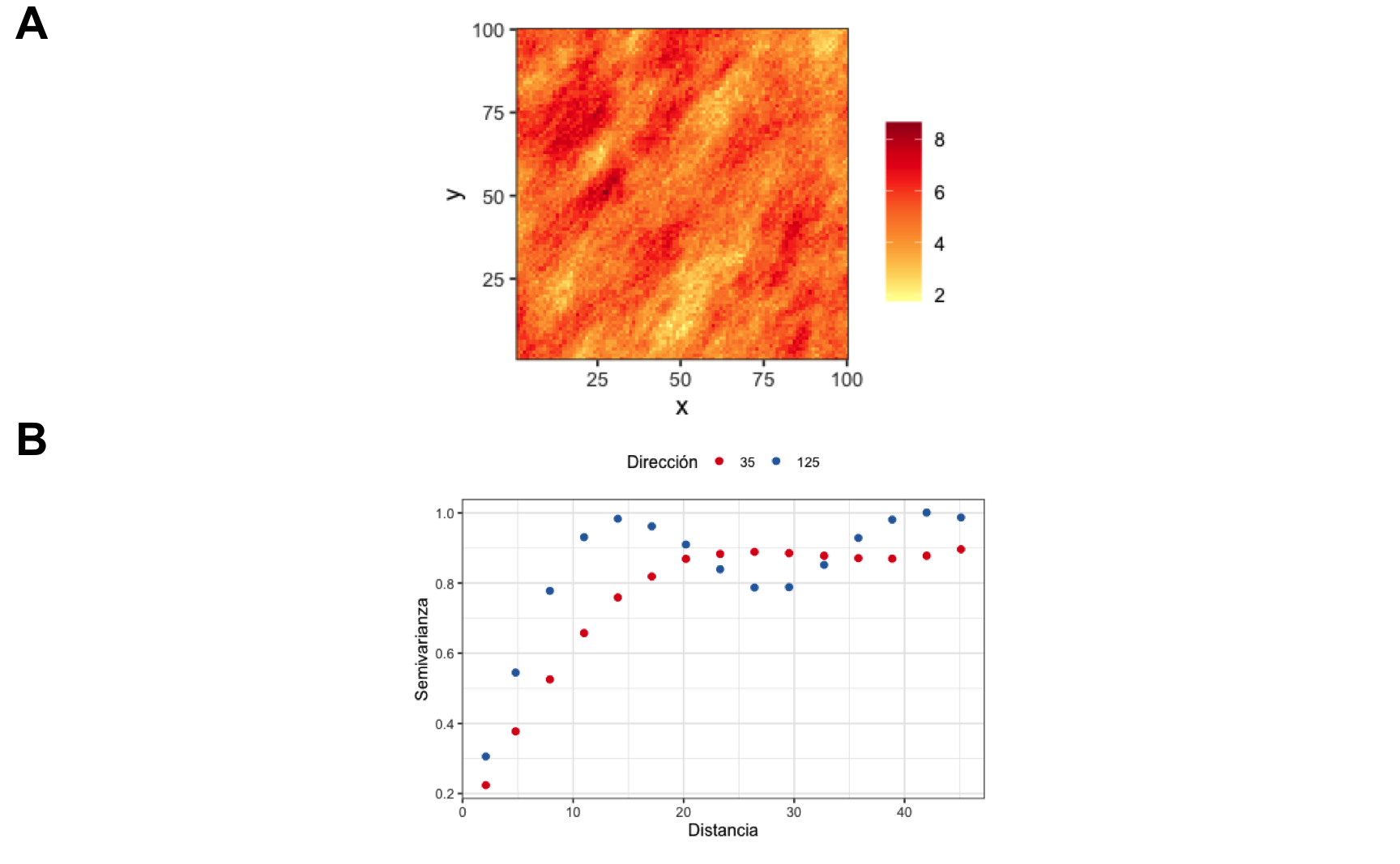


Figura 10: **A** Ejemplo de mapa de la superficie de variograma, mostrando anisotropía donde el eje principal ocurre en la dirección 35 y el eje menor ocurre en la dirección 125. **B** Variogramas direccionales donde se observa como en la dirección de 35 se alcanza un rango mayor (), mientras que en la dirección perpendicular (125) el rango es menor ()

### Validación cruzada

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 y con el modelo ajustado se predice el valor de la variable para esas observaciones (Chilès & Delfiner, [1999](#ref-chiles1999); Goovaerts, [1997](#ref-goovaerts1997); Isaaks & Srivastava, [1989](#ref-isaaks1989); Oliver & Webster, [2014](#ref-oliver2014c); Webster & Oliver, [2007](#ref-webster2007)).

El tipo de validación cruzada más usado es *LOO* (leave-one-out), donde se deja por fuera una observación a la vez y se predice el valor de la variable para cada observación por separado (Goovaerts, [1997](#ref-goovaerts1997); Isaaks & Srivastava, [1989](#ref-isaaks1989); Oliver & Webster, [2014](#ref-oliver2014c); Webster & Oliver, [2007](#ref-webster2007)). 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, donde se predice el valor de variable para todas las observaciones del grupo; este proceso se repite K veces.

Una vez realizado el ajuste y la validación cruzada del mismo se obtienen valores predecidos 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 & Delfiner, [1999](#ref-chiles1999); Goovaerts, [1997](#ref-goovaerts1997); Isaaks & Srivastava, [1989](#ref-isaaks1989); Oliver & Webster, [2014](#ref-oliver2014c); Webster & Oliver, [2007](#ref-webster2007)).

Dentro de las métrica más usadas están (Oliver & Webster, [2014](#ref-oliver2014c); Webster & Oliver, [2007](#ref-webster2007); Yao et al., [2013](#ref-yao2013po)):

En estas métricas es el total de observaciones, es el valor observado en el punto , es el valor predecido en el punto , es el error/varianza de Kriging, y es la media (promedio) de la variable.

* Error medio (): El error corresponde con los residuales de lo observado menos lo predecido, 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.
* Error cuadrático medio (): Este valor corresponde con la desviación promedio de los errores al cuadrado; idealmente se prefieren valores pequeños. Se calcula mediante la Ecuación (7) y comparando modelos se escogería el modelo que presente un menor.
* Razón de desviación cuadrática media (): Esta valor compara la diferencia entre la predicción y valor actual con respecto a la varianza (error) obtenida de la interpolación (). Se esperaría que este valor ande cerca de 1. Se calcula mediante la Ecuación (8) y comparando modelos se escogería el que presente un más cercano a 1.
* Error Porcentual Absoluto Medio (): Es una medida porcentual de la diferencia entre lo observado y lo predecido, 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 más bajo.
* Estadístico de bondad de predicción (): 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 la media, valores negativos indican que el modelo es menos efectivo que 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 más cercano a 1.

### Kriging

Kriging es un método de interpolación (estimación) donde la idea es obtener valores de la variable en lugares donde no se obtuvo muestra. 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 influencia mínima o nula (Chilès & Delfiner, [1999](#ref-chiles1999); Goovaerts, [1997](#ref-goovaerts1997); Isaaks & Srivastava, [1989](#ref-isaaks1989); Webster & Oliver, [2007](#ref-webster2007)). Lo anterior se presenta de manera gráfica en la Figura 11.

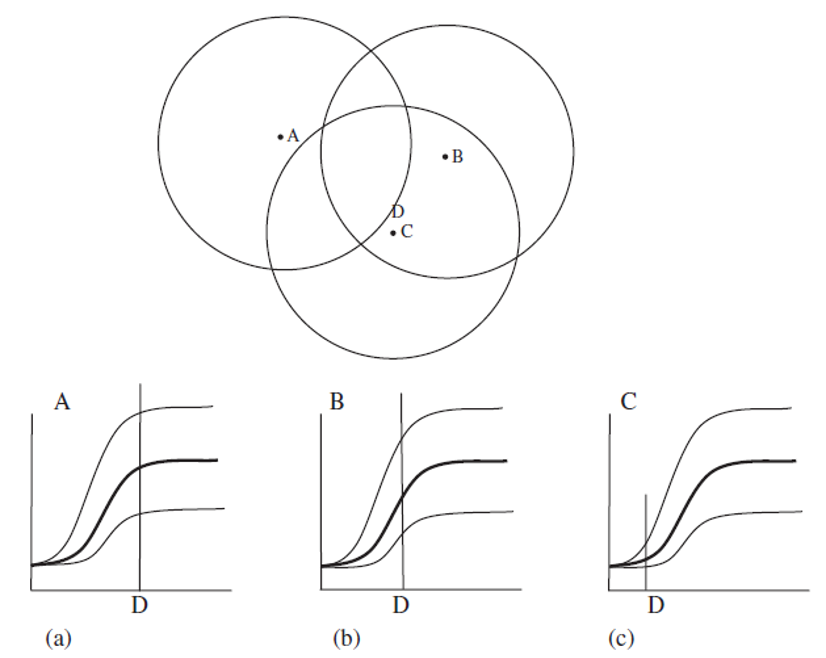


Figura 11: Visualización del proceso de interpolación mediante Kriging, donde para el punto a interpolar (D), el punto que está más cercano (C) tiene más peso (influencia, baja semivarianza), y el punto más lejano (A) prácticamente no tiene peso ya que cae fuera del rango. Tomado de McKillup & Darby Dyar ([2010](#ref-mckillup2010)).

Dentro de las ventajas del Kriging están (Chilès & Delfiner, [1999](#ref-chiles1999); Goovaerts, [1997](#ref-goovaerts1997); Isaaks & Srivastava, [1989](#ref-isaaks1989); Trauth, [2015](#ref-trauth2015); Webster & Oliver, [2007](#ref-webster2007)): 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).

El resultado de la interpolación por medio Kriging presenta las siguientes características (Figura 12): suaviza los resultados, y sobre-estima valores pequeños y sub-estima valores grandes (Oliver & Webster, [2014](#ref-oliver2014c); Webster & Oliver, [2007](#ref-webster2007)).

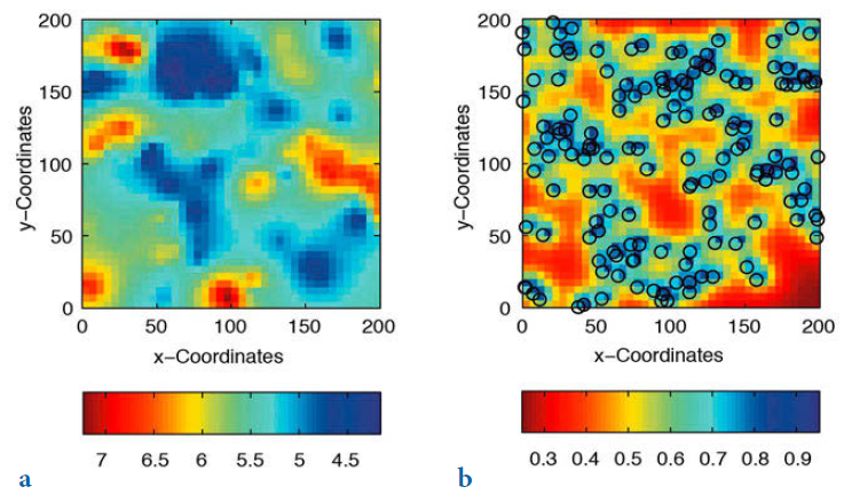


Figura 12: Ejemplo del resultado de interpolación (a), y su error (b), por medio de Kriging. Tomado de Trauth ([2015](#ref-trauth2015)).

