USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Modelos geoestadísticos para la determinación del espesor del hielo en Groenlandia

Víctor Sloth Sixto Poulsen

Curso 2024-25

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Modelos geoestadísticos para la determinación del espesor del hielo en Groenlandia

Víctor Sloth Sixto Poulsen

Enero, 2025

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Coñecemento: Estadística e Investigación Operativa
Título: Modelos geoestadísticos para la determinación del espesor del hielo en Groenlandia
Breve descrición do contido
A lo largo de esta disertación, presentamos y exploramos el uso de modelos Kriging como solución para el problema geoestadístico de estimación del tamaño total de la capa de hielo de Groenlandia, tanto en volumen como en extensión. Además, evaluamos el rendimiento de predicción de estos y otros modelos, comparando su precisión en relación a su respectiva complejidad.
Recomendacións
Outras observacións

Índice

Resumen viii

Introducción xi

1. Recolección de datos 1

1.1. Variables 1

1.1.1. Espesor del hielo (E), m 1

1.1.2. Altitud (A), m 4

1.1.3. Anomalía gravitatoria (AG), mGal 4

1.1.4. Balance de masa (BM), mmWE 5

1.1.5. Distancia a la costa (DC), km 5

1.1.6. Velocidad del hielo (VH), m/a 5

1.2. Procesado 6

1.2.1. Covariables 7

1.2.2. Variable de interés 8

1.2.3. Matriz de observaciones 10

2. Modelos geoestadísticos 11

2.1. Conceptos previos de regresión 11

2.1.1. Validación cruzada 12

2.2. Dependencia espacial 13

2.2.1. Estacionariedad 13
2.2.2. Isotropía 14
2.2.3. El semivariograma 14
2.3. Kriging 16

3. Aplicación al cálculo del espesor del hielo **19**
3.1. Análisis preliminar 19
3.1.1. Estimación del semivariograma 22
3.2. Predicción 25
3.2.1. Otros modelos 27
3.2.2. Comparación 33

4. Conclusiones **35**

I. Figuras **37**

II. Código de R **39**
II.1. Procesado de covariables 39
II.2. Procesado del espesor de hielo 40
II.3. Estimación de semivariogramas 42
II.4. Predicción de modelos 43
II.5. Comparación de métricas 44

Resumen

A lo largo de esta disertación, presentamos y exploramos el uso de modelos Kriging como solución para el problema geoestadístico de estimación del tamaño total de la capa de hielo de Groenlandia, tanto en volumen como en extensión. Además, evaluamos el rendimiento de predicción de estos y otros modelos, comparando su precisión en relación a su respectiva complejidad.

Abstract

Over the course of this dissertation, we present and explore the use of Kriging models in fitting a solution to the geostatistical problem of estimating the total size of the Greenland ice sheet, both in volume and extention. In addition, we evaluate the prediction performance of these and other models, comparing their precision in relation to their respective complexity.

Introducción

A Groenlandia, nación constituyente al reino de Dinamarca, lo cubre la segunda capa de hielo más grande del mundo. Durante décadas, ha sido el objetivo de expediciones científicas, con una proliferación reciente en el uso de radares capaces de penetrar la superficie para obtener información sobre su espesor o composición. Esta facilidad en la obtención de datos ha motivado muchos estudios sobre diversas propiedades del hielo, como la presencia de agua líquida o los movimientos internos de glaciares, que nos detallan los factores más influyentes en su evolución (Karlsson et al. 2024).

En nuestro caso, usaremos algunas de estas mediciones para estimar el volumen y la extensión total de la capa de hielo a través de una discretización de la superficie de Groenlandia, que nos aportará un número finito de N regiones con área a_i en las que calcularemos para cada una el espesor aproximado del hielo, e_i, haciendo uso de modelos geoestadísticos. De esta manera, el volumen total buscado será $\sum_{i=1}^{N} a_i e_i$ y el área $\sum_{i=1}^{N} a_i$. Por tanto, este método necesita un mapa del espesor de la capa de hielo groenlandesa.

Para estimar este espesor del hielo en cada región, usaremos modelos propios de la geoestadística que extienden el poder de predicción de las técnicas de regresión a situaciones con dependencia espacial. Además, compararemos su precisión con otros modelos, como BedMachine (Morlighem et al. 2017) o la misma regresión, para determinar la eficacia de considerar modelos cada vez más complejos.

Quiero agradecer a Vivi K. Pedersen (Aarhus University) y a Nanna B. Karlsson (Geological Survey of Denmark and Greenland) por proporcionarme la idea de este proyecto, así como los datos necesarios para su realización. Además, quiero también agradecer a Manuel F. Bande (Universidad de Santiago de Compostela) por darme las pautas necesarias para enfocar este trabajo.

Some data used in this study were acquired by NASA's Operation IceBridge.

Capítulo 1

Recolección de datos

Tal y como recoge la expresión "Basura entra, basura sale" (*"Garbage in, garbage out"*), la calidad de los datos es esencial para la obtención de resultados fiables a través del análisis estadístico. En este capítulo, elegiremos las variables, procedentes de diversas fuentes, que nos permitirán establecer una estimación del espesor del hielo que recubre Groenlandia, unificaremos sus formatos y las prepararemos para su incorporación en un modelo estadístico.

1.1. Variables

Como buscamos aproximar el espesor del hielo, esta será nuestra variable respuesta, también llamada variable de interés, de la que lógicamente debemos buscar mediciones. Además, para conseguir una mejor predicción en toda la superficie, necesitamos también encontrar covariables con alguna relación de causa o efecto con el espesor del hielo. Hecho esto, damos una breve descripción de cada variable, mencionando sus fuentes y unidades de medida, y las representamos gráficamente.

1.1.1. Espesor del hielo (E), m

Elegimos los datos disponibles públicamente en el CReSIS, centro que ha diseñado, desarrollado y desplegado sondeos de profundidad por radar desde 1987. Además, su colaboración en proyectos como la operación IceBridge (*NASA's Operation IceBridge* 2009-2019) de la NASA han sido clave para la adquisición de gran parte de los datos que usaremos.

Entre los 1.2 petabytes de datos sin procesar que alberga CReSIS en su base de datos, nuestro interés recae en una porción de mediciones de espesor del hielo de alta calidad entre los años

1993 y 2017 (Open Polar Radar 2024). Los datos son muy extensos, tanto que incluirlos todos solo ralentizaría los tiempos de ejecución, sin aportar información adicional significativa. Por lo tanto, emplearemos únicamente los archivos `Browse_[...].csv`, versiones reducidas de las tablas completas de datos donde se toma uno de cada 50 puntos. Veamos en la Figura 1.1 el porcentaje de mediciones que aporta cada expedición con respecto al total.

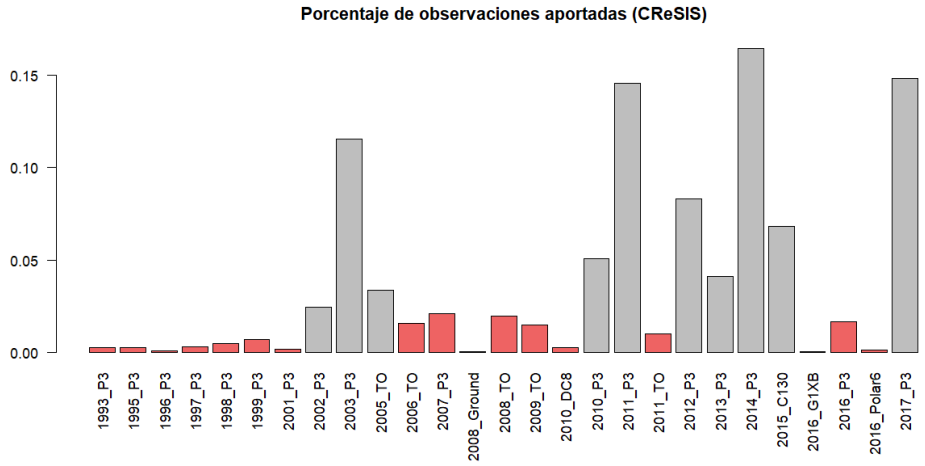


Figura 1.1: Porcentaje de datos aportado por cada expedición a las mediciones del espesor del hielo de CReSIS. Marcado en rojo tenemos a las que no superan el 2,4 %.

Elegimos continuar trabajando solamente con las 10 expediciones más exitosas, representadas en gris, que suponen el 87,5 % del número de observaciones totales. En la Figura 1.2 dibujamos sobre un mapa de Groenlandia el espesor del hielo, medido en metros, de cada una de estas expediciones, acompañándolas de su histograma y diagrama de caja.

Adicionalmente, en la documentación asociada (*CReSIS RDS Radar Guide* 2024) se incluye una sección de análisis de errores, donde proveen la siguiente fórmula para estimar la raíz del error cuadrático medio (RMSE) de las mediciones:

$$RMSE_{CReSIS} = \sqrt{\left(\frac{k_t c}{2B\sqrt{3{,}15}}\right)^2 + \left(\frac{T}{200}\right)^2},$$

donde k_t es el coeficiente de ensanchamiento de la ventana (1,53), calculado numéricamente para compensar los efectos del método usado; c es la velocidad de la luz en el vacío, en metros por segundo; B es el ancho de banda de la onda usada por la sonda; T es el espesor del hielo, en metros, y $\sqrt{3{,}15}$ es el índice de refracción de hielo uniforme utilizado.

Acotando la fórmula por el valor de B más bajo posible para nuestras mediciones, 1 MHz, y aplicándolo a todos los espesores obtenemos un RMSE relativo máximo del 0,779 % con respecto al valor medido. Concluimos, por tanto, que el error es despreciable.