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. **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), y son estacionarios**, esto último implica que la media y varianza de la variable no varían significativamente (Chilès & Delfiner, [1999](#ref-chiles1999); Goovaerts, [1997](#ref-goovaerts1997); Isaaks & Srivastava, [1989](#ref-isaaks1989); Webster & Oliver, [2007](#ref-webster2007)).

Los tipos de Kriging más comunes son (Chilès & Delfiner, [1999](#ref-chiles1999); Goovaerts, [1997](#ref-goovaerts1997); Isaaks & Srivastava, [1989](#ref-isaaks1989); Webster & Oliver, [2007](#ref-webster2007)):

* *Simple ()*: 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 ()*: Esta variante es la más usada, donde se asume una media constante pero desconocida, y adicionalmente los datos no deben presentar una tendencia.
* *Lognormal ()*: 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. **Para obtener el resultado de la interpolación en la escala original de la variable NO es tan simple como exponenciar los resultados. Cressie (**[**1993**](#ref-cressie1993)**), Webster & Oliver (**[**2007**](#ref-webster2007)**), Laurent (**[**1963**](#ref-laurent1963jasa)**), y Yamamoto (**[**2007**](#ref-yamamoto2007cg)**) brindan más detalles de cómo realizar la transformación inversa de la manera más apropiada.**
* *Universal ()*: 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 (en función de las coordenadas), como el caso típico de niveles piezométricos. Lark et al. ([2006](#ref-lark2006ejss)) brinda más detalles y técnicas más actualizadas de como lidiar con este tipo de situación.
* *CoKriging ()*: 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 ()*: 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 & Garcia ([2010](#ref-eldeiry2010jide)), Kravchenko & Bullock ([1999](#ref-kravchenko1999a)), Meng et al. ([2013](#ref-meng2013cagis)), Wang et al. ([2017](#ref-wang2017rs)), y Yao et al. ([2013](#ref-yao2013po)) hacen uso de varios de los tipos de Kriging 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 variable en el espacio) en un set de datos simulados. Se usan datos simulados para que el enfoque sea más en el proceso y no tanto en la variable en si, la cual puede ser porosidad, densidad, espesor, etc., o cualquier variable numérica de interés.

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 et al., [2015](#ref-finley2015jss); Jing & Oliveira, [2015](#ref-jing2015jss); Pebesma & Graeler, [2020](#ref-R-gstat); Ribeiro et al., [2003](#ref-ribeiro2003p3iwdsc)). Aquí se presenta el uso del paquete *gstat* (Pebesma & Graeler, [2020](#ref-R-gstat)), que es uno de los más usados, ya presenta una gran cantidad de funciones disponibles. Para la manipulación de los datos se usan principalmente *dplyr* (Wickham, François, et al., [2020](#ref-R-dplyr)), *tidyr* (Wickham & Henry, [2020](#ref-R-tidyr)) y *broom* (Robinson & Hayes, [2020](#ref-R-broom)), y para la creación de gráficos se usa *ggplot2* (Wickham, [2016](#ref-ggplot22016); Wickham, Chang, et al., [2020](#ref-R-ggplot2)).

## 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 aproximad de la meseta total del variograma.

Primeramente se deben importar los datos en un data frame (tabla) al cual se le llama datos. En este caso la variable de interés es z. Una vez los datos están listos se procede a analizar la variable y ver su distribución como se muestra a continuación.

datos = import(here('data','datos.csv'))  
  
myvar = 'z' # variable a modelar  
  
descr(datos[myvar],style = 'rmarkdown') %>%   
 as.data.frame() %>%   
 slice(-c(12,13,15)) %>%   
 kable(digits = 2,  
 format = 'simple',  
 caption = 'Resumen estadístico de la variable.')

Tabla 1: Resumen estadístico de la variable.

|  |  |
| --- | --- |
|  | z |
| Mean | 29.90 |
| Std.Dev | 0.86 |
| Min | 28.19 |
| Q1 | 29.20 |
| Median | 29.86 |
| Q3 | 30.38 |
| Max | 31.95 |
| MAD | 0.94 |
| IQR | 1.15 |
| CV | 0.03 |
| Skewness | 0.47 |
| N.Valid | 60.00 |

(S = var(datos[[myvar]])) # varianza de la variable

## [1] 0.7452631

gg.hist = ggplot(datos, aes\_string(myvar)) +   
 geom\_histogram(aes(y = stat(density)), bins = 10,   
 col = 'black', fill = 'blue', alpha = .5) +   
 geom\_vline(xintercept = mean(datos[[myvar]]), col = 'red') +  
 geom\_density(col = 'blue') +  
 labs(y = 'Densidad')  
gg.hist

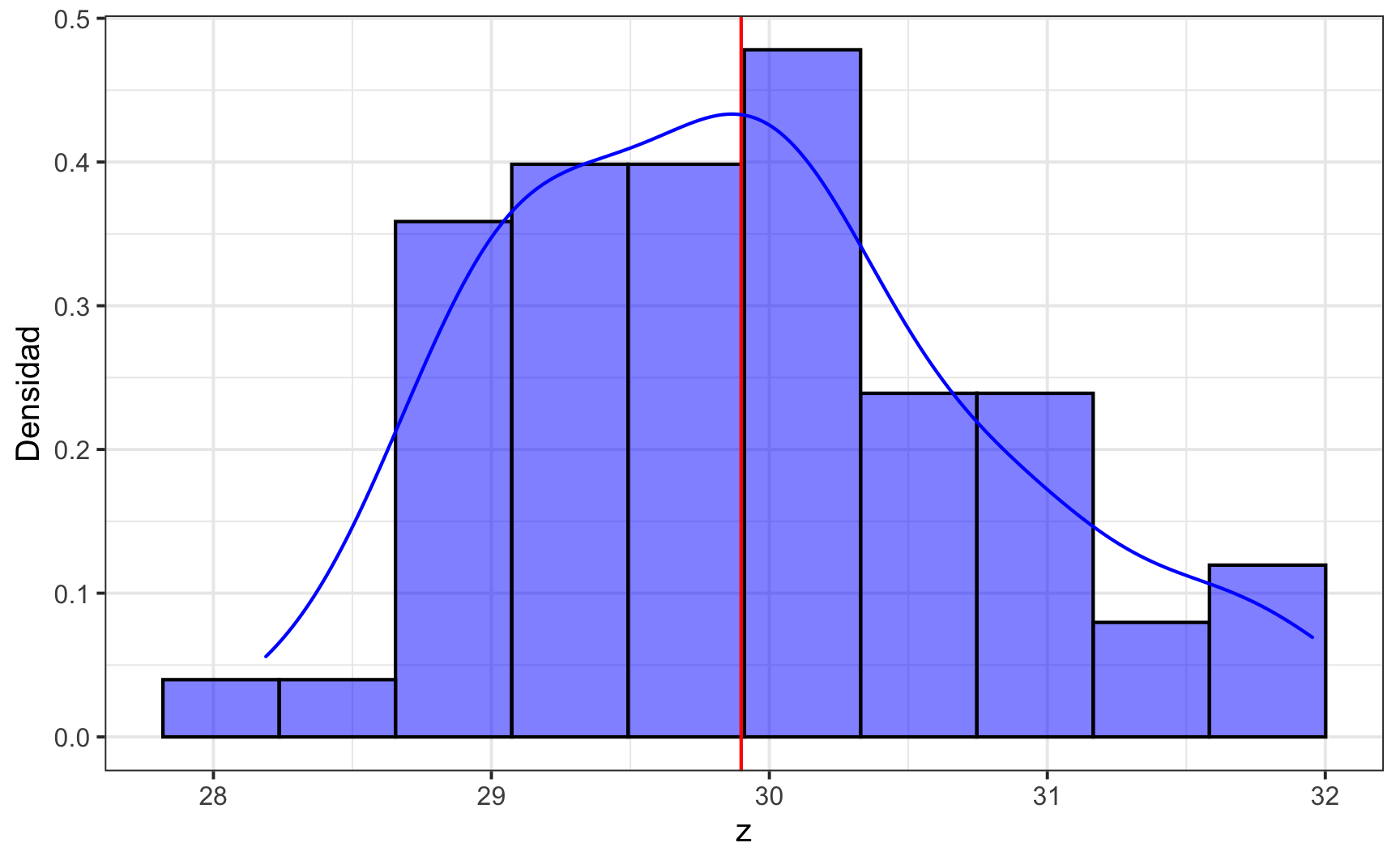


Figura 13: Distribución de los datos.

El resumen estadístico (Tabla 1) y el histograma (Figura 13) muestran que los datos tienen una distribución aproximadamente normal, donde la media y mediana son similares y el histograma presenta una forma general de campana, por lo que no es necesaria ninguna transformación.

Como los datos iniciales están en una tabla, es necesario convertir estos datos en datos/objetos espaciales, para poder realizar operaciones y análisis previos al análisis geoestadístico.

En este caso los datos tienen una grilla de coordenadas local la cual no corresponde con ningún sistema de coordenadas reconocido. De manera general se recomienda trabajar los datos en sistemas de coordenadas planas (x,y) por lo que si se tienen en geográficas (long, lat) se recomienda convertirlas a planas conforme la zona de estudio. Para esto se puede consultar Garnier-Villarreal ([2020](#ref-garnier-villarreal2020)), donde se presenta un capítulo dedicado al trato de datos espaciales en **R**, y se indica como transformar los datos de un sistema de coordenadas a otro.