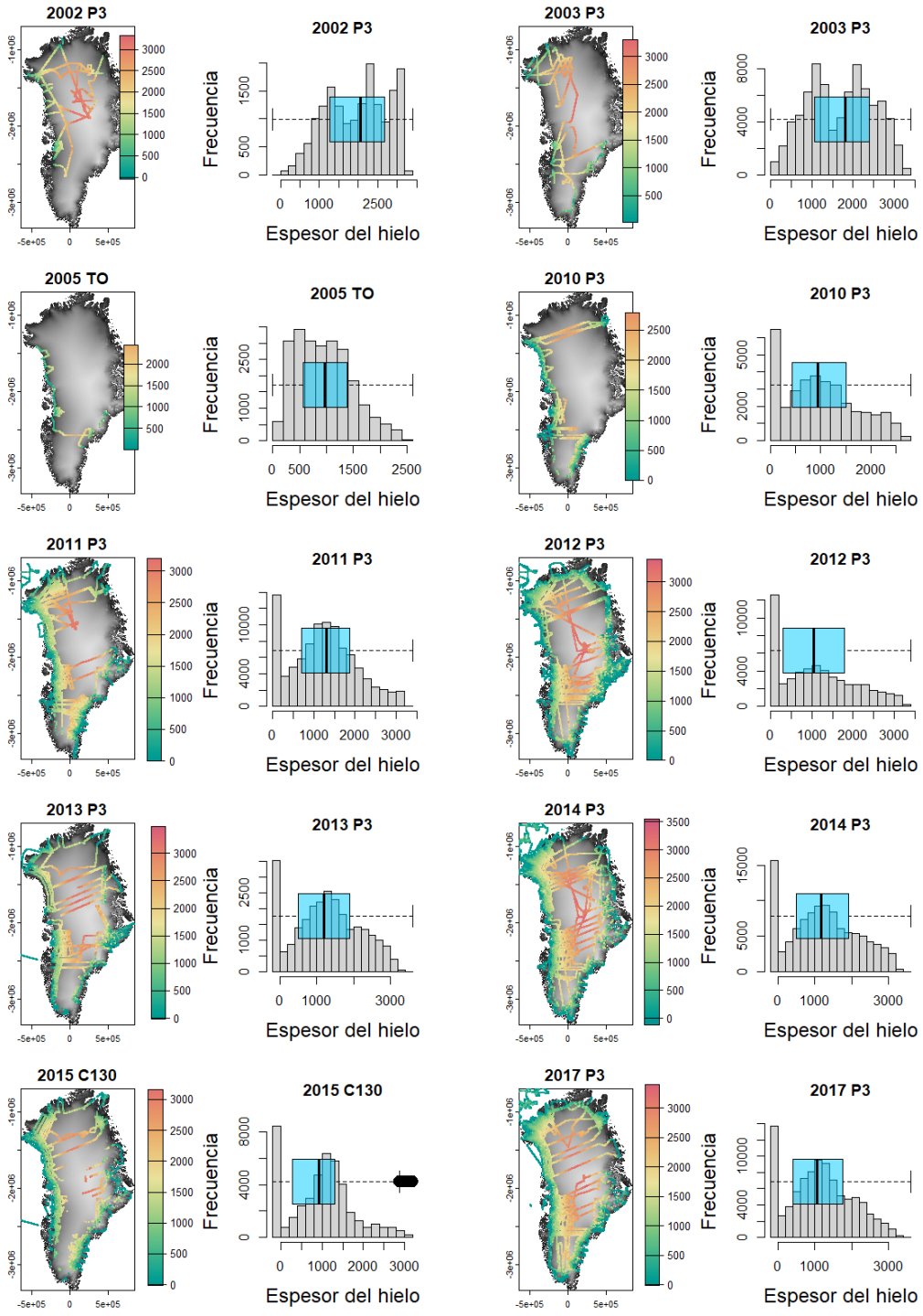


Figura 1.2: Datos de CReSIS del espesor del hielo en Groenlandia, acompañados de sus histogramas y diagramas de caja, por expediciones.

1.1.2. Altitud (A), m

Nos referimos a la altura respecto al elipsoide terrestre de referencia WGS84 (*World Geodetic System* 1984), en metros, de la superficie de Groenlandia. Tomamos los datos del Greenland Mapping Project Digital Elevation Model (Howat et al. 2015) (Howat et al. 2014), que forma parte del programa MEaSUREs de la NASA, pionera en el uso científico de mediciones por satélite. Podemos verlos representados en la Figura 1.3.

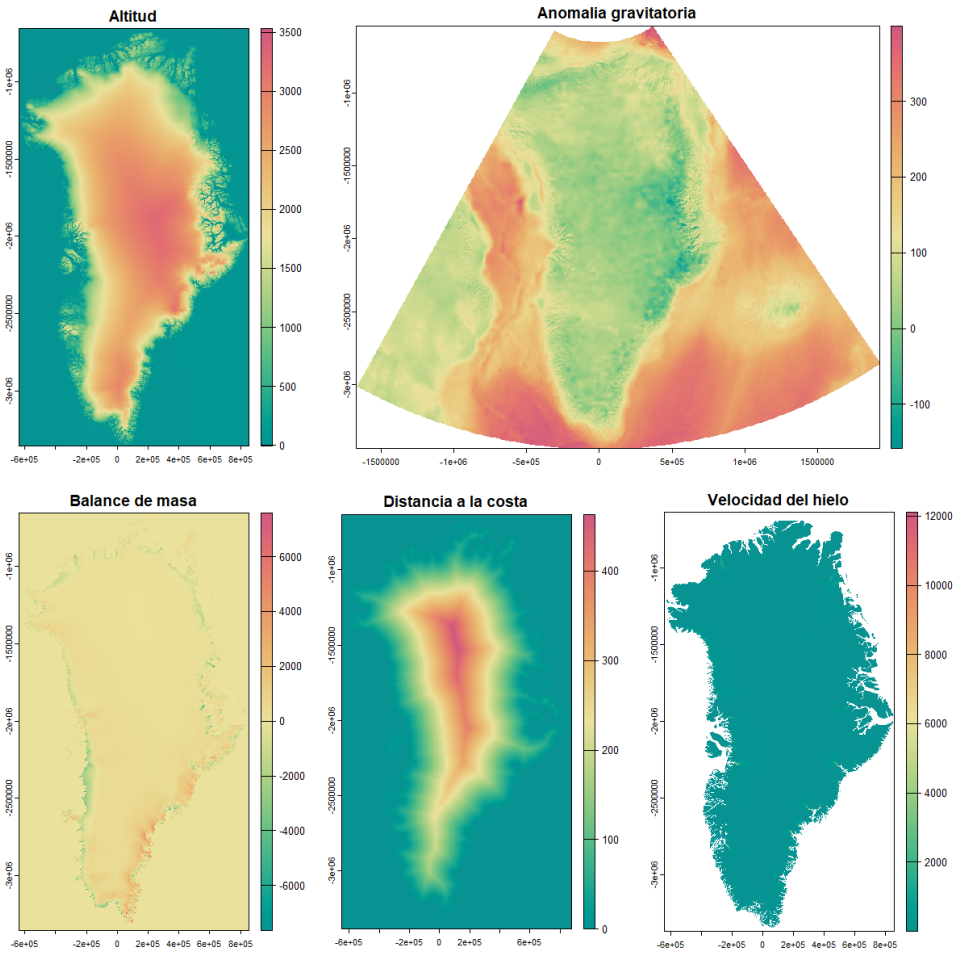


Figura 1.3: Mapas de las covariables, proyectadas al sistema de coordenadas común, en sus propias resoluciones y extensiones.

1.1.3. Anomalía gravitatoria (AG), mGal

El campo gravitatorio de la Tierra no es uniforme, mostrando variaciones causadas principalmente por las alteraciones en densidad de los materiales bajo la superficie. Estas inconsistencias pueden medirse a través de gravímetros, instrumentos capaces incluso de detectar cambios en la

gravedad desde satélites de baja órbita.

La anomalía gravitatoria se refiere a la diferencia entre la fuerza de gravedad teórica calculada a partir del WGS84 y la gravedad real medida. La unidad usada para cuantificar estas perturbaciones gravitatorias son los Gals o Galileos, definidas como 1 $Gal = 1\ cm/s^2$.

El mapa que usamos para nuestra covariable de anomalía gravitatoria es la propuesta por el modelo WGM2012 (Bonvalot et al. 2012) (Balmino et al. 2011), creada por el BGI gracias a las mediciones tomadas por misiones como los satélites GOCE y GRACE. Los datos están en miliGals, y se pueden ver en la Figura 1.3.

1.1.4. Balance de masa (BM), mmWE

La definimos como la diferencia entre la aportación y el retiro de masa a causa de diversos procesos naturales. Consideramos el modelo descrito por Brice Noël (Noël et al. 2018), que con datos del modelo climático especializado de Groenlandia, RACMO2, le resta a la precipitación acumulada los efectos de la erosión, el deshielo, y la sublimación.

Tomamos como nuestros datos a la versión promediada durante décadas, expresados en la Figura 1.3 en milímetros equivalentes de agua (mmWE), una unidad de medida numéricamente idéntica a kilogramo por metro cuadrado usada para representar el grosor de una masa cuya densidad es la misma que el agua, ya que 1 kg de agua tiene una altura de 1 mm al distribuirse uniformemente sobre 1 m^2 de superficie.

1.1.5. Distancia a la costa (DC), km

Para obtener el mapa de distancia a la costa, partimos de la delimitación de Groenlandia usada por BedMachine (Morlighem et al. 2017), uno de los modelos con los que compararemos nuestros resultados al final de este trabajo. Por tanto, tomamos como línea de costa al contorno de la unión formada por las zonas sin hielo y con hielo continental según la Figura I.1 del Anexo I. Calculamos ahora la distancia mínima, en kilómetros, de cada punto del interior a la costa que hemos definido, con lo que llegamos al resultado que puede verse en la Figura 1.3.

1.1.6. Velocidad del hielo (VH), m/a

El hielo que recubre Groenlandia está en constante movimiento debido, en gran parte, a su propio peso. Estos movimientos han sido medidos por el programa MEaSUREs de la NASA (Joughin et al. 2017), que ya mencionamos. Los datos que empleamos, dados en metros por año,

constan del promedio ponderado de la velocidad del hielo medido durante dos décadas (Joughin et al. 2016) y se pueden ver en la Figura 1.3.

Como consecuencia de este gran alcance temporal, los valores promediados de la zona costera, en constante evolución, tendrán la mayor varianza. Según la documentación, el error de medida puede alcanzar un 3 % por simplificaciones como el flujo paralelo del hielo.

1.2. Procesado

Como el formato de los datos varía por provenir de varias fuentes, las debemos unificar en una base de datos con estructura consistente para su uso. Agruparemos los datos por observaciones, donde cada observación se corresponde con un punto en la superficie de Groenlandia. Cada observación constará de tres partes: la variable respuesta que nos interesa predecir, las covariables que nos ayudarán a hacerlo y las coordenadas del punto.

Para minimizar la sobrerepresentación de observaciones con una densidad anómalamente alta en una determinada zona, que provocaría un sesgo indeseado, construiremos una rejilla regular a la que ajustaremos los datos. Esta rejilla estará formada por celdas rectangulares de área a_i que cubrirán la región de estudio de forma matricial. Además, la utilizaremos como máscara en el sentido informático, dejando vacías las celdas sin interés para nuestro estudio. Su construcción depende de tres parámetros:

1. El **sistema de coordenadas**. Debe ser el mismo para todos nuestros datos espaciales, garantizando así la unicidad de los puntos, representados por el par (x, y). En la rejilla, las coordenadas de una celda son equivalentes a las coordenadas de su centro.

   En nuestro caso, como la zona a estudiar se sitúa en una franja muy próxima al polo norte, de latitudes comprendidas entre 59°N y 84°N, el sistema de coordenadas que minimiza la distorsión ocasionada al proyectar esta región de la superficie terrestre sobre un plano es el NSIDC Sea Ice Polar Stereographic North (EPSG:3413), que emplea una proyección estereográfica polar partiendo del modelo global elipsoidal WGS84 (*World Geodetic System* 1984), el modelo geodésico estándar usado, por ejemplo, en los sistemas GPS. Esta será la proyección que tomaremos para nuestra rejilla y a la cual se ajustarán todos los datos.

2. La **resolución**. Entenderemos la resolución de una rejilla como una medida del número de celdas que la forman. Como consecuencia, las celdas serán más pequeñas cuanto mayor sea su número. Como luego ajustaremos los datos a la rejilla, elegir una resolución alta resultaría en más celdas sin observaciones, obligándonos a recurrir a la interpolación para completar la superficie. Por otro lado, elegir una resolución baja nos proporciona celdas de

mayor tamaño en las que será más probable encontrar varias observaciones, en cuyo caso se promedian, obteniendo así una mayor representación global a coste de precisión. Por tanto, alejarnos del intervalo de resoluciones en las que se comprenden los datos originales introducirá incertidumbre innecesaria.

Además, dada la abundancia de datos, reducir el número de observaciones nos beneficia ya que disminuirá notablemente el tiempo de computación. Por tanto, elegiremos como resolución de la rejilla a la menor resolución de todas las covariables, que resulta ser la de Anomalía gravitatoria, con celdas de 1.57 km $\times$ 1.57 km.

3. La **región de estudio**. Debe estar constituida por celdas en las que disponemos de todas las covariables para el correcto funcionamiento del modelo de predicción. Adicionalmente, intentaremos evitar zonas sin interés, cuyas celdas mantendremos vacías.

   Al aplicar la primera restricción intersecando las covariables, obtenemos la región definida por la covariable Velocidad del hielo, por lo que será la región que heredará la rejilla.

Modificaremos ahora los datos para ajustarlos a la rejilla que hemos caracterizado. Los cálculos se realizarán en el programa estadístico `R, versión 4.4.2`, empleando el paquete de análisis espacial de datos `terra, versión 1.7-83`.

1.2.1. Covariables

Comenzamos importando las covariables en `R`, transformando en cada caso su formato a un `raster` a través del paquete `terra`, asegurándonos de usar siempre la misma proyección. Hecho esto, las representamos gráficamente en la Figura 1.3, que ya mostramos. A continuación, remuestramos los `rasters` a la resolución acordada a través de una interpolación bilinear de celdas contiguas y aplicamos la máscara que definimos a través de nuestra región de estudio, dejando así vacías las celdas sobre el océano, irrelevantes para nuestro estudio.

Si representamos sus histogramas, detectamos que la distribución de VelHielo no es ideal. Para remediarlo, consideramos la transformación por logaritmo, cuya distribución mejora notablemente, por lo que de ahora en adelante utilizaremos siempre log(VelHielo) en lugar de VelHielo. Otro aspecto importante es que los intervalos de valores de las variables no difieran en demasiadas órdenes de magnitud, algo que hemos conseguido mitigar representando a Distancia a la costa en kilómetros en lugar de metros. Representamos sus gráficas en la Figura 1.4, donde acortamos los nombres de las covariables.

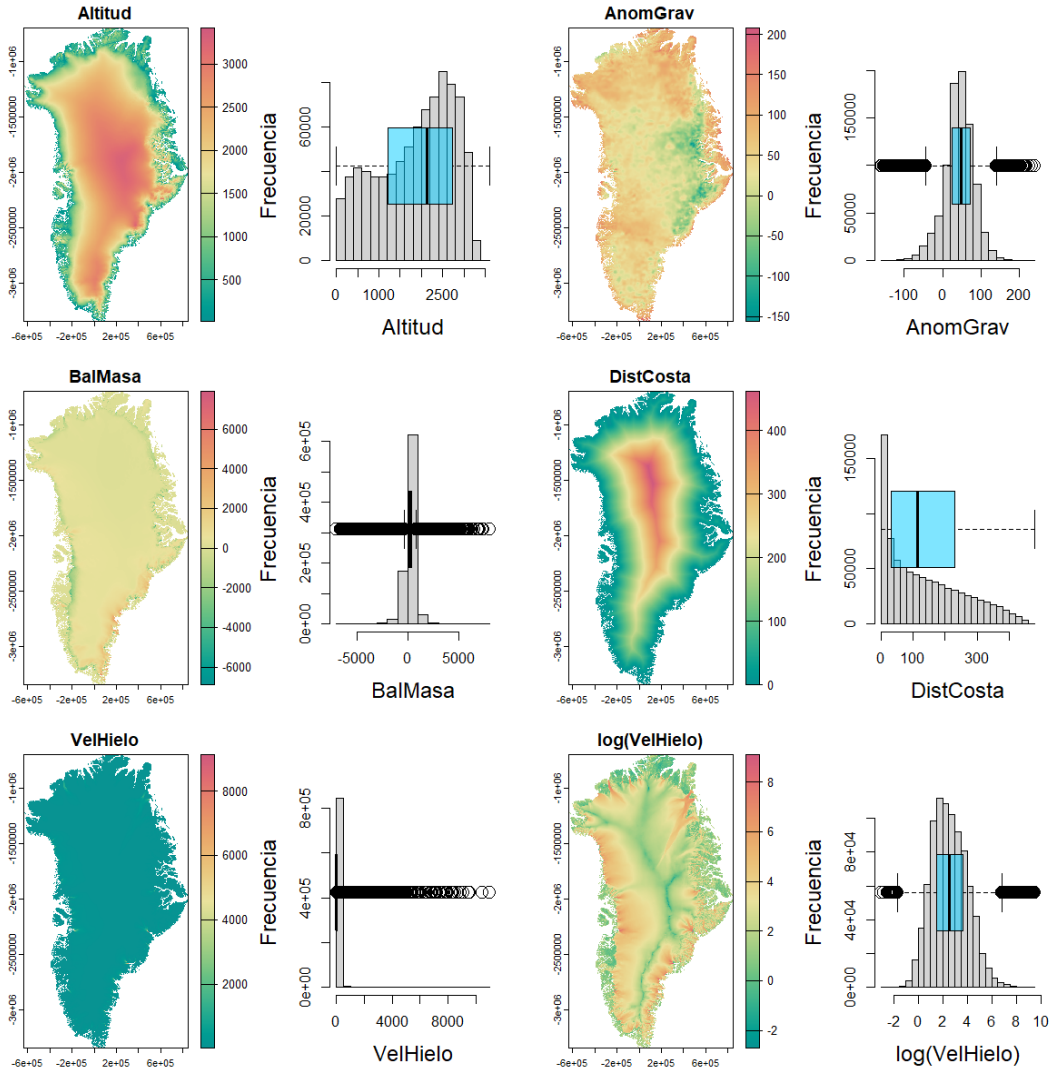


Figura 1.4: Mapas de las covariables en formato unificado, incluyendo la transformación logaritmo de VelHielo, acompañadas de sus histogramas y diagramas de caja.

1.2.2. Variable de interés

Continuaremos importando las mediciones del espesor del hielo de CReSIS, almacenadas por observaciones en ficheros `.csv`. Como ejemplo, mostramos algunas de las mediciones de CReSIS efectuadas por la expedición Greenland Platform 3 de la NASA en el año 2014:

```
      LAT       LON     TIME THICK ELEVATION        FRAME SURFACE  BOTTOM
 81.53761 -30.31336 59698.17  0.00  5100.011 2.014031e+12 4605.78 4605.78
 81.53596 -30.35328 59703.39 44.58  5101.157 2.014031e+12 4614.32 4658.90
 81.53459 -30.39788 59708.61 46.87  5100.818 2.014031e+12 4608.27 4655.15
```

```
81.53343 -30.44277 59713.82 11.25  5099.926 2.014031e+12 4488.24 4499.49
81.53187 -30.48708 59719.03 85.62  5100.118 2.014031e+12 4307.59 4393.21
81.53010 -30.53102 59724.25 20.86  5100.647 2.014031e+12 4283.92 4304.78
```

Las columnas que nos interesan son las dos primeras, ya que expresan la posición de cada observación en el sistema de coordenadas `LatLong` (EPSG:4326), y la cuarta, que recoge las mediciones de espesor del hielo en cada punto. Tras comprobar que la estructura es consistente para cada archivo, extraemos estas columnas, omitiendo observaciones incompletas, para crear un vector espacial de cada expedición con el paquete `terra`. Esto nos permite representarlos individualmente ante un fondo de Groenlandia proporcionado por la covariable Altitud en la Figura 1.2, que ya hemos visto, y de forma conjunta en la primera columna de la Figura 1.5.

Antes ajustar la unión de todas las expediciones a la rejilla, debemos filtrar valores atípicos. La única anomalía que observamos es la abundante inclusión de mediciones hechas en zonas sin presencia de hielo por las expediciones más recientes, dando lugar a espesores nulos o incluso negativos. Para solucionar esto, filtramos los datos, extrayendo los valores menores o iguales a cero. El resultado, expresado de forma conjunta, se comprueba en la segunda columna de la Figura 1.5.

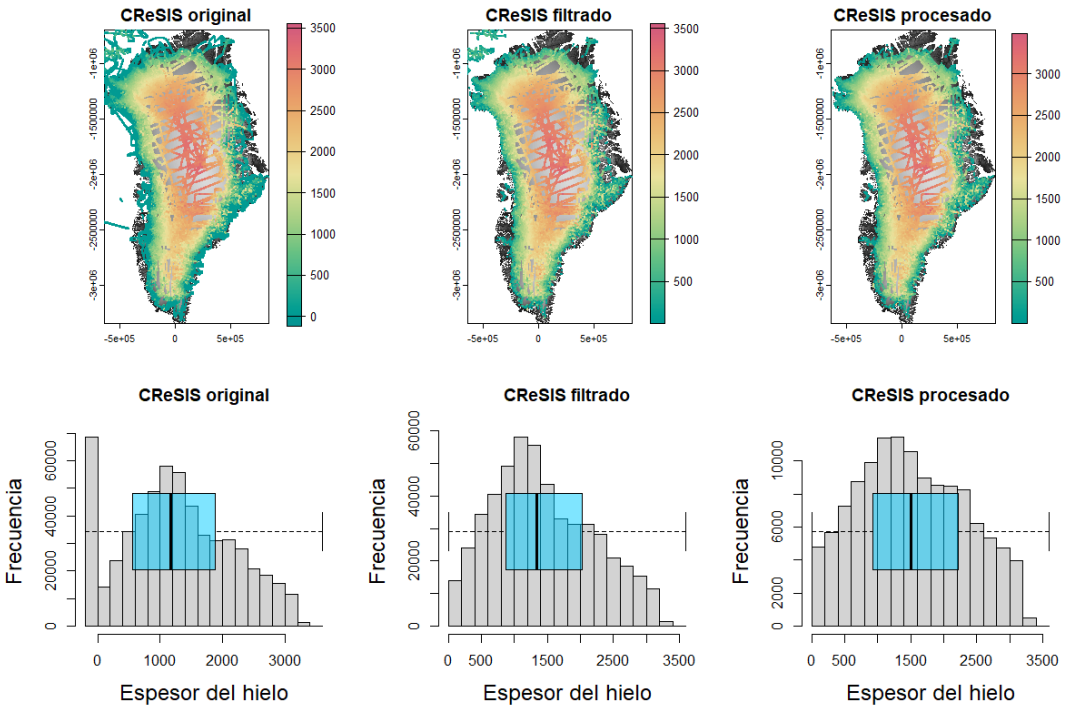


Figura 1.5: Comparación entre los datos originales de CReSIS procedentes de las 10 expediciones más extensas, los resultantes de filtrar las mediciones menores o iguales a cero y los datos finales, ajustados a la rejilla, acompañados de sus histogramas y diagramas de caja.

Por último, ajustamos las mediciones a nuestra rejilla en la tercera columna de la Figura 1.5, promediando las observaciones contenidas en cada celda e ignorando a las que caen fuera de la región de estudio, tal y como comprobamos en la esquina superior izquierda, correspondiente al norte de Canadá. A través de esto, logramos reducir el número de observaciones de 525 332 a 124 711 puntos, que serán nuestros datos de CReSIS procesados que usaremos en el resto del trabajo. Además, incluimos en la Figura I.2 del Anexo I la densidad de las mediciones en términos de observaciones por celda.

1.2.3. Matriz de observaciones

Hasta ahora hemos estado trabajando con los datos como objetos espaciales gracias al paquete `terra`. Esto nos facilitó su manejo y nos ha permitido visualizarlos sobre un mapa de Groenlandia. Sin embargo, ahora que hemos terminado con la organización de los datos por separado estamos listos para convertirlos en una matriz conjunta que será fácilmente leída por los modelos. Para ello, debemos simplemente juntar las tres partes que forman cada observación: la variable de interés, las covariables y las coordenadas del punto. Mostramos a continuación las primeras 6 filas de la matriz de observaciones resultante:

Espesor	Altitud	AnomGrav	BalMasa	DistCosta	VelHielo	x	y
101.4600	34.22641	92.66726	-140.20325	2.021291	128.11993	446708.0	-777528.5
104.7000	36.21686	90.71111	-111.79719	3.237049	93.17938	446708.0	-779101.1
111.1500	35.95683	93.74339	-112.30405	4.169230	92.82845	448280.6	-779101.1
111.9500	41.67732	91.64365	-82.32557	5.413191	47.28630	448280.6	-780673.7
124.5800	48.98146	95.26985	-33.39499	5.765606	49.67364	449853.1	-780673.7
172.6133	62.62717	94.57361	74.68323	6.986169	26.66465	451425.7	-782246.2

Comprobamos que no tiene valores vacíos y que consta de 124 711 filas, coincidiendo con el número de observaciones de la variable de interés, confirmando que el procedimiento se ha realizado correctamente.