Para crear el objeto espacial se usa la función st\_as\_sf del paquete *sf* (Pebesma, [2020a](#ref-R-sf)), donde los argumentos requeridos son los datos, las columnas donde están las coordenadas (x,y), y opcionalmente el código del sistema de coordenadas. Como en este caso los datos no corresponden con ninguna sistema de coordenadas se pone NA, pero si no se usaría el código epsg.

datos\_sf = st\_as\_sf(datos, coords = 1:2, crs = NA) %>%   
 mutate(X = st\_coordinates(.)[,1], Y = st\_coordinates(.)[,2]) %>%   
 relocate(X, Y)  
datos\_sp = as(datos\_sf, 'Spatial')  
coordnames(datos\_sp) = c('X','Y')

Una vez transformados los datos a datos espaciales 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. El siguiente código calcula las distancias.

dists = st\_distance(datos\_sf) %>% .[lower.tri(.)] %>% unclass()  
distancias = signif(c(min(dists), mean(dists), max(dists)),6) # rango de distancias  
names(distancias) = c('min', 'media', 'max')   
distancias

## min media max   
## 1.41421 51.09850 128.03500

Una forma de mostrar el área de influencia de los resultados es tener un polígono que encierra a los datos, ya que sería esta el área donde los resultados son realmente válidos. El siguiente código realiza el cálculo de dicho polígono.

outline = st\_convex\_hull(st\_union(datos\_sf))

Como se desea tener una superficie con datos interpolados (en puntos donde no se tiene muestra) es necesario generar una grilla a rellenar. Se pueden tener grillar regulares (rectangulares) o irregulares (conforme el polígono que encierra a los datos). El siguiente código crea estas grillas, generando objetos stars (Pebesma, [2020b](#ref-R-stars)) que son similares a objetos raster (Hijmans, [2020](#ref-R-raster)) pero más versátiles ya que permiten tener arreglos espaciales más complejos y diversos.

bb = st\_bbox(datos\_sf)  
dint = max(c(bb[3]-bb[1],bb[4]-bb[2])/nrow(datos\_sf))  
dx = seq(bb[1],bb[3],dint) # coordenadas x  
dy = seq(bb[4],bb[2],-dint) # coordenadas y  
st\_as\_stars(matrix(0, length(dx), length(dy))) %>%  
 st\_set\_dimensions(1, dx) %>%  
 st\_set\_dimensions(2, dy) %>%  
 st\_set\_dimensions(names = c("X", "Y")) %>%   
 st\_set\_crs(st\_crs(datos\_sf)) -> datosint  
  
datosint2 = st\_crop(datosint, outline)

Ya con los datos espaciales es bueno visualizar su distribución en el espacio. La Figura 14 muestra la ubicación de los datos, donde los puntos se encuentran rellenados de acuerdo al valor de la variable.

gg.map.pts = ggplot() +   
 geom\_sf(data = outline, col = 'cyan', alpha = .1, size = .75) +   
 geom\_sf(data = datos\_sf, aes\_string(col = myvar), size = 3, alpha = 0.6) +   
 scale\_color\_viridis\_c() +   
 labs(x = x\_map, y = y\_map) +  
 if (!is.na(st\_crs(datos\_sf))) {  
 coord\_sf(datum = st\_crs(datos\_sf))  
 }  
gg.map.pts

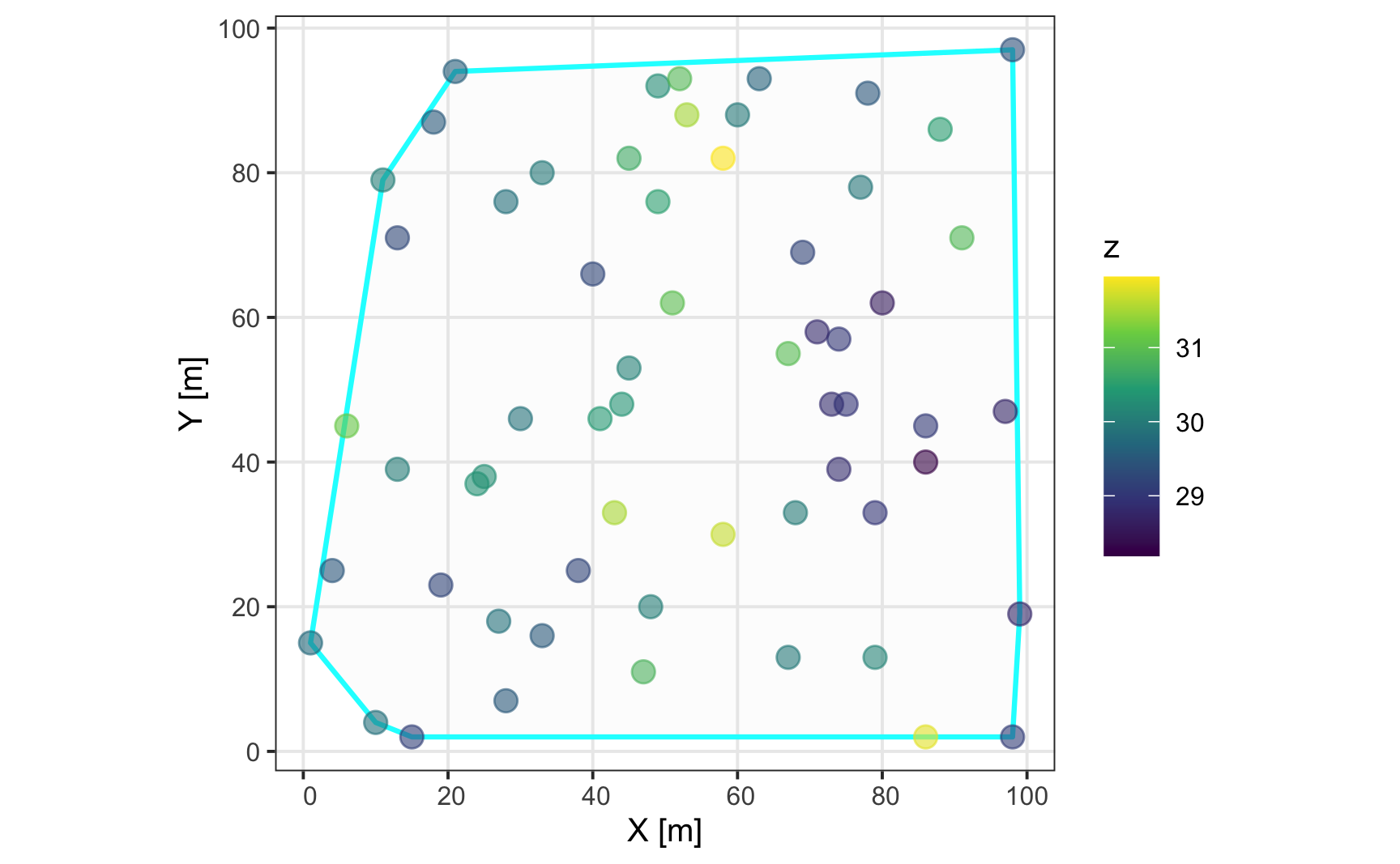


Figura 14: Mapa estático de la distribución espacial de los datos.

El siguiente código no se ejecuta, ya que genera un mapa interactivo que no se puede desplegar aquí, pero se incluye para referencia y en caso de que se quiera utilizar. Para este mapa se usa el paquete *mapview* (Appelhans et al., [2020](#ref-R-mapview)) y para los colores se usa el paquete *RColorBrewer* (Neuwirth, [2014](#ref-R-RColorBrewer)).

mapview(outline, alpha.regions = 0, layer.name='Borde',   
 homebutton = F, legend = F, native.crs = F) +   
 mapview(datos\_sf, zcol = myvar, alpha=0.1,   
 layer.name = myvar, native.crs = F,   
 col.regions = brewer.pal(9,'YlOrRd'))

## 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.

### Variograma experimental

El primer paso es crear un variograma experimental omnidireccional (Figura 15). Aquí se empieza a hacer uso de *gstat*, donde es conveniente crear un objeto gstat en el cual se definen los datos a usar y la variable de interés. Para definir la variable de interés se usa la sintaxis de fórmula de la siguiente forma: variable ~ 1, que sería similar a definir un modelo lineal para sólo el intercepto.

Una vez definido este objeto, que se va a usar en varias instancias, se construye el variograma experimental con la función variogram. Esta función tiene como argumentos son el objeto gstat, el intervalo de distancia (width) y la distancia máxima a la cual calcular la semivarianza (cutoff); y si recordamos la distancia máxima era 128.03.

myformula = as.formula(paste(myvar,'~1'))  
g = gstat(formula = myformula,   
 data = datos\_sf) # objeto gstat para hacer geoestadistica  
  
# variograma experimental cada cierta distancia (width), y hasta cierta distancia (cutoff)  
dat.vgm = variogram(g,   
 width = 5,  
 cutoff = 60)

gg.omni.exp = ggplot(dat.vgm,aes(x = dist, y = gamma)) +   
 geom\_point(size = 2) +   
 labs(x = x\_var, y = y\_var) +  
 geom\_hline(yintercept = S, col = 'red', linetype = 2) +  
 ylim(0, max(dat.vgm$gamma)) +  
 xlim(0, max(dat.vgm$dist)) +   
 geom\_text\_repel(aes(label = np), size = 3)  
gg.omni.exp

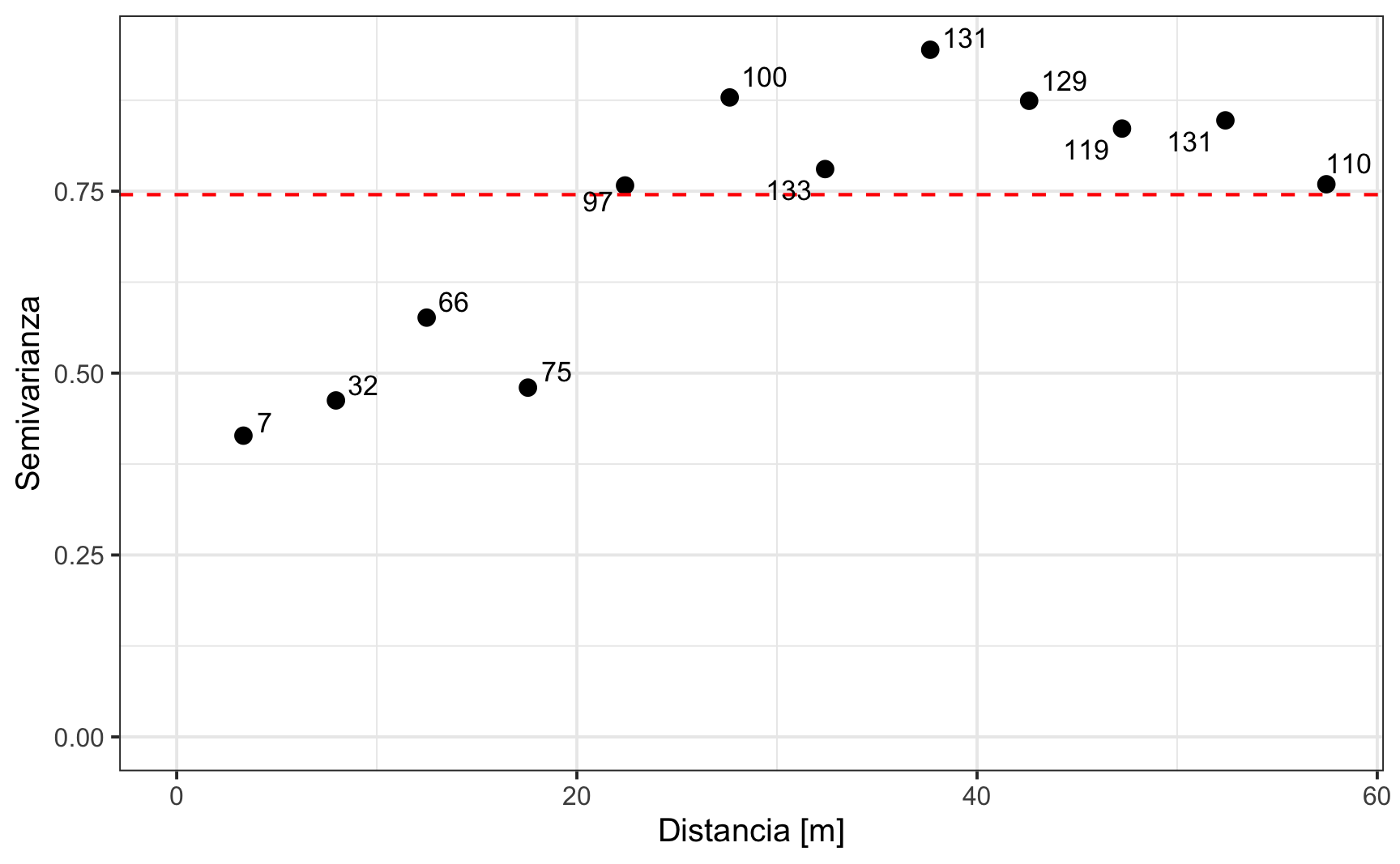


Figura 15: Variograma omnidireccional de los datos. Las etiquetas de los puntos corresponden con el número de pares de puntos usados para el cálculo de la semivarianza. La línea roja punteada corresponde con la varianza de los datos, que es una aproximación a la meseta de los datos.