Capítulo 2

Modelos geoestadísticos

Con los datos preparados, nos preguntamos ahora qué método estadístico debemos emplear para maximizar la calidad de nuestras predicciones. En este capítulo recordaremos el procedimiento estándar de los modelos de regresión, introduciremos nociones geoestadísticas de dependencia espacial y estos nos llevarán a exponer las ventajas y el funcionamiento de los modelos Kriging.

2.1. Conceptos previos de regresión

Comencemos recordando el modelo clásico para predecir a través de la influencia de covariables: la regresión. En nuestro caso, nos limitaremos a relaciones lineales con la variable respuesta, por la que consideraremos un modelo de regresión lineal múltiple, ya que hemos adquirido varias covariables. Su ecuación, expresada en forma matricial, con n observaciones y p covariables, será la siguiente:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & q_{11} & \cdots & q_{1p} \\ 1 & q_{21} & \cdots & q_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & q_{n1} & \cdots & q_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \mathbf{Q}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

donde $\mathbf{y}$ es el vector de mediciones de la variable respuesta; $\mathbf{Q}$ es la matriz de diseño, con las observaciones por filas y las covariables por columnas, acompañadas de una columna unidad para acomodar al intercepto β_0; $\boldsymbol{\beta}$ es el vector de los parámetros del modelo y $\boldsymbol{\varepsilon}$ es el vector de residuos, que recoge el error entre las predicciones y las observaciones.

Para obtener un vector de predicciones $\hat{\mathbf{y}}$, debemos estimar $\boldsymbol{\beta}$, que haremos mediante su

mejor estimador lineal insesgado (BLUE), el de mínimos cuadrados ordinario (OLS):

$$\hat{\boldsymbol{\beta}}_{OLS} = \left(\mathbf{Q}^{\mathbf{T}}\mathbf{Q}\right)^{-1}\mathbf{Q}^{\mathbf{T}}\mathbf{y}.$$

Sin embargo, este método de estimación supone ciertas varias hipótesis:

- **Linealidad.** La variable respuesta debe poder explicarse con una combinación lineal de las covariables, mientras que estas pueden ser transformadas de antemano.
- **Homocedasticidad.** La varianza de los errores es constante e independiente de la matriz de diseño u otros factores.
- **Independencia.** Las mediciones de la variable de interés o, análogamente, de los residuos, son independientes entre sí.
- **Normalidad.** Los residuos siguen una distribución normal. Sin embargo, no obliga a la variable de interés a seguirla.

2.1.1. Validación cruzada

Para evaluar la precisión de las predicciones de un modelo de aprendizaje estadístico como el de regresión, se suele aplicar la validación cruzada. Este método se compone de varias iteraciones, donde en cada una se dividen las observaciones disponibles a través de una muestra para formar los conjuntos de datos de entrenamiento y de evaluación. A continuación, se usan los datos de entrenamiento para estimar los parámetros del modelo y poder predecir sobre los datos de evaluación, con los que se calcula un error de predicción entre los resultados obtenidos con los valores conocidos de la variable respuesta.

De esta manera, tras repetir muchas veces el muestreo, se pueden promediar los parámetros del modelo calculados en cada iteración de acuerdo con su error de predicción para alcanzar una estimación más imparcial y precisa que la que se obtendría inicialmente. Por tanto, para su aplicación distinguimos tres conjuntos de datos:

1. **Datos de entrenamiento.** Un subconjunto de filas de la matriz de observaciones, en las que disponemos tanto de la variable respuesta como de las covariables, con las que entrenaremos el modelo para establecer relaciones entre ellas.
2. **Datos de evaluación.** Datos con el mismo formato que los datos de entrenamiento, que usaremos para comprobar la fiabilidad del modelo a través del error entre sus predicciones y los valores conocidos de la variable de interés.

3. **Datos de predicción.** A diferencia de las anteriores, no proceden de la matriz de observaciones. Este conjunto consta únicamente de los datos necesarios para la predicción. Por esta razón, no diremos que está formada por observaciones sino por instancias de predicción.

2.2. Dependencia espacial

La hipótesis de independencia de residuos resulta demasiado restrictiva para el análisis de datos espacialmente autocorrelados. La geoestadística relaja esta suposición estructurando una dependencia espacial a través de la estacionariedad y la isotropía.

Debido a las complejas interacciones que ocurren en los procesos naturales, interpretaremos una medición de la variable de interés como la realización espacialmente dependiente de una función de variable aleatoria Z evaluada en las coordenadas $s = (x, y)$ de un punto perteneciente a la región de estudio D. Consideraremos que esta función depende de una componente determinista m y una aleatoria e, pudiendo ser representada de la forma $Z(s) = m(s) + e(s)$.

2.2.1. Estacionariedad

Las estacionariedad se refiere a la idea de que algo no varíe al trasladarse a través del espacio. Nos permite simplificar el problema de autocorrelación espacial al asumir que la media y la variación se comportan de la misma manera en todos los puntos. La primera formulación, llamada estacionariedad estricta o fuerte, supone una misma distribución en todos los puntos. Asumir esto puede ser demasiado restrictivo, en cuyo caso debe ser relajada a **estacionariedad débil**, que supone una media constante en toda la región D y que la covarianza C entre dos puntos sólo dependa del vector h que los separe:

$$\begin{aligned} E\left[Z(s)\right] &= m, \\ C(h) = E\left[\left(Z(s) - m\right)\left(Z(s+h) - m\right)\right] &= E\left[e(s)e(s+h)\right], \end{aligned} \qquad \forall s \in D.$$

Sin embargo, en los casos donde no podemos suponer una media constante debemos recurrir a la hipótesis de **estacionariedad intrínseca**, que supone en su lugar que la esperanza de la diferencia entre dos puntos distanciados por h sea cero (Oliver y Webster 2014). Además, reemplaza la covarianza por una medida más general: la semivarianza.

$$\begin{aligned} E\left[Z(s) - Z(s+h)\right] &= 0, \\ \gamma(h) = \frac{1}{2} Var\left[Z(s) - Z(s+h)\right] &= \frac{1}{2} E\left[\left(Z(s) - Z(s+h)\right)^2\right], \end{aligned} \qquad \forall s \in D.$$

A la función $\gamma(h)$ se le llama semivariograma, que suele abreviarse a simplemente variograma, causando una posible confusión con $2\gamma(h)$. Nosotros sólo usaremos los nombres de forma sinónima si es indiferente indicar a cuál nos referimos. Se deduce de su fórmula que $\gamma(0) = 0$, $\gamma(h) \geq 0$ y $\gamma(h) = \gamma(-h)$ para todo vector h. Además, cuando se cumple la estacionariedad débil, se comprueba que $\gamma(h) = C(0) - C(h)$ (Oliver y Webster 2014). De cara a la siguiente subsección, nótese que $\gamma(h)$ y $C(h)$ pueden depender tanto de la magnitud como de la orientación del vector h.

2.2.2. Isotropía

Si el variograma de las realizaciones de Z depende sólo de la magnitud de h y no de su dirección, se dice que presenta **isotropía**. Puede pensarse como una homoscedasticidad espacial. En cualquier otro caso, se dice anisótropa. Por comodidad de modelado y de cálculos, se suele intentar reducir el efecto de la anisotropía a través de una transformación de coordenadas o estableciendo una subdivisión de direcciones (Bárdossy 1997).

2.2.3. El semivariograma

La hipótesis de estacionariedad hace que el semivariograma experimental, también llamado semivariograma muestral o empírico, sea adecuada para el modelado, mientras que suponer isotropía nos permitirá simplificar su implementación al ignorar los cambios en dirección de h. Su fórmula será, para realizaciones conocidas de Z y un número arbitrariamente fijado de intervalos de distancia $\tilde{h}_l = [h_l, h_l + \delta)$ (Bivand 2008):

$$\hat{\gamma}(\tilde{h}_l) = \frac{1}{2N_l} \sum_{i=1}^{N_l} (Z(s_i) - Z(s_i + h))^2, \quad s_i, s_i + h \in D, \quad \forall h \mid \|h\| \in \tilde{h}_l,$$

donde N_l es el número de pares de puntos $s_i, s_i + h \in D$ que distan algún h tal que $\|h\| \in \tilde{h}_l$. En la práctica usaremos una representación equivalente, $\hat{\gamma}(\bar{h}_l) = \hat{\gamma}(\tilde{h}_l)$, donde $\bar{h}_l$ es la media de los $\|h\| \in \tilde{h}_l$ (`gstat` *user manual* 2025).

La naturaleza puntual del semivariograma muestral no se presta para el uso en modelos, por lo que necesitamos una función continua homóloga. Para esto se elige un semivariograma teórico, modificando sus parámetros para ajustar su traza a los puntos definidos por el semivariograma experimental. Mostramos en el Cuadro 2.1 las fórmulas de algunos de los modelos teóricos más simples (Oliver y Webster 2014) (Bárdossy 1997), que además representaremos gráficamente en la Figura 2.1:

Nugget: $\gamma(h) = \begin{cases} 0, & h = 0 \\ C(0), & h > 0 \end{cases}$ **Esférica:** $\gamma(h) = C(0)\left(\frac{3h}{2a} - \frac{h^3}{2a^3}\right), \quad h \geq 0$

Lineal: $\gamma(h) = \lambda h, \quad h \geq 0$ **Estable:** $\gamma(h) = C(0)\left(1 - e^{-\frac{|h|^\alpha}{a}}\right), \quad \begin{matrix} h \geq 0 \\ 0 < \alpha < 2 \end{matrix}$

Cuadro 2.1: Algunos semivariogramas teóricos.

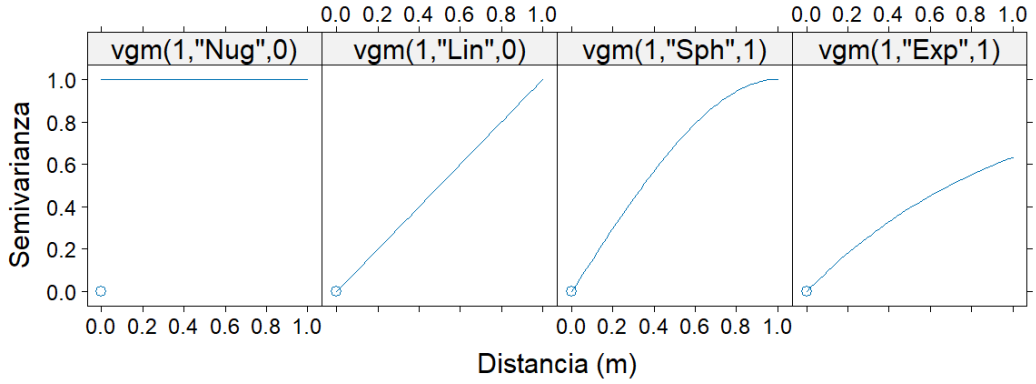


Figura 2.1: Semivariogramas teóricos: Nugget ("Nug"), Lineal ("Lin"), Esférico ("Sph") y Exponencial ("Exp"), equivalente al modelo Estable con $\alpha = 1$.

El alcance a es la distancia $\|h\|$ en la que el variograma alcanza su valor máximo, indicando que puntos más lejanos ya no contribuyen a la dependencia espacial. El umbral $C(0)$ es la cantidad de covarianza en $h = 0$ y, en modelos con umbral bajo la estacionariedad débil, es el valor del semivariograma que se obtiene en el alcance, $\gamma(a)$. Mientras que el modelo exponencial nunca para de crecer, se aproxima asintóticamente a $C(0)$, por lo que se considera un "alcance efectivo" de $r = \frac{a}{3}$, en donde se supera al 95 % del umbral. Por otro lado, el modelo lineal no tiene umbral, por lo que se suele usar en situaciones donde únicamente se cumple la hipótesis de estacionariedad intrínseca (Bárdossy 1997). El nugget se suele añadir a otros modelos teóricos para compensar la incertidumbre provocada por errores de medición o por una tendencia subyacente no espacial, donde se llamará "efecto nugget".

Veamos en la Figura 2.2 un ejemplo donde hemos ajustado datos predeterminados procedentes del paquete *sp* de R a un semivariograma esférico. En el eje x tenemos la distancia $\|h\|$, mientras que en la y tenemos la semivarianza. Para los 15 intervalos $\tilde{h}_l$ considerados, obtenemos valores de $\hat{\gamma}(\bar{h}_l)$, dibujadas en forma de puntos, a las que se ajusta el semivariograma teórico, la función continua del gráfico.

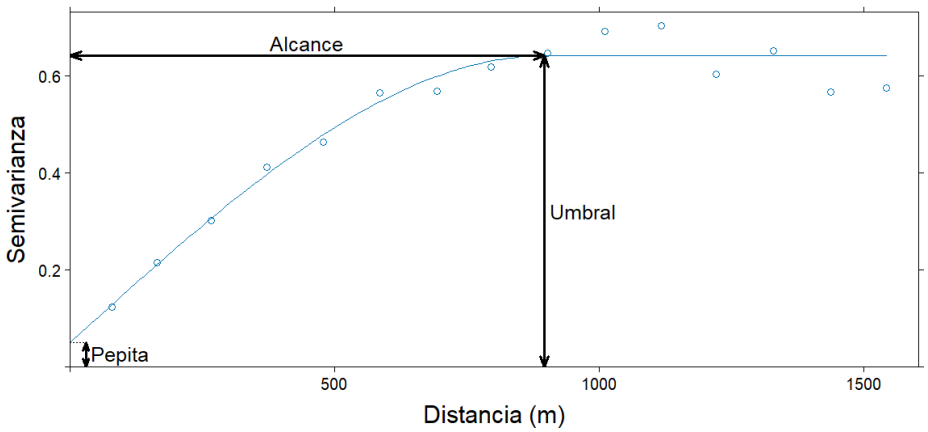


Figura 2.2: Ejemplo de semivariograma experimental ajustada a través del método de mínimos cuadrados a un modelo esférico.

En la práctica puede ser muy difícil determinar el mejor modelo, ya que nuestra elección de semivariograma teórico depende de la forma que presenta el semivariograma experimental, que a su vez varía en función del número de intervalos $\tilde{h}_l$ y la acotación superior de $\|h\|$ que elijamos.

2.3. Kriging

La interpolación Kriging es la familia de modelos estándar de la geoestadística que permite extender el método de regresión a situaciones de dependencia espacial a través del uso de un semivariograma. Para lograrlo, se relaja la hipótesis de independencia de residuos, estableciendo en su lugar la hipótesis de estacionariedad débil, que para datos normalmente distribuidos implica estacionariedad fuerte al depender exclusivamente de la media y la varianza, que bajo la hipótesis son estacionarias. Adicionalmente, se suele buscar la suposición de isotropía para simplificar el semivariograma usado.

Partimos de la ecuación de regresión lineal múltiple 2.1, escribiéndola para una sola observación: $y = \beta_0 + \sum_{k=1}^{p} \beta_k q_k + \varepsilon$, en la que q_k es la covariable k-ésima. Si consideramos el residuo ε como una realización de la función de variable aleatoria $Z(s) = m(s) + e(s)$, formada por una componente determinista y una aleatoria y dependiente de la localización s, obtenemos la siguiente igualdad para las observaciones:

$$y(s_j) = \beta_0 + \sum_{k=1}^{p} \beta_k q_k(s_j) + m(s_j) + e(s_j),$$

donde se considera s_j las coordenadas en las que se sitúa cada observación. Como se cumplen

las hipótesis de regresión, $m(s_j)$ puede considerarse nulo, por lo que los residuos serán nuestra componente aleatoria, $e(s_j) = \varepsilon(s_j)$, mientras que $\beta_0 + \sum_{k=1}^{p} \beta_k q_k(s)$ se tomará como la tendencia en cualquier punto de coordenadas s.

De cara a la predicción en una instancia de coordenadas s_0, tenemos un problema: solo conocemos los residuos en los puntos donde disponemos de observaciones. La gran ventaja de los modelos Kriging es que nos permiten definir $e(s_0)$ a partir de una combinación lineal de los n residuos del modelo de regresión, que denotaremos $\varepsilon_j = \varepsilon(s_j)$:

$$e(s_0) = \sum_{j=1}^{n} \omega_j(s_0)\varepsilon_j.$$

Estos $\omega_j(s_0)$, llamados pesos Kriging, son calculados para cada instancia de predicción de tal manera que se minimice la varianza del error de predicción $Var(Z(s_0) - \hat{Z}(s_0))$ (Hengl et al. 2007), restringido a que el estimador sea insesgado, $E(Z(s_0) - \hat{Z}(s_0)) = 0$, para lo que se usa el método de multiplicadores de Lagrange (Oliver y Webster 2014). De esta manera, obtenemos la siguiente estimación para los pesos:

$$\hat{\boldsymbol{\omega}}(\boldsymbol{s_0}) = \begin{bmatrix} \hat{\omega}_1(s_0) \\ \vdots \\ \hat{\omega}_n(s_0) \end{bmatrix} = \begin{bmatrix} \hat{C}(s_1, s_1) & \cdots & \hat{C}(s_1, s_n) \\ \vdots & \ddots & \vdots \\ \hat{C}(s_n, s_1) & \cdots & \hat{C}(s_n, s_n) \end{bmatrix}^{-1} \begin{bmatrix} \hat{C}(s_0, s_1) \\ \vdots \\ \hat{C}(s_0, s_n) \end{bmatrix} = \hat{\mathbf{C}}^{-1}\hat{\mathbf{c}}_{\mathbf{0}},$$

donde $C(s_i, s_j) = C(Z(s_i), Z(s_i + h)) = C(0) - \gamma(h)$ para un semivariograma ajustado, como estamos bajo estacionariedad débil. Este procedimiento requiere disponer un semivariograma ajustado e invertir una matriz $n \times n$, una operación de complejidad temporal $O(n^2)$, por lo que puede suponer un problema a la hora del cálculo. A través de estos pesos llegamos al mejor predictor lineal insesgado (BLUP), mostrada a continuación en forma matricial.

$$\hat{y}(s_0) = \mathbf{q}(s_0)\hat{\boldsymbol{\beta}}_{GLS} + \hat{\mathbf{c}}_{\mathbf{0}}{}^{\mathbf{T}}\hat{\boldsymbol{C}}^{-1}\left(y(\mathbf{s_J}) - \boldsymbol{Q}\hat{\boldsymbol{\beta}}_{GLS}\right), \tag{2.2}$$

siendo $\mathbf{q}(s_0)$ el vector de covariables asociado a la instancia de predicción en las coordenadas s_0, $y(\mathbf{s_J})$ el vector de mediciones de la variable respuesta en las n observaciones con las que entrenamos el modelo de regresión y $\hat{\boldsymbol{\beta}}_{GLS}$ la estimación por mínimos cuadrados generalizado (GLS) de $\boldsymbol{\beta}$, que no supone la hipótesis de independencia (Bivand 2008):

$$\hat{\boldsymbol{\beta}}_{GLS} = \left(\mathbf{Q}^{\mathbf{T}}\hat{\mathbf{C}}^{-1}\mathbf{Q}\right)^{-1}\mathbf{Q}^{\mathbf{T}}\hat{\mathbf{C}}^{-1}y(\mathbf{s_J}).$$

Como podemos comprobar, el semivariograma es esencial para el cálculo de los parámetros

de los modelos Kriging, a través de la estimación de $\mathbf{C}$. Además, en las observaciones, $\hat{\boldsymbol{c}}_{\boldsymbol{j}}{}^{\mathbf{T}}\hat{\boldsymbol{C}}^{-1} = \begin{bmatrix}\delta_{1j} & \cdots & \delta_{nj}\end{bmatrix}$, siendo δ_{ij} el delta de Kronecker, lo que garantiza que $\hat{y}(s_j) = y(s_j)$. Por tanto, Kriging es un llamado interpolador exacto.

Existen dos métodos más restrictivos que el que hemos formulado que prescinden de añadir una tendencia a través de la regresión, trabajando en su lugar directamente sobre la variable respuesta, suponiéndola como realizaciones de la función de variable aleatoria (Bivand 2008). Estos son el Kriging Simple (KS), que en lugar de la tendencia de regresión supone una componente determinista constante $m(s) = m$, conocida en toda la región de estudio, y el Kriging Ordinario (OK), que permite que $m(s) = m$ sea desconocida, habilitando así su estimación a partir de los datos.

La formulación que hemos visto se corresponde propiamente con el popularmente denominado Kriging Universal (UK), aunque un nombramiento más estricto lo consideraría un Kriging Ordinario con deriva externa (Hengl et al. 2007), dejando el término Kriging Universal para referirse a un modelo que incorpora una función de las coordenadas en la estimación de la tendencia (Hengl et al. 2003), como homenaje a *Le krigeage universel* (Matheron 1969). Asimismo, otros autores lo llaman Regresión-Kriging, aunque este suele implicar algún proceso iterativo en el cálculo de $\hat{\boldsymbol{\beta}}$.

Además, todos los métodos Kriging son aplicables en un entorno local, reduciendo el tiempo de cálculo en el caso de gran cantidad de observaciones (Bivand 2008) con una pérdida mínima de precisión o como manera de limitar el alcance del semivariograma teórico. Por último, la varianza del error de predicción del BLUP de Kriging Universal, la ecuación 2.2, es la siguiente (Bivand 2008) (Hengl et al. 2003):

$$\hat{\sigma}^2(s_0) = C(0) - \mathbf{c}_0^{\mathbf{T}}\hat{\mathbf{C}}^{-1}\mathbf{c}_0 + \left(\mathbf{q}(s_0) - \mathbf{c}_0^{\mathbf{T}}\hat{\mathbf{C}}^{-1}\mathbf{Q}\right)\left(\mathbf{Q}^{\mathbf{T}}\hat{\mathbf{C}}^{-1}\mathbf{Q}\right)^{-1}\left(\mathbf{q}(s_0) - \mathbf{c}_0^{\mathbf{T}}\hat{\mathbf{C}}^{-1}\mathbf{Q}\right)^{\mathbf{T}}.$$

Capítulo 3

Aplicación al cálculo del espesor del hielo

Llevemos ahora los procedimientos que hemos estudiado en el anterior capítulo a la práctica con los datos que preparamos al comienzo de este trabajo. Comenzaremos este capítulo comprobando la adecuación de los datos a los modelos que hemos introducido, continuaremos calculando una predicción para los modelos Kriging considerados y finalizaremos comparando su precisión con otros posibles modelos.

Por limitaciones de computación, consideramos solamente una muestra aleatoria de $M = 50\ 000$ filas de nuestra matriz de observaciones, aproximadamente el 40 % del número total, que dividiremos en datos de entrenamiento y de evaluación en una proporción del 70 % y 30 %, respectivamente. Así, nos limitamos a realizar una sola iteración de validación cruzada, en cuyo caso podemos llamarlo método de retención.