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 (Figura 16), como los variogramas direccionales (Figura 17).

Para el mapa de la superficie de variograma los argumentos necesarios son la extensión (cutoff, misma que le variograma experimental), el ancho del pixel (width, no es el mismo que para el variograma, por lo general mayor), y definir que es un mapa (map=TRUE).

map.vgm <- variogram(g,   
 width = 10,   
 cutoff = 60,   
 map = TRUE)

gg.map.exp = ggplot(data.frame(map.vgm), aes(x = map.dx, y = map.dy, fill = map.var1)) +  
 geom\_raster() +   
 scale\_fill\_gradientn(colours = plasma(20)) +  
 labs(x = x\_vmap, y = y\_vmap, fill = "Semivarianza") +  
 coord\_equal()  
gg.map.exp

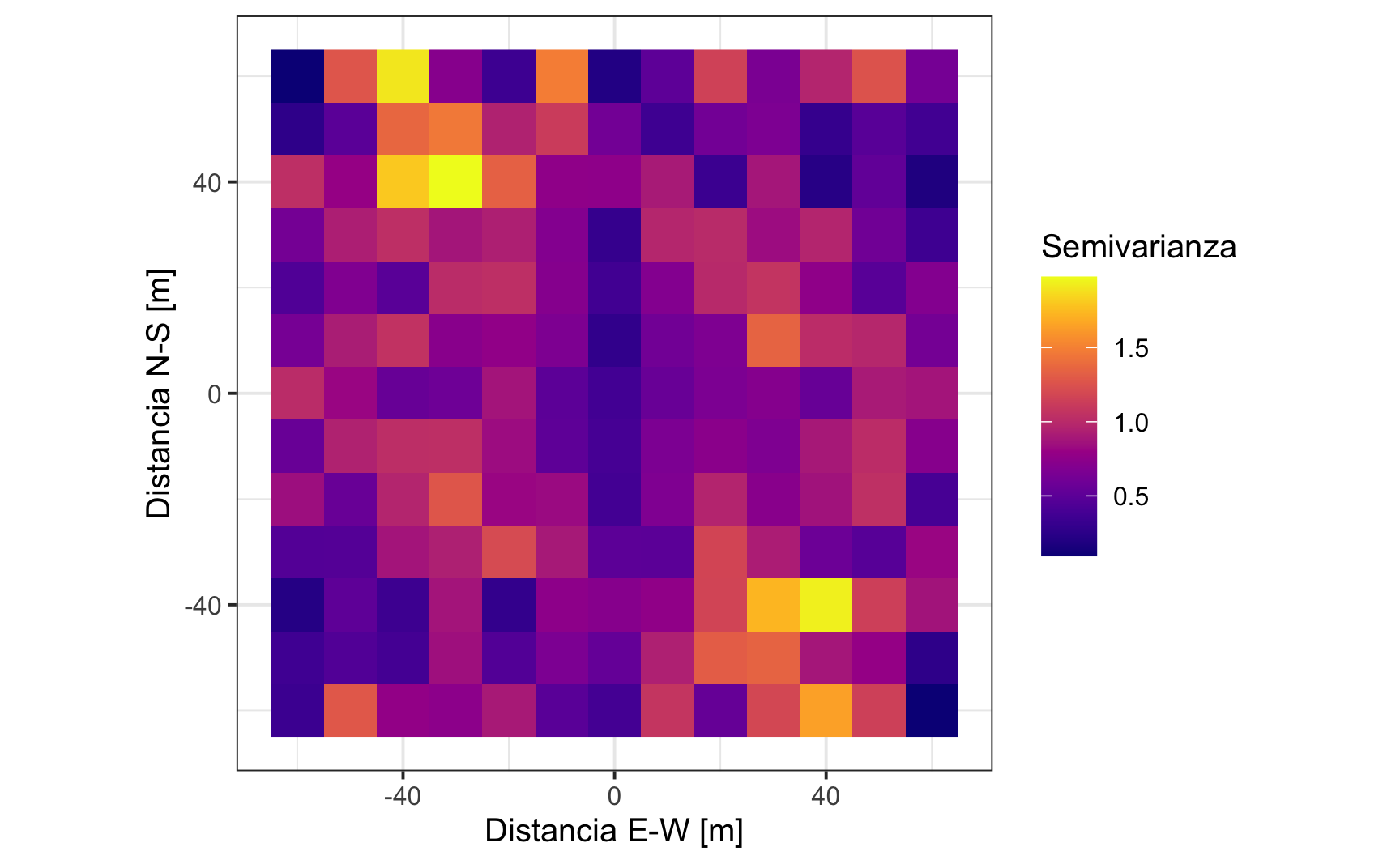


Figura 16: Mapa de la superficie de variograma para los datos.

Para los variogramas direccionales hay que definir los argumentos de las direcciones (alpha) y la tolerancia angular (tol.hor), donde lo más usado son direcciones cada 45° y la tolerancia angular es la mitad del intervalo entre direcciones.

# con direcciones y tolerancia angular  
d = c(0,45,90,135) # direcciones  
dat.vgm2 = variogram(g,   
 alpha = d,  
 tol.hor = 22.5,   
 cutoff = 60)   
  
dat.vgm2.gg = dat.vgm2 %>%   
 mutate(dir.hor = factor(dir.hor, labels = as.character(d)))

gg.dir.exp = ggplot(dat.vgm2.gg,aes(x = dist, y = gamma,  
 col = dir.hor, shape = dir.hor)) +   
 geom\_point(size = 2) +   
 labs(x = x\_var, y = y\_var, col = "Dirección", shape = 'Dirección') +  
 geom\_hline(yintercept = S, col = 'red', linetype = 2) +  
 ylim(0, max(dat.vgm2$gamma)) +  
 xlim(0, max(dat.vgm2$dist)) +   
 scale\_color\_brewer(palette = 'Dark2') +  
 facet\_wrap(~dir.hor) +   
 geom\_text\_repel(aes(label = np), size = 3, show.legend = F) +  
 theme(legend.position = 'top')  
gg.dir.exp

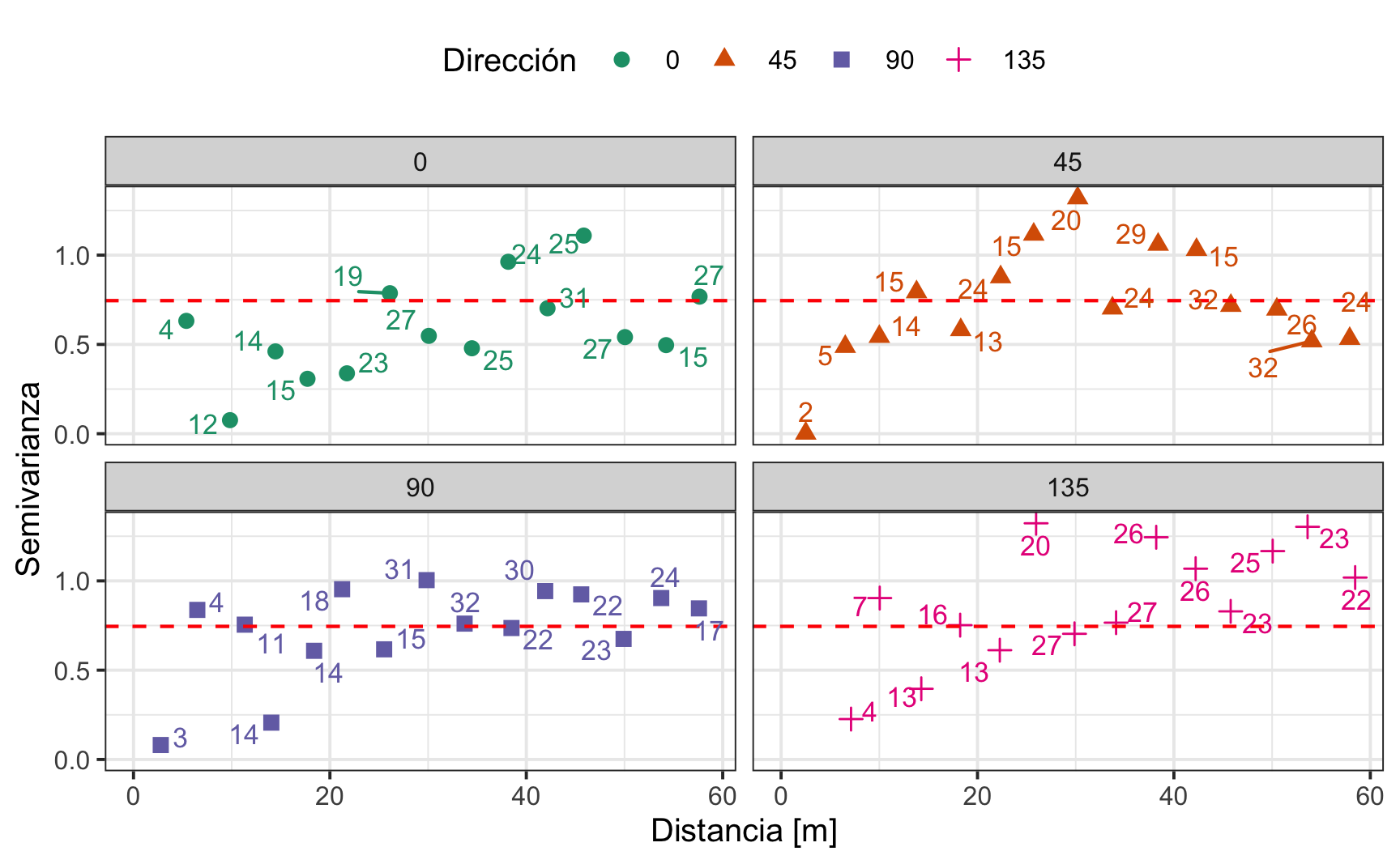


Figura 17: Variogramas direccionales cada 45°. La línea roja punteada representa la varianza de los datos, lo que se aproxima a la meseta total.

Analizando el mapa y los variogramas direccionales se concluye que no muestran señas de anisotropía, por lo que el modelado se puede continuar con el variograma omnidireccional.

### 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) y determinar valores iniciales, para posteriormente realizar el ajuste.