3.1. Análisis preliminar

Analicemos las relaciones entre los datos para comprobar que se puedan aplicar los métodos descritos. Con el M considerado producimos un conjunto de datos de entrenamiento de $0{,}7M = 35\ 000$ observaciones. Dibujamos en la Figura 3.1 la matriz de gráficas de dispersión para analizar sus relaciones internas y asegurarnos de que la dependencia entre covariables sea tal que la inversión de su matriz $\mathbf{Q}$ esté bien definida.

A través de las correlaciones con Espesor vemos las covariables más influyentes en su predicción lineal. Además, observamos que la correlación entre Altitud y Distancia a la costa es elevada

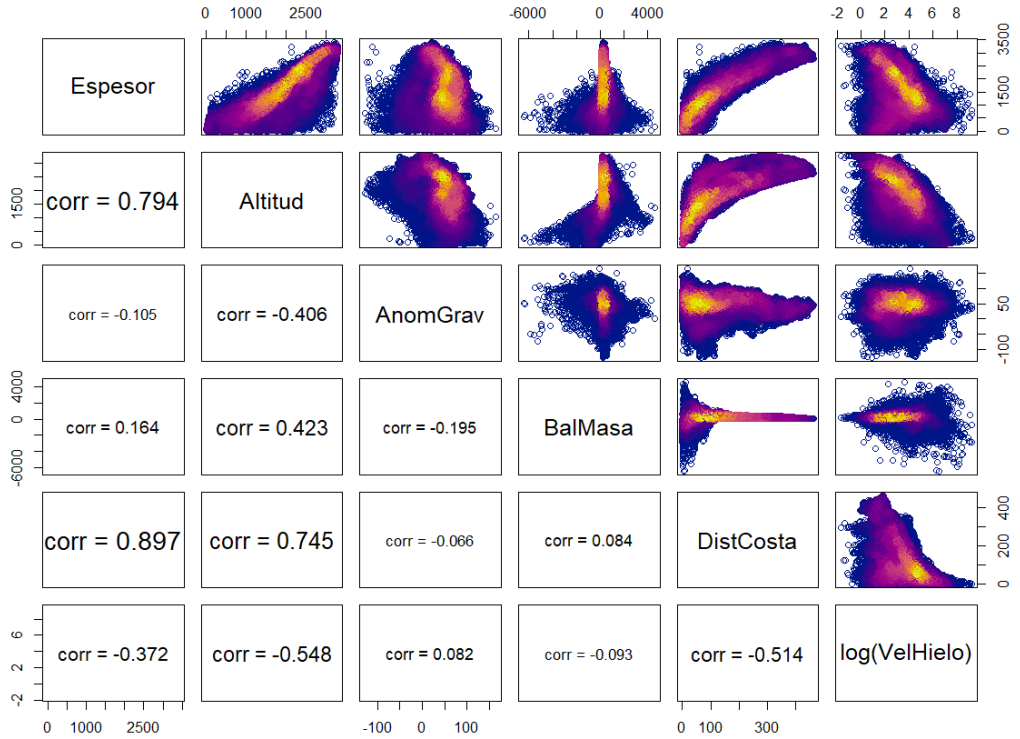


Figura 3.1: Matriz de gráficas de dispersión de todos los pares de variables, cuya correlación se corresponde al simétrico respecto a la diagonal. Colores más cálidos reflejan una mayor densidad de puntos.

por su similitud. Sin embargo, elegimos conservar las dos, ya que no impiden la inversión de la matriz $\mathbf{Q}$.

De cara al resto de este trabajo, consideraremos el modelo de regresión lineal múltiple (MLR) con la siguiente fórmula, en la que representamos las variables con sus siglas:

$$E = \beta_0 + \beta_1 A + \beta_2 AG + \beta_3 BM + \beta_4 DC + \beta_5 log(VH) + \varepsilon. \tag{3.1}$$

Comprobemos ahora, usando el paquete `performance, versión 0.13.0` (Lüdecke et al. 2021), si este modelo se ajusta a las hipótesis. Decidiremos mediante un análisis visual de las gráficas de la Figura 3.2, algo preferible al uso de tests estadísticos en el caso de trabajar con grandes cantidades de datos.

En la Figura 3.2, la gráfica en posición $(1, 1)$ compara la distribución empírica y la teórica generada por el modelo, donde podemos ver que se ajusta bien a los datos. Comprobamos una linealidad aceptable en la gráfica $(1, 2)$, aunque los residuos se curvan bastante en predicciones

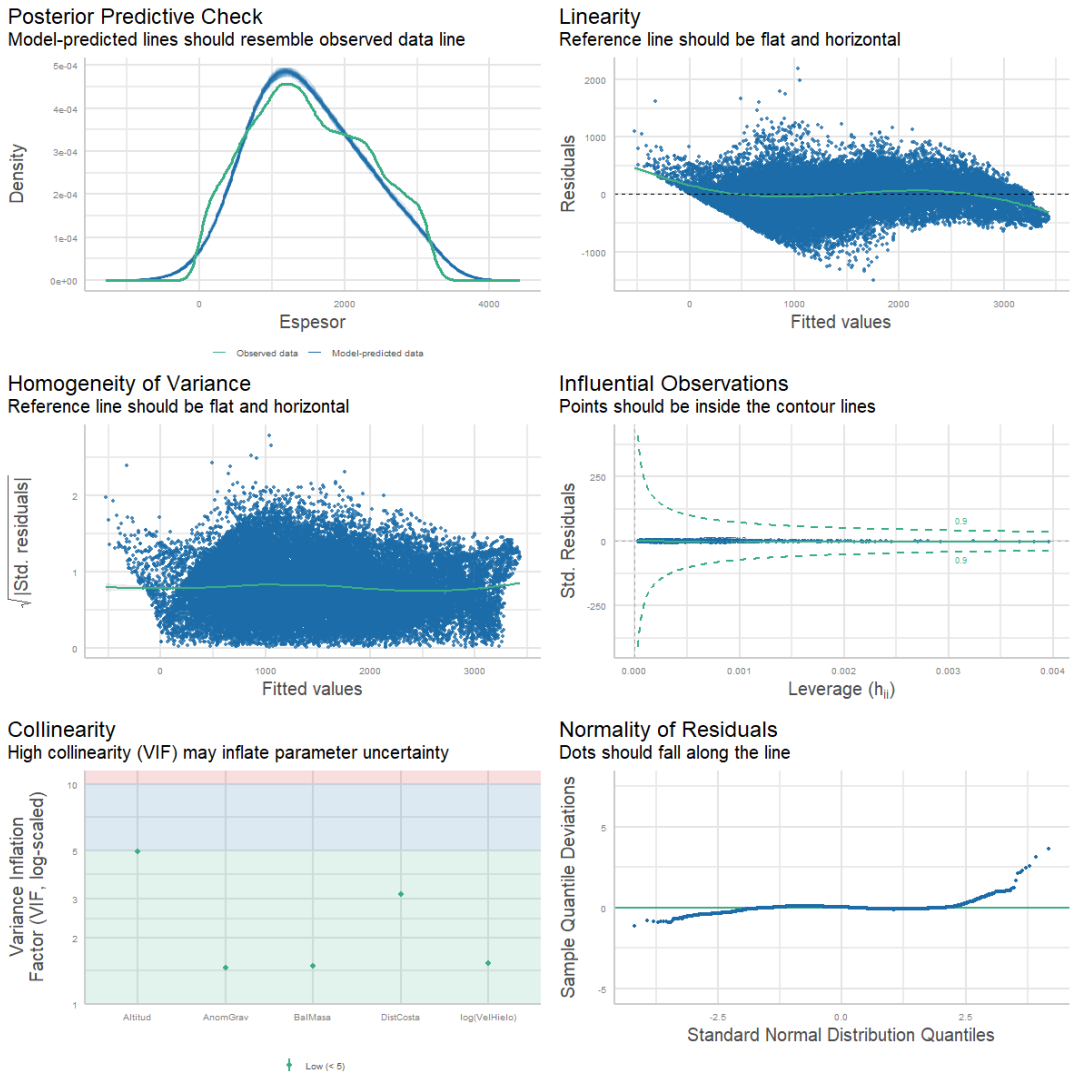


Figura 3.2: Diagnóstico del modelo de regresión lineal múltiple 3.1 sobre nuestros datos procesados, a través de la función `performance::check_model`.

extremas, lo que nos indica que aún hay cierta influencia no lineal. Además, como podemos ver mejor en la gráfica $(2, 1)$, su variación es bastante constante, lo que nos lleva a concluir que presentan un buen nivel de homocedasticidad.

La gráfica $(2, 2)$ no marca ninguna observación demasiado influyente por medio de la distancia de Cook. La $(3, 1)$ nos muestra la correlación relativamente elevada que ya conocemos entre las covariables Altitud y Distancia a la costa, aunque sigue sin suponer un problema. Por último, la gráfica $(3, 2)$ comprueba la normalidad de los residuos, hipótesis al cual se ajusta relativamente bien excepto en los extremos, como la linealidad. En conclusión, el entrenamiento del modelo 3.1 con nuestros datos resulta en un cumplimiento razonable de las hipótesis de regresión.

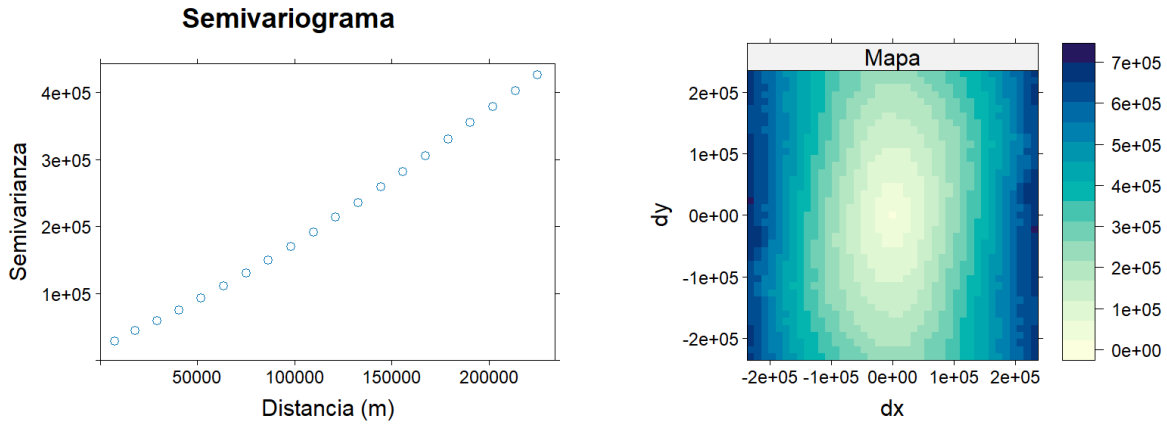
3.1.1. Estimación del semivariograma

Efectuaremos los cálculos con el paquete `gstat, versión 2.1-2`, de `R`, en donde podemos comprobar que sus ecuaciones, escritas en el Apéndice A de su manual (*gstat user manual* 2025), son idénticas a las indicadas en este trabajo. Además, provee el método que usaremos para estimar los semivariogramas teóricos a partir de los experimentales, el de mínimos cuadrados iterativamente ponderados, para el cual seleccionaremos los pesos $\frac{N_l}{\hat{\gamma}(\tilde{h}_l)^2}$. Representaremos a todos los semivariogramas en términos de 20 intervalos $\tilde{h}_l$.

Comenzamos extendiendo el modelo de regresión al caso de dependencia espacial con Kriging Ordinario, donde estimamos su componente constante $m(s) = m$ con la media muestral. La ecuación resultante es:

$$\hat{E}(s_0) = \hat{m} + \sum_{j=1}^{n} \hat{\omega}_j(s_0)\varepsilon_j, \tag{3.2}$$

donde ε_j es el residuo resultante en la observación j. Para analizar la dependencia espacial, calcularemos el semivariograma experimental sobre los residuos, en la Figura 3.3a, así como un mapa que lo visualiza espacialmente, en la Figura 3.3b. Fijamos una distancia máxima de $\|h\| \leq \frac{\text{máx}\{DistCosta\}}{2} \approx 231$ km en ambos.