Usando el variograma omnidireccional (Figura 15), se puede estimar una pepita de aproximadamente 0.25, una meseta parcial de 0.5, un rango de 30, y se puede usar un modelo tipo esférico (‘Sph’). Lo anterior se define de la siguiente manera:

pep = .25 # pepita  
meseta = .5 # meseta parcial  
mod = "Sph" # modelo a ajustar (esférico)  
rango = 30 # rango

El paquete *gstat* ya viene con modelos definidos, por lo que el usuario debe seleccionar el modelo que considera apropiado. Usando la función show.vgms (Figura 18) se pueden desplegar los diferentes modelos disponibles (nombre en comillas). Para los modelos mencionados en la parte de teoría los nombres usados por *gstat* serían: ‘Sph’ para esférico, ‘Exp’ para exponencial, ‘Gau’ para gaussiano, y ‘Pot’ para potencia.

show.vgms()

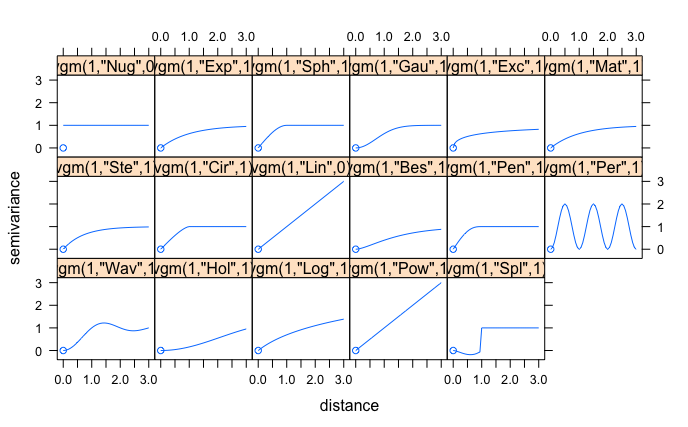


Figura 18: Modelos disponibles en *gstat*

Una vez definidos los valores iniciales se usa la función fit.variogram para realizar el ajuste automático. Los argumentos de la función son el variograma experimental y el modelo que se define por medio de la función vgm. Con el modelo ajustado (Tabla 2) se puede calcular un error del ajuste inicial, pero es más confiable el que se obtiene usando la validación cruzada, ya que el obtenido acá es un valor optimista.

dat.fit = fit.variogram(dat.vgm, model = vgm(meseta, mod, rango, pep))  
fit.rmse = sqrt(attributes(dat.fit)$SSErr/(nrow(datos))) # error del ajuste  
varmod = dat.fit # modelo ajustado

fit.rmse # error del ajuste

## [1] 0.01287286

varmod

Tabla 2: Modelo ajustado al variograma experimental

|  |  |  |
| --- | --- | --- |
| Modelo | Meseta | Rango |
| Nug | 0.336 | 0.000 |
| Sph | 0.533 | 44.838 |

Con el modelo ajustado se puede visualizar éste sobre el variograma omnidireccional (Figura 19) y los variogramas direccionales (Figura 20). Para poder graficar el modelo ajustado es necesario definir la dirección a la cual se quiere estimar y para esto hay que definirlo en un vector unitario. Para simplificar este paso se creó la función unit\_vector, que posteriormente se usa para crear un objeto (variog.dir) que va a contener la semivarianza ajustada para las direcciones definidas (en este caso el vector d).

unit\_vector = function(th) {  
 if (th == 0) {  
 uv = c(0,1,0)  
 } else if (th == 90) {  
 uv = c(1,0,0)  
 } else if (th == 180) {  
 uv = c(0,-1,0)  
 } else if (between(th,0,90)) {  
 uv = c(sin(th\*pi/180),cos(th\*pi/180),0)  
 } else if (between(th,90,180)) {  
 uv = c(cos((th-90)\*pi/180),-1\*sin((th-90)\*pi/180),0)  
 }  
 return(uv)  
}  
  
variog.dir = map\_dfr(d, ~variogramLine(object = varmod,   
 maxdist = max(dat.vgm$dist),  
 min = 0.001, n = 100,   
 dir = unit\_vector(.x)),   
 .id = 'dir.hor') %>%   
 as\_tibble() %>%   
 mutate(dir.hor = factor(dir.hor, labels = as.character(d)))

# plot(dat.vgm, dat.fit, xlab = x\_var, ylab = y\_var)   
  
# omnidireccional  
  
gg.omni.fit = variog.dir %>%   
 ggplot(aes(dist,gamma)) +   
 geom\_point(data = dat.vgm,shape=1,size=2) +  
 geom\_line(size=1) +  
 labs(x = x\_var, y = y\_var) +  
 coord\_cartesian(ylim = c(0,max(dat.vgm$gamma)))  
gg.omni.fit

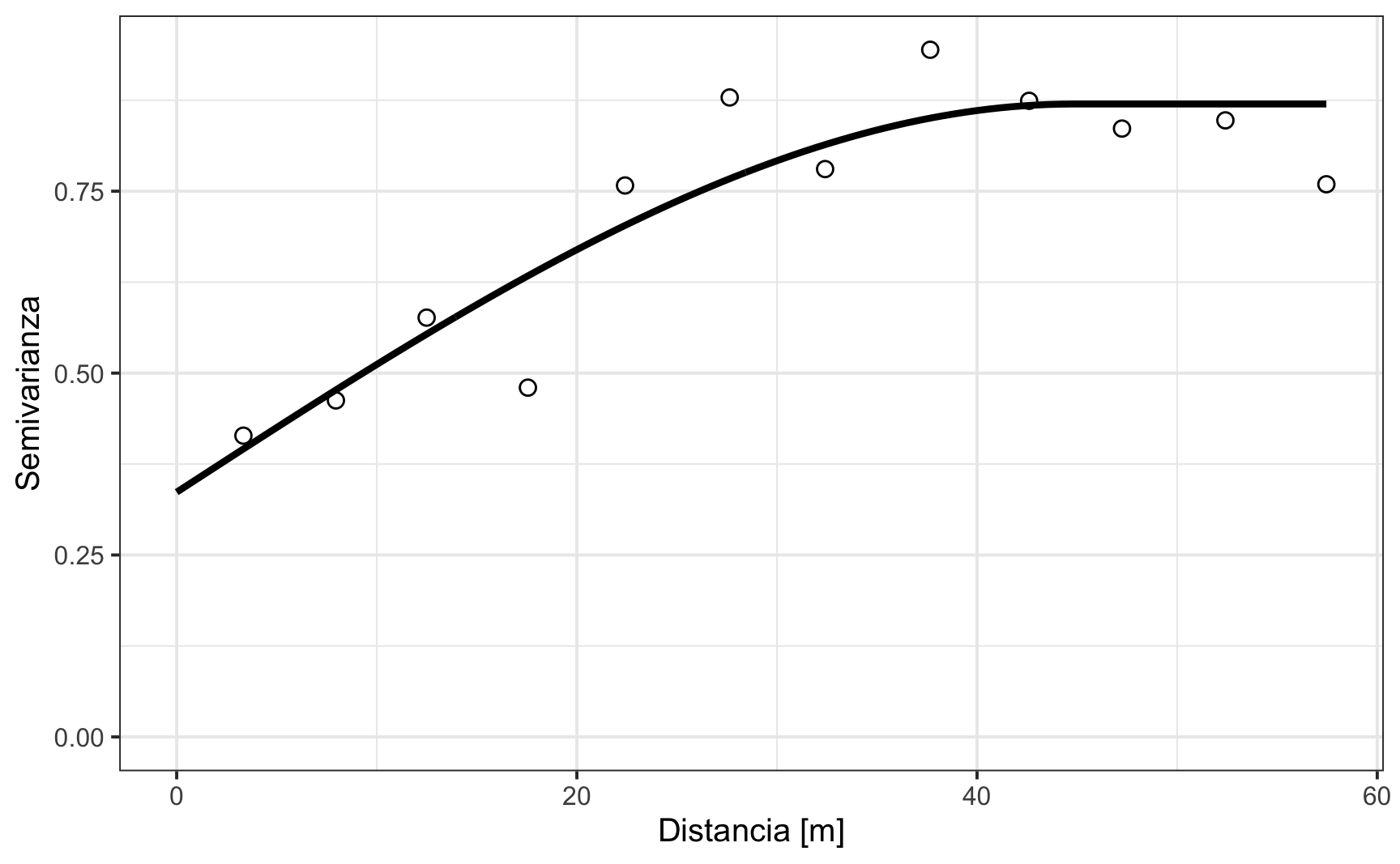


Figura 19: Variograma omnidireccional y modelo esférico ajustado.

# plot(dat.vgm2, dat.fit, as.table = T, xlab = x\_var, ylab = y\_var)   
  
# direccionales  
  
gg.dir.fit = variog.dir %>%   
 ggplot(aes(dist,gamma,col=dir.hor)) +   
 geom\_point(data = dat.vgm2.gg,shape=1,size=1.5) +  
 geom\_line(size=1) +   
 scale\_color\_brewer(palette = 'Dark2') +  
 labs(x = x\_var, y = y\_var, col = 'Dirección') +  
 facet\_wrap(~dir.hor) +  
 theme(legend.position = 'top')  
gg.dir.fit

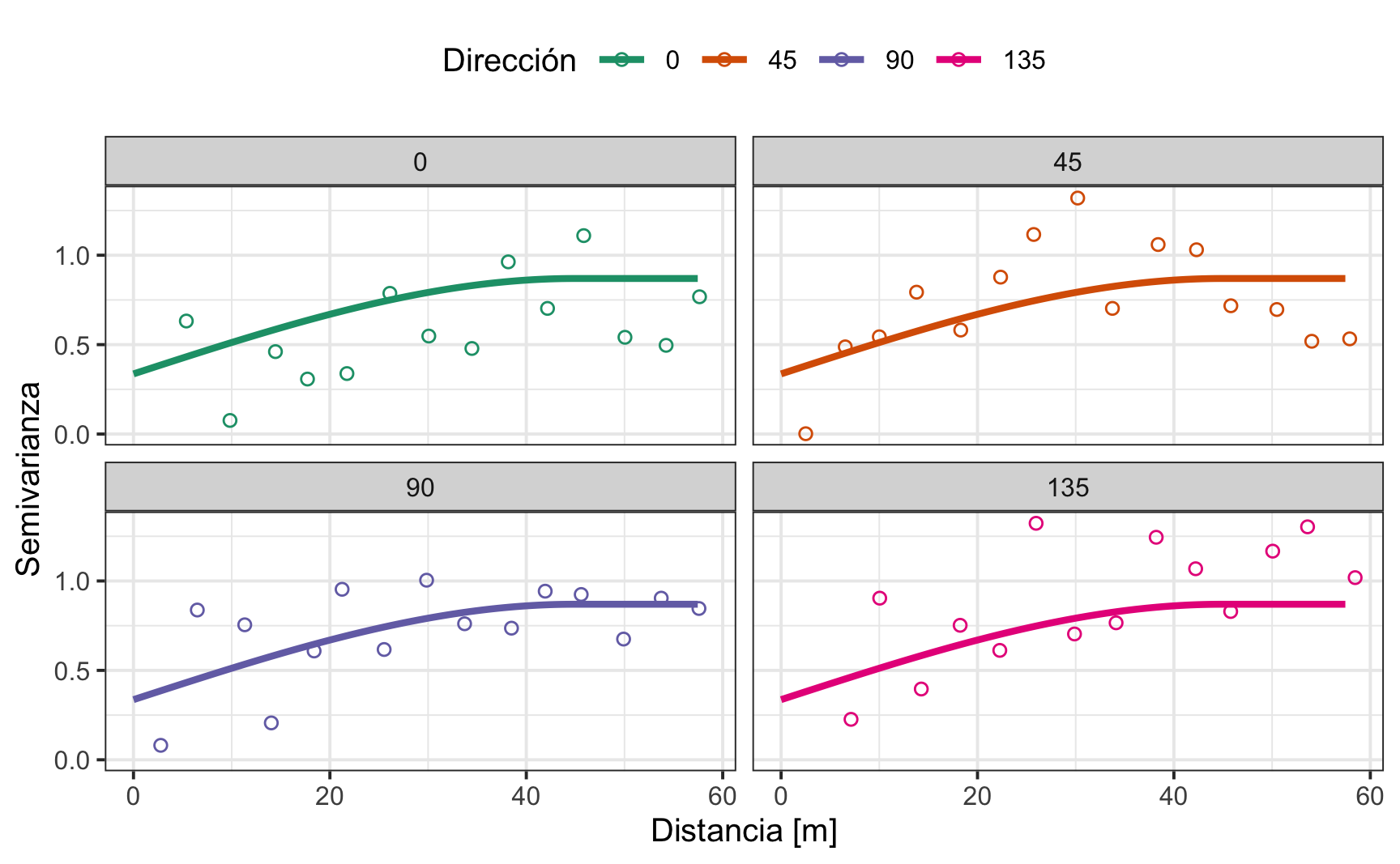


Figura 20: Variogramas direccionales y modelo esférico ajustado, mostrando que el modelo aplica para todas las direcciones.

### 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 el error cuadrático medio (), la razón de desviación cuadrática media (), y la correlación () entre los valores observados y predecidos, donde se busca es determinar qué tan similares son los valores entre si (Figura 21).

Para realizar la validación cruzada se usa la función krige.cv, con los argumentos de la fórmula, datos, y el modelo ajustado.

kcv.ok = krige.cv(myformula,   
 locations = datos\_sf, model = varmod)

El siguiente bloque de código muestra cómo se calculan las diferentes métricas, ya sea usando funciones ya disponibles o escribiendo la fórmula necesaria.

cl = .95 # nivel de confianzaa  
decimales = 3 # decimales a usar  
  
xval.rmse = sqrt(mean(kcv.ok$residual^2)) # RMSE - menor es mejor  
  
xval.msdr = mean(kcv.ok$residual^2/kcv.ok$var1.var) # MSDR - ~1 es mejor  
  
xval.mod = lm(observed ~ var1.pred, as.data.frame(kcv.ok))  
  
xval.r2 = xval.mod %>%   
 broom::glance() %>%   
 pull(r.squared) # r2 - mayor es mejor  
  
xval.g = 1 - (sum((kcv.ok$var1.pred-kcv.ok$observed)^2)/sum((kcv.ok$observed-mean(kcv.ok$observed))^2)) # G - mayor y positivo es mejor  
  
xval.mape = MAPE(xval.mod) # MAPE - menor es mejor  
  
correl = signif(CorCI(cor(kcv.ok$observed,   
 kcv.ok$var1.pred),   
 nrow(kcv.ok)),  
 decimales)  
  
metricas = tibble(metric = c('$RMSE$','$MSDR$','$r$','$R^2$','$MAPE$','$G$'),   
 estimate = c(apa(xval.rmse,decimales),  
 apa(xval.msdr,decimales),  
 apa(correl[1],decimales,F),  
 apa(xval.r2,decimales,F),  
 apa(xval.mape,decimales,F),  
 apa(xval.g,decimales,F)))

Tabla 3: Métricas de ajuste para la validación cruzada

|  |  |
| --- | --- |
| Métrica | Valor |
|  | 0.747 |
|  | 0.929 |
|  | .492 |
|  | .242 |
|  | .019 |
|  | .238 |

Como se mencionó arriba las métricas son más útiles cuando se comparan modelos, pero para este caso se puede decir que presentan valores aceptables (Tabla 3), la está muy cerca de 1, la correlación () es moderada-alta (.492), el es menor a la desviación estándar de los datos (0.863), el es bajo cercano a 0, y el estadístico es positivo.

gg.xval1 = ggplot(as.data.frame(kcv.ok), aes(var1.pred, observed)) +   
 geom\_point(col = "blue", shape = 1, size = 1.25) +   
 coord\_fixed() +   
 geom\_abline(slope = 1, col = "red", size = 1) +   
 geom\_smooth(se = F, method = 'lm', col = 'green', size = 1.25) +  
 labs(x = "Predecidos", y = "Observados")  
gg.xval1

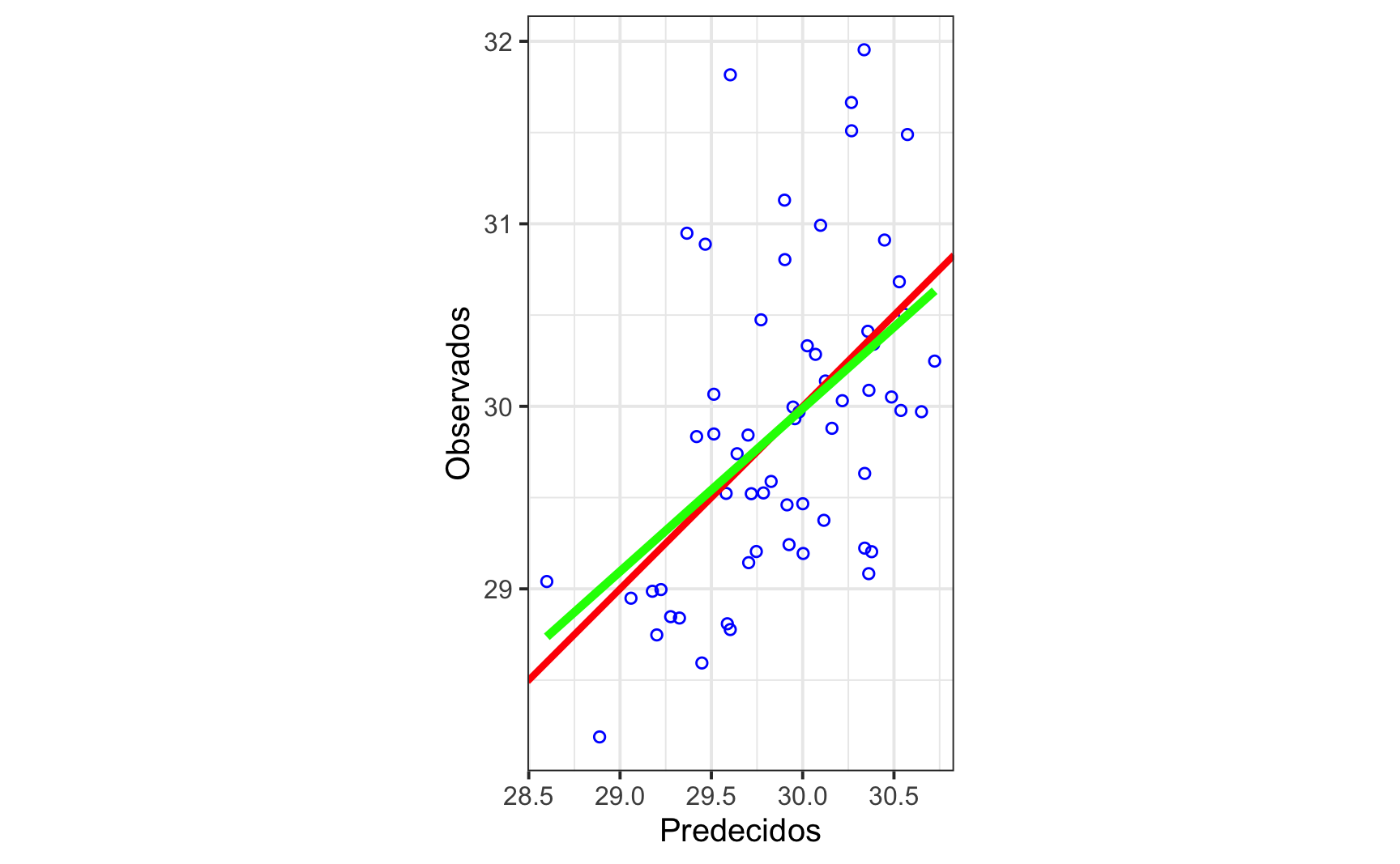


Figura 21: Relación entre los valores observados y predecidos por la validación cruzada. La línea roja es la línea 1:1 y la línea verde es la regresión entre los datos.

Adicionalmente se pueden explorar los residuales ya que idealmente se esperaría que presenten una distribución normal. Lo anterior se puede apreciar en la Figura 22, donde le histograma aunque no es perfectamente normal, no presenta valores extremos y se encuentra moderadamente centrado alrededor de 0.

gg.xval2 = ggplot(as.data.frame(kcv.ok), aes(residual)) +   
 geom\_histogram(bins = 15, col = 'black', fill = "blue") +   
 labs(x = "Residuales", y = "Frecuencia") +   
 geom\_vline(xintercept = mean(kcv.ok$residual), col = 'red')  
gg.xval2