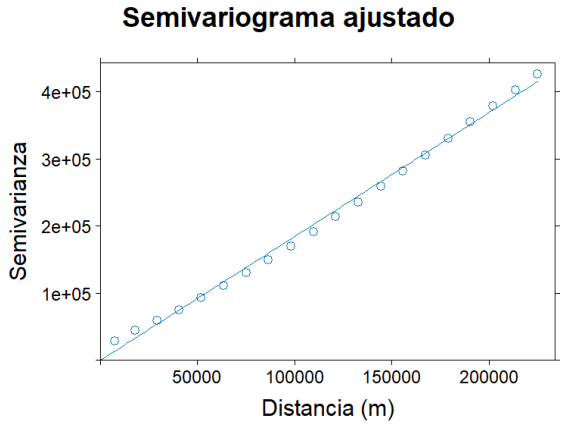
(a) Semivariograma muestral.

(b) Mapa del semivariograma muestral.

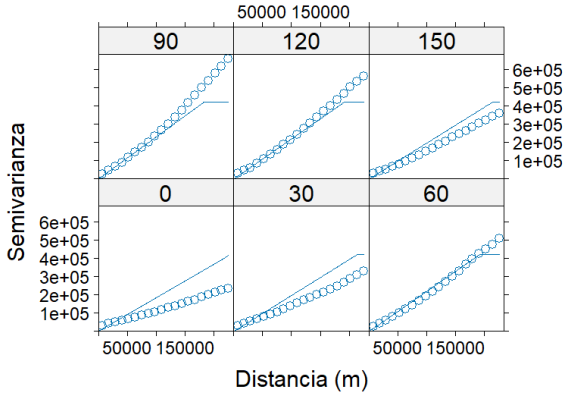
Figura 3.3: Semivariograma experimental de los residuos definidos al aplicar Kriging Ordinario a los datos de entrenamiento.

Los residuos son claramente anisotrópicos, con una dependencia espacial visiblemente más fuerte en la componente horizontal de la Figura 3.3b. Considerando la forma alargada de Groenlandia, esto puede atribuirse a que sea la dirección asociada a una menor distancia entre costas,

ya que estas presentan la mayor cantidad de variación. Para compensar esta anisotropía, aplicaremos una transformación de coordenadas $\tilde{y} = \frac{4}{5}y; \tilde{x} = x$. Por otro lado, en la Figura 3.3a vemos que el semivariograma no parece estabilizarse en un umbral sino que crece constantemente, una indicación de que es un error suponer una media constante, por lo que habría que limitarse a una estacionariedad intrínseca o tratar de estimar una tendencia externa mediante covariables. Por ahora, continuemos con la primera opción, continuando así con el ajuste en la Figura 3.4, considerando un modelo lineal.



(a) Modelo lineal ajustado sobre el semivariograma muestral.



(b) Modelo lineal ajustado sobre el mapa del semivariograma muestral, en incrementos de 30° en sentido horario comenzando en orientación Norte.

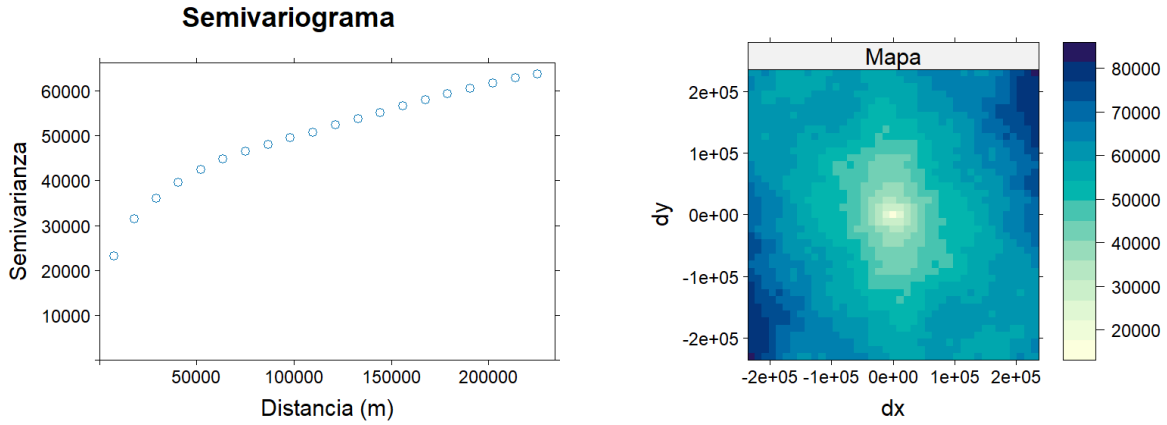
Figura 3.4: Semivariograma teórico lineal ajustado sobre los experimentales de la Figura 3.3.

En la Figura 3.4a vemos un ajuste donde el efecto nugget termina siendo cero, y la Figura 3.4b nos muestra que el semivariograma muestral y el modelo lineal difieren notablemente, en especial tras los 100 km. Como alternativa, tratemos ahora de reducir el efecto de esta tendencia a través de considerar la regresión lineal múltiple de ecuación 3.1 a través del modelo Kriging Universal, con lo que nos queda la siguiente ecuación:

$$\hat{E}(s_0) = \hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 AG + \hat{\beta}_3 BM + \hat{\beta}_4 DC + \hat{\beta}_5 log(VH) + \sum_{j=1}^{n} \hat{\omega}_j(s_0)\varepsilon_j. \tag{3.3}$$

Procederemos sobre sus residuos de forma idéntica al caso de Kriging Ordinario, calculando los semivariogramas de la Figura 3.5.

En la Figura 3.5a identificamos inmediatamente una dependencia espacial mejor definida que en el caso anterior, ya que vemos una clara tendencia a umbral tras superar los 50 km, resultando en una curva que recuerda al modelo teórico exponencial. Desplazando la vista hasta la Figura

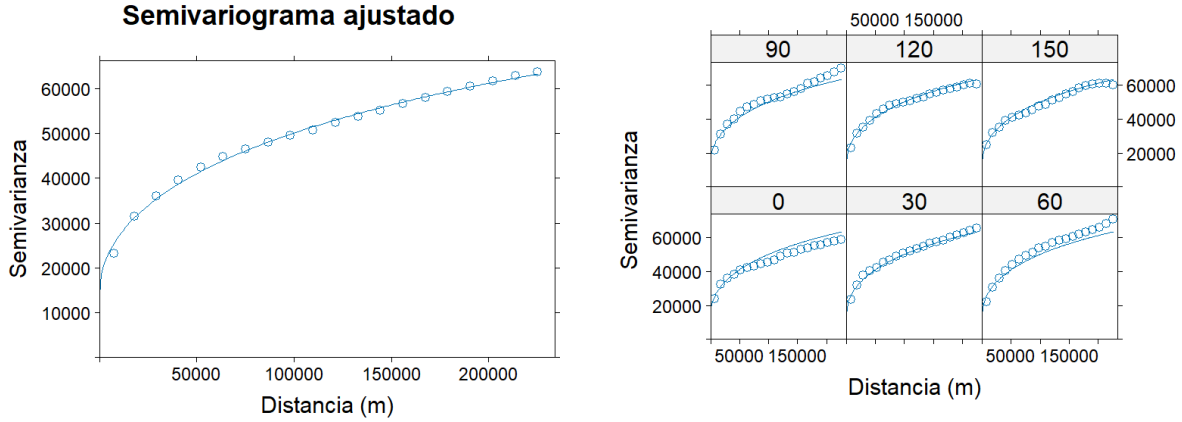


(a) Semivariograma muestral.

(b) Mapa del semivariograma muestral.

Figura 3.5: Semivariograma experimental de los residuos definidos al aplicar el Kriging Universal considerado a los datos de entrenamiento.

3.5b, no apreciamos apenas anisotropía, lo que significa que la inclusión de covariables también lo han podido mitigar. Por tanto, en la Figura 3.6 ajustamos un semivariograma teórico estable, suponiendo la ausencia de anisotropía.



(a) Modelo estable ajustado sobre el semivariograma muestral.

(b) Modelo estable ajustado sobre el mapa del semivariograma muestral, en incrementos de 30° en sentido horario comenzando en orientación Norte.

Figura 3.6: Semivariograma teórico estable con $\alpha = 0{,}5$ ajustado sobre el experimental de la Figura 3.5.

La Figura 3.6a nos muestra un buen ajuste, apoyada por la Figura 3.6b, en la cual el modelo no se aleja demasiado del semivariograma experimental para ninguna dirección en concreto. Tras los 200 km, sin embargo, el semivariograma muestral comienza a inestabilizarse en algunas direcciones.

3.2. Predicción

Finalmente, estamos listos para predecir usando los modelos Kriging Ordinario y Universal concretados, en donde los mapas de covariables serán nuestros datos de predicción, ya que incluyen implícitamente sus coordenadas.

Tal y como mencionamos, el semivariograma lineal de nuestro Kriging Ordinario empeora tras los 100 km, por lo que consideramos el modelo localmente, en un entorno de 100 km de radio al rededor de cada instancia de predicción.

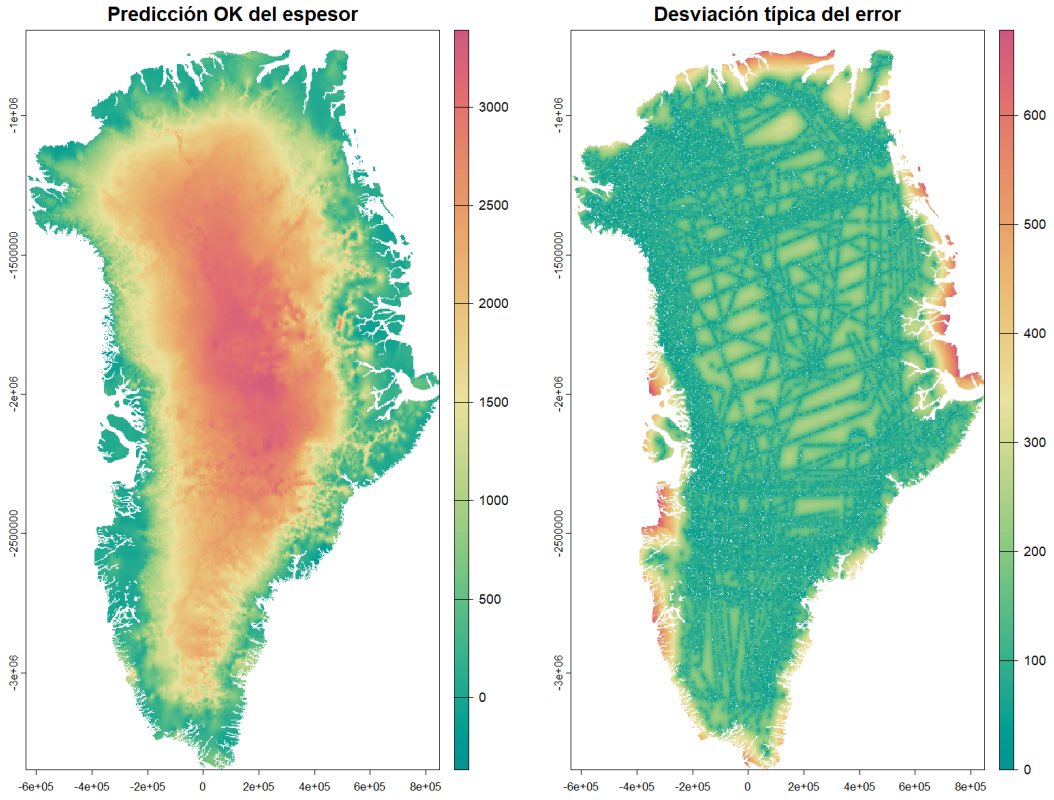


Figura 3.7: Predicción final de nuestro modelo Kriging Ordinario.