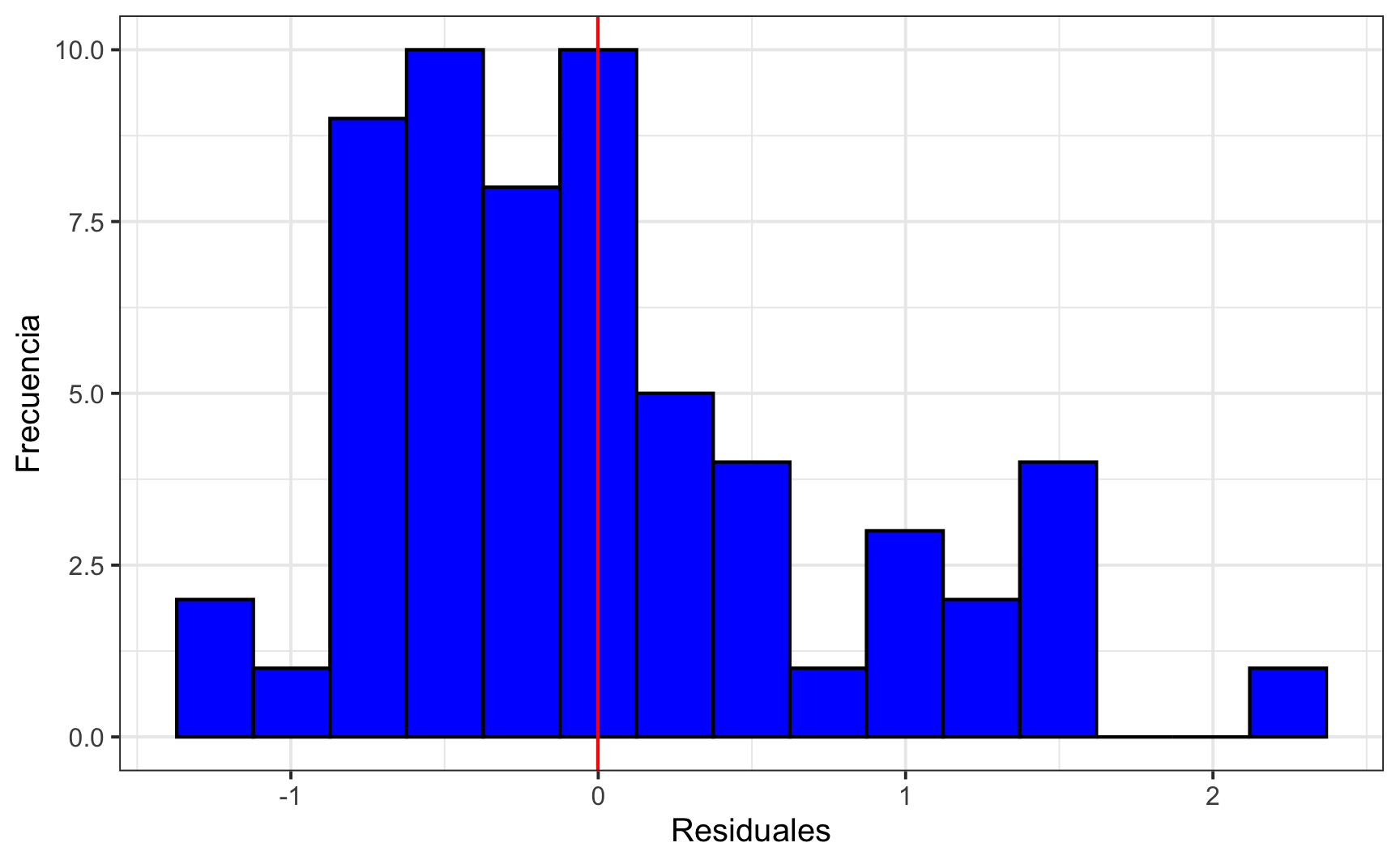


Figura 22: Histograma de los residuales, donde lo que se busca es que esten centrados alrededor de 0 y sigan una distribución aproximadamente normal.

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

## Interpolación (Kriging)

Para recalcar nuevamente, el análisis geoestadístico es un proceso que conlleva el calculo del variograma, el ajuste de un modelo, la validación del modelo, 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.

La interpolación por Kriging es se realiza por medio de la función krige, la cual tiene como argumentos la fórmula, los datos, la grilla a interpolar, y el modelo ajustado seleccionado. El resultado en objeto stars, igual al formato de la grilla a interpolar, con dos atributos o columnas: los valores predecidos (estimados) en var1.pred y la varianza (error) de la predicción (estimación) en var1.var.

ok = krige(as.formula(paste(myvar,'~1')),   
 locations = datos\_sp,  
 newdata = datosint2,   
 model = varmod)

## [using ordinary kriging]

Los mapas finales tanto de la predicción (estimación) como de la varianza (error de estimación) se presentan en las Figuras 23 y 24, respectivamente. El el argumento más importante para la visualización de los resultados es el aes(fill) donde se define el atributo a visualizar (predicción o varianza).

gg.pred = ggplot() +   
 geom\_stars(data = ok, aes(fill = var1.pred, x = x, y = y)) +   
 scale\_fill\_gradientn(colours = viridis(10), na.value = NA) +   
 coord\_sf() +   
 labs(x = x\_map, y = y\_map,   
 title = 'Predicción',   
 fill = myvar)  
gg.pred

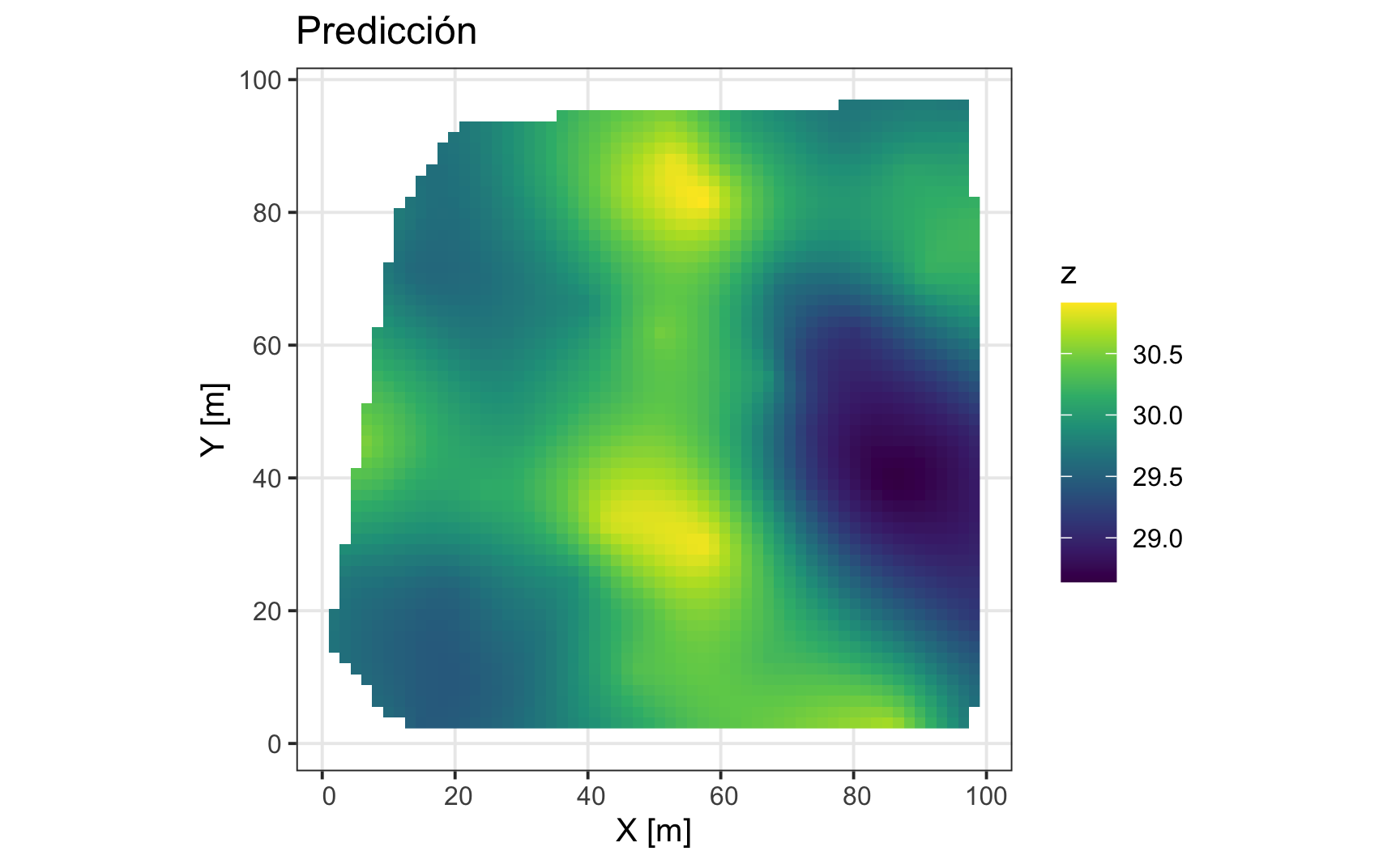


Figura 23: Mapa de predicción de la variable de interés.

gg.var = ggplot() +   
 geom\_stars(data = ok, aes(fill = var1.var, x = x, y = y)) +   
 scale\_fill\_gradientn(colours = brewer.pal(9,'RdPu'), na.value = NA) +   
 coord\_sf() +   
 labs(x = x\_map, y = y\_map,   
 title = 'Varianza',  
 fill = myvar)  
gg.var

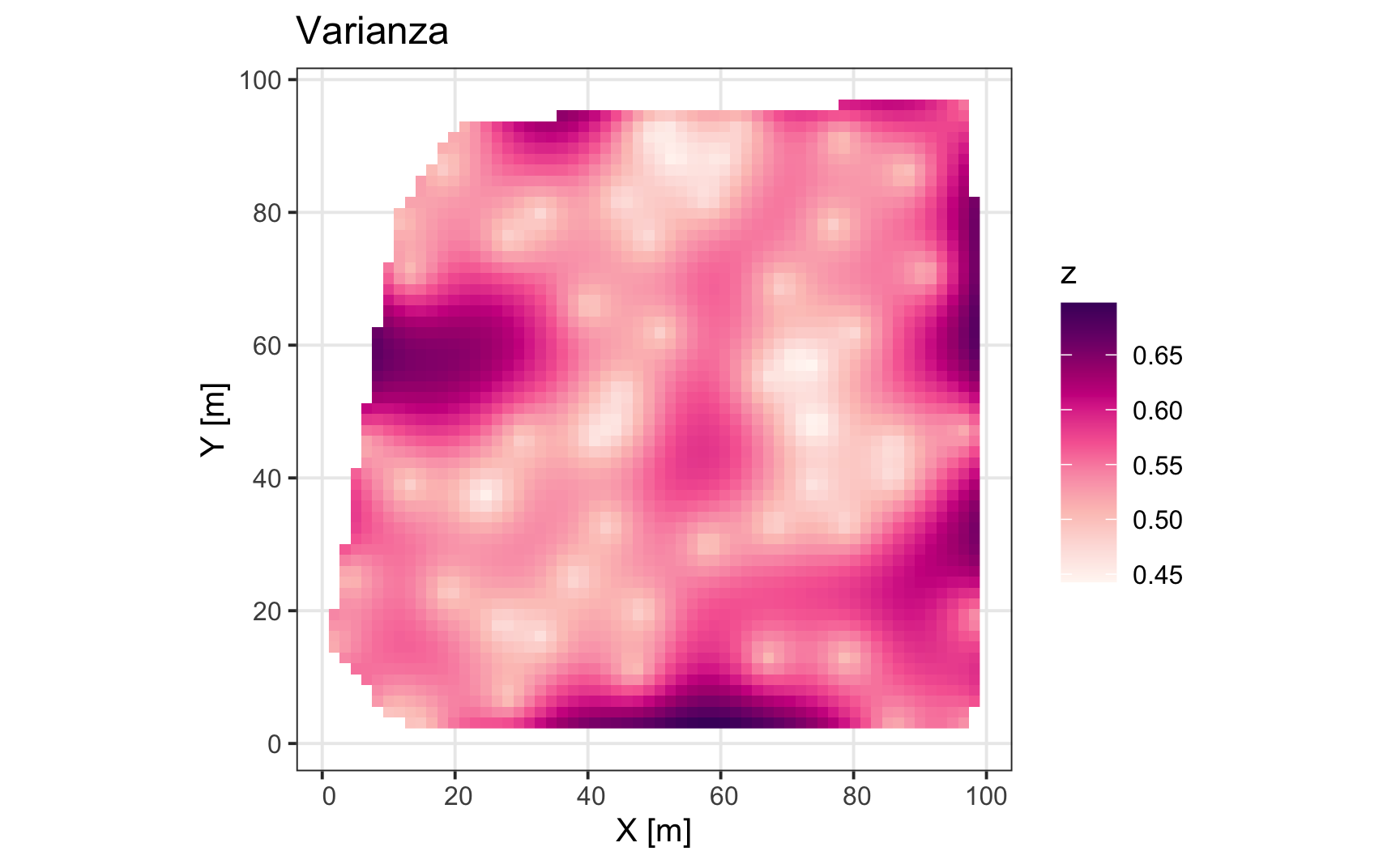


Figura 24: Mapa de varianza (error) de la variable de interés.

# Aplicación web

Lo demostrado acá se encuentra implementado en una aplicación web (Garnier-Villarreal, [2019](#ref-garnier-villarreal2019c)), la cual puede ser usada accediendo a esta 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 25 muestra la interfaz de la aplicación, donde el rectángulo amarillo resalta las viñetas (pasos) a seguir durante el análisis, a como se presentaron en este trabajo.

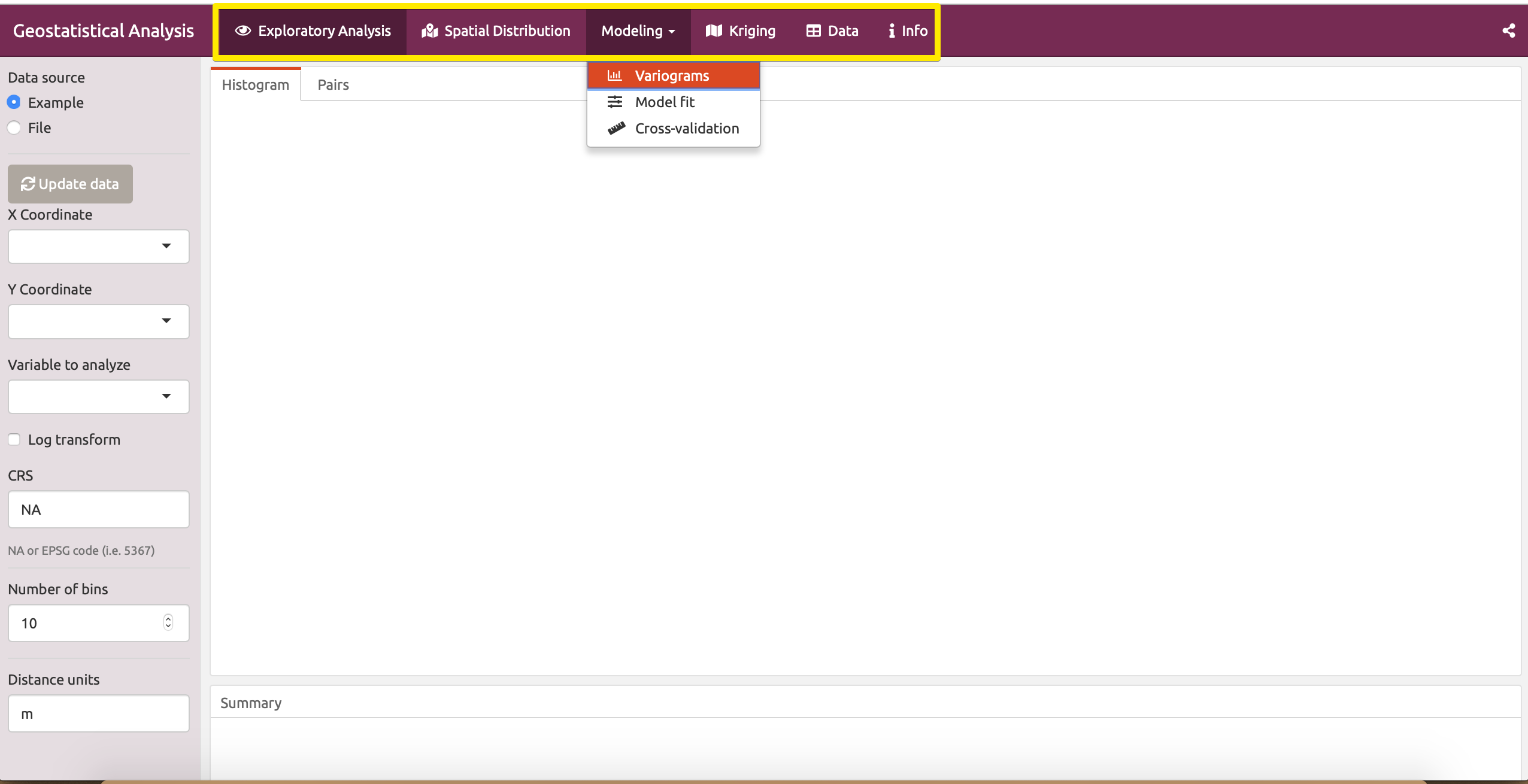


Figura 25: Interfaz de la aplicación web para realizar análisis geoestadístico.

# 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 un 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 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.

El código presentado aquí puede ser utilizado libremente, para mayor facilidad existe un repositorio en GitHub (<https://github.com/maxgav13/intro_geostats>) que contiene todo lo relacionado con este trabajo y que puede ser clonado o descargado para su uso. 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 siente cómodos con lenguajes de programación.

# Referencias

Appelhans, T., Detsch, F., Reudenbach, C., & Woellauer, S. (2020). *mapview: Interactive Viewing of Spatial Data in R*. <https://github.com/r-spatial/mapview>

Borradaile, G. J. (2003). *Statistics of Earth Science Data: Their Distribution in Time, Space and Orientation*. Springer-Verlag Berlin Heidelberg.

Chilès, J.-P., & Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty* (2.ª ed.). John Wiley & Sons.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley & Sons.

Davis, J. C. (2002). *Statistics and Data Analysis in Geology* (3.ª ed.). John Wiley & Sons.

Eldeiry, A. A., & 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. <https://doi.org/10.1061/(ASCE)IR.1943-4774.0000208>

Finley, A. O., Banerjee, S., & E.Gelfand, A. (2015). spBayes for Large Univariate and Multivariate Point-Referenced Spatio-Temporal Data Models. *Journal of Statistical Software*, *63*(13). <https://doi.org/10.18637/jss.v063.i13>

Garnier-Villarreal, M. (2019). *Geostatistical Analysis* (Versión 1) [Computer software]. <https://maximiliano-01.shinyapps.io/geostatistics/>

Garnier-Villarreal, M. (2020). *Geología Numérica: Ciencia de Datos Para Geociencas*. <https://geologia-numerica.netlify.app>

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. Oxford University Press.

Hijmans, R. J. (2020). *raster: Geographic Data Analysis and Modeling*. <https://rspatial.org/raster>

Isaaks, E. H., & Srivastava, R. M. (1989). *Applied Geostatistics*. Oxford University Press.

Jing, L., & Oliveira, V. D. (2015). geoCount: An R Package for the Analysis of Geostatistical Count Data. *Journal of Statistical Software*, *63*(11), 1-33. <https://doi.org/10.18637/jss.v063.i11>

Kravchenko, A., & Bullock, D. G. (1999). A Comparative Study of Interpolation Methods for Mapping Soil Properties. *Agronomy Journal*, *91*(3), 393-400. <https://doi.org/10.2134/agronj1999.00021962009100030007x>

Lark, R. M., Cullis, B. R., & 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. <https://doi.org/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. <https://doi.org/10.15517/rgac.v0i38.4220>

McKillup, S., & Darby Dyar, M. (2010). *Geostatistics Explained: An Introductory Guide for Earth Scientists*. Cambridge University Press. <www.cambridge.org/9780521763226>

Meng, Q., Liu, Z., & 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. <https://doi.org/10.1080/15230406.2013.762138>

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes*. <https://CRAN.R-project.org/package=RColorBrewer>

Nowosad, J. (2019). *Geostatystyka w R*. <https://bookdown.org/nowosad/Geostatystyka/>

Oliver, M., & Webster, R. (2014). A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *CATENA*, *113*, 56-69. <https://doi.org/10.1016/j.catena.2013.09.006>

Pebesma, E. (2020a). *sf: Simple Features for R*. <https://CRAN.R-project.org/package=sf>

Pebesma, E. (2020b). *stars: Spatiotemporal Arrays, Raster and Vector Data Cubes*. <https://CRAN.R-project.org/package=stars>

Pebesma, E., & Bivand, R. (2020). *Spatial Data Science*. <https://keen-swartz-3146c4.netlify.app>

Pebesma, E., & Graeler, B. (2020). *gstat: Spatial and Spatio-Temporal Geostatistical Modelling, Prediction and Simulation*. <https://github.com/r-spatial/gstat/>

Pyrcz, M. J., & Deutsch, C. V. (2014). *Geostatistical Reservoir Modeling* (2.ª ed.). Oxford University Press.

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>

Ribeiro, P. J., Christensen, O. F., & Diggle, P. J. (2003). geoR and geoRglm: Software for Model-Based Geostatistics. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 16.

Robinson, D., & Hayes, A. (2020). *broom: Convert Statistical Analysis Objects into Tidy Tibbles*. <http://github.com/tidyverse/broom>

Sarma, D. D. (2009). *Geostatistics with Application in Earth Sciences* (2.ª ed.). Springer.

Swan, A., & Sandilands, M. (1995). *Introduction to Geological Data Analysis*. Blackwell Science.

Trauth, M. (2015). *MATLAB® Recipes for Earth Sciences* (4.ª ed.). Springer-Verlag Berlin Heidelberg.

Varouchakis, E., Hristopulos, D., & Karatzas, G. (2012). Improving kriging of groundwater level data using nonlinear normalizing transformations—a field application. *Hydrological Sciences Journal*, *57*(7), 1404-1419. <https://doi.org/10.1080/02626667.2012.717174>

Wackernagel, H. (2003). *Multivariate Geostatistics* (3.ª ed.). Springer-Verlag Berlin Heidelberg.

Wang, G., Gertner, G., Fang, S., & 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., Wu, Z., & 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. <https://doi.org/10.3390/rs9121278>

Webster, R., & Oliver, M. A. (2007). *Geostatistics for Environmental Scientists* (2.ª ed.). John Wiley & Sons.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., & Dunnington, D. (2020). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. <https://CRAN.R-project.org/package=ggplot2>

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *dplyr: A Grammar of Data Manipulation*. <https://CRAN.R-project.org/package=dplyr>

Wickham, H., & Henry, L. (2020). *tidyr: Tidy Messy Data*. <https://CRAN.R-project.org/package=tidyr>

Yamamoto, J. K. (2007). On unbiased backtransform of lognormal kriging estimates. *Computational Geosciences*, *11*(3), 219-234. <https://doi.org/10.1007/s10596-007-9046-x>

Yao, X., Fu, B., Lü, Y., Sun, F., Wang, S., & Liu, M. (2013). Comparison of Four Spatial Interpolation Methods for Estimating Soil Moisture in a Complex Terrain Catchment. *PLoS ONE*, *8*(1), e54660. <https://doi.org/10.1371/journal.pone.0054660>