En la Figura 3.7 vemos como el Kriging Ordinario, gracias a la gran cantidad de observaciones y a la restricción de radio, es capaz de proporcionarnos un mapa razonable del espesor de la capa de hielo. Sin embargo, detectamos deficiencias en la predicción en las cuales, por ser un interpolador exacto, ha tenido que alcanzar súbitamente el espesor medido en las observaciones o en donde presenta una elevada desviación típica, como zonas del contorno costero en donde no disponemos de tantos datos. En adición, podemos sospechar de que la anomalía en los puntos al rededor de las coordenadas $(500\ 000, 1\ 750\ 000)$ sea debido a valores atípicos que no fueron identificados y eliminados de antemano.

Habíamos visto que el semivariograma teórico de Kriging Universal se mantiene ajustado en todas las distancias hasta los 200 km, tras el cual comienza a desviarse en algunas direcciones, por lo que para sus predicciones elegimos considerar entornos locales con un radio de 200 km.

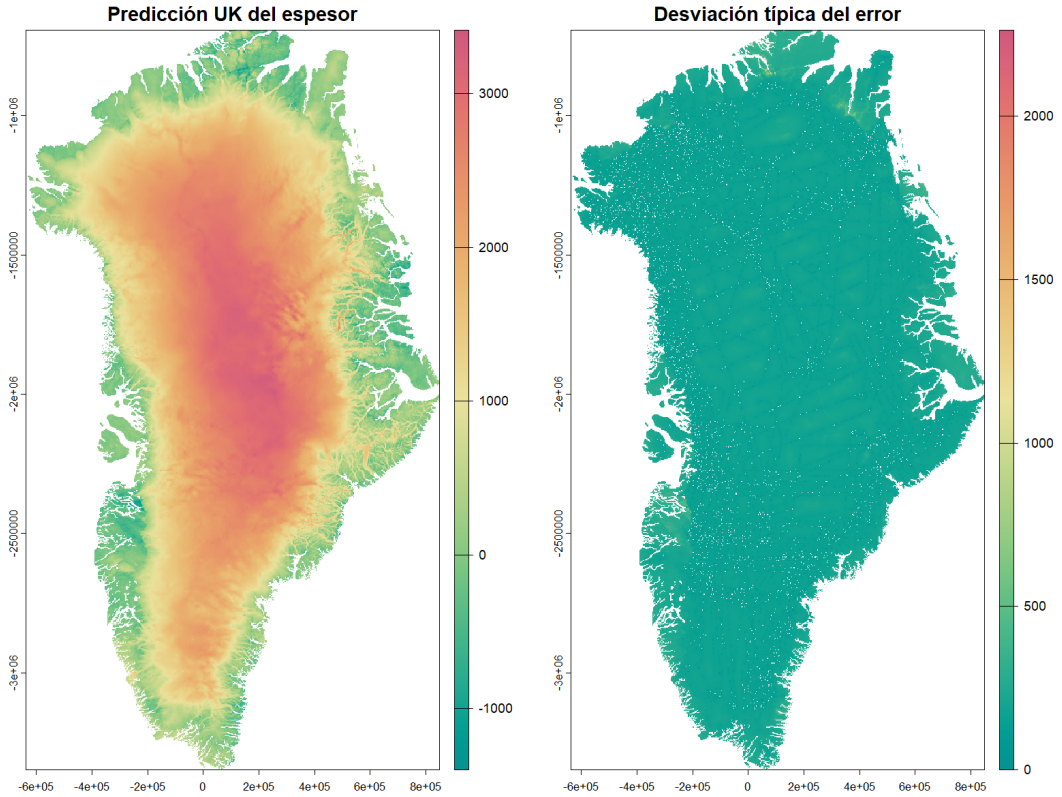


Figura 3.8: Predicción final de nuestro modelo Kriging Universal.

En la Figura 3.8 observamos una diferencia notable en términos de suavidad y detalle, especialmente en la costa. Mientras que la predicción de OK tenía manchas con un cambio repentino de predicciones al rededor de varias observaciones, Kriging Universal las incorpora detallando su apariencia a través de las covariables. Como ejemplo de esto, vemos que en la zona al rededor del punto $(500\,000, 1\,750\,000)$, mencionado anteriormente, el espesor atípico es explicado como un fiordo profundo.

Por el otro lado, no podemos interpretar adecuadamente la desviación típica por culpa de varios casos atípicos en el norte de la isla. Para solucionar esto, establecemos la desviación típica máxima de OK como un límite arbitrario, eligiendo no representar a las predicciones cuya desviación típica supera este valor. Además, tampoco mostraremos en el mapa los puntos en donde la predicción es negativa o nula, entendiendo esto como una ausencia de hielo. Implementaremos estos cambios para el resto de mapas, comenzando con la Figura 3.9.

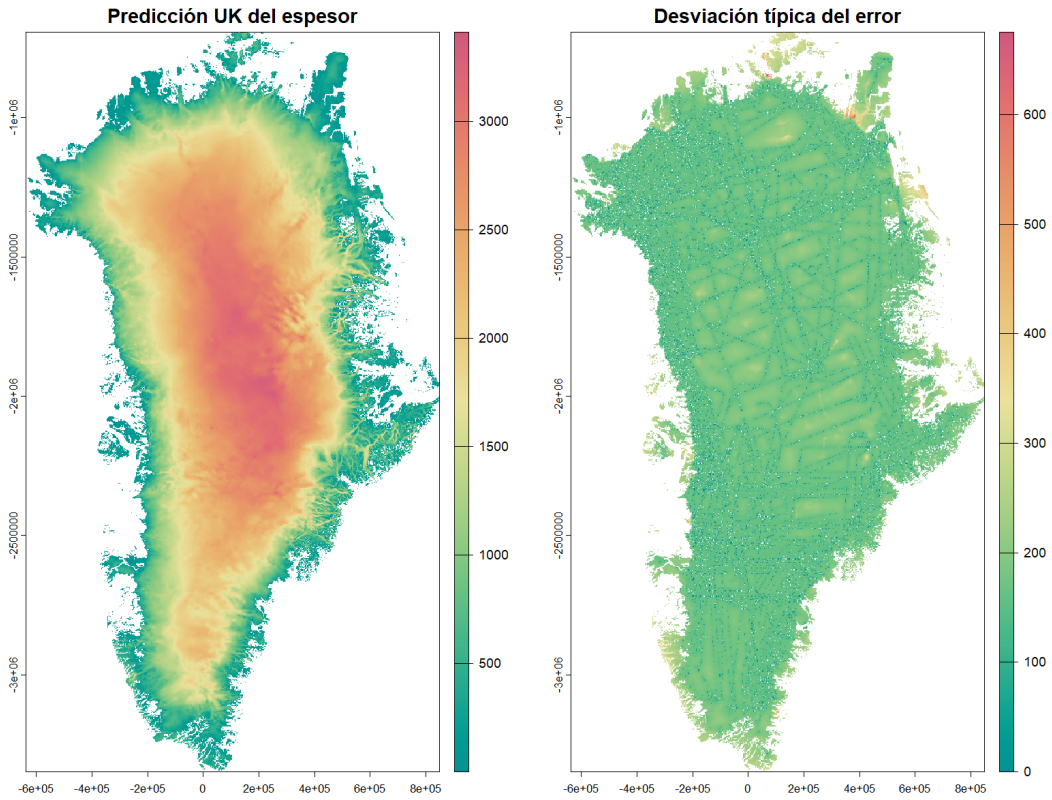


Figura 3.9: Predicción final de nuestro modelo Kriging Universal, sin considerar 93 565 puntos en donde la predicción es negativa o nula ni 123 en las que se supera la desviación típica máxima de OK.

Por tanto, el Kriging Universal demuestra claramente ser mejor predictor para zonas alejadas de las observaciones, resultando en menos desviación típica, gracias a las covariables.

3.2.1. Otros modelos

Aparte de los modelos Kriging Ordinario (OK), de ecuación 3.2, y Kriging Universal (UK), de ecuación 3.3, representaremos también el resultado de predecir utilizando la regresión lineal múltiple (MLR), de ecuación 3.1, sobre nuestros datos de entrenamiento, en la Figura 3.10.

Al no considerar dependencia espacial, el modelo MLR pierde mucho detalle con respecto a Kriging, resultando en superficies más suaves. Introduciremos a continuación tres modelos adicionales, que compararemos con las demás.

Primero, el método de ponderación por el inverso de la distancia (IDW) es una extensión de MLR a los casos de dependencia espacial, al igual que Kriging. Sin embargo, mientras que parte de la misma idea de inferir el residuo en un punto desconocido a partir de una combinación lineal

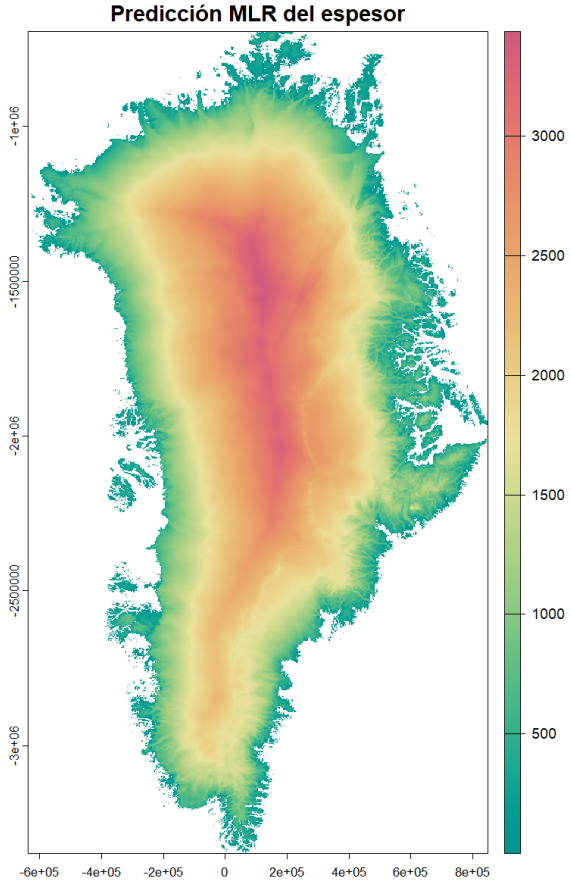


Figura 3.10: Predicciones positivas del modelo MLR.

de los residuos conocidos en las observaciones, los pesos que utiliza son más simples que en el caso Kriging. La ecuación general de un modelo IDW, a partir de un modelo de regresión lineal múltiple, será la siguiente:

$$Z(s_0) = \beta_0 + \sum_{k=1}^{p} q_k(s_0)\beta_k + \sum_{j=1}^{n} \omega_j(s_0)\varepsilon_j,$$

donde considera $\omega_j = \frac{1}{d(s_0,s_j)^p}$ la inversa de la distancia euclidiana entre la instancia de predicción y la observación j, elevada a la potencia p. Nosotros fijaremos $p = 2$, por lo que la ecuación que consideraremos para nuestros datos es:

$$\hat{E}(s_0) = \hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 AG + \hat{\beta}_3 BM + \hat{\beta}_4 DC + \hat{\beta}_5 log(VH) + \sum_{j=1}^{n} \frac{\varepsilon_j}{d(s_0, s_j)^2}.$$

Al aplicar esta fórmula sobre nuestros datos de entrenamiento, obtenemos la Figura 3.11.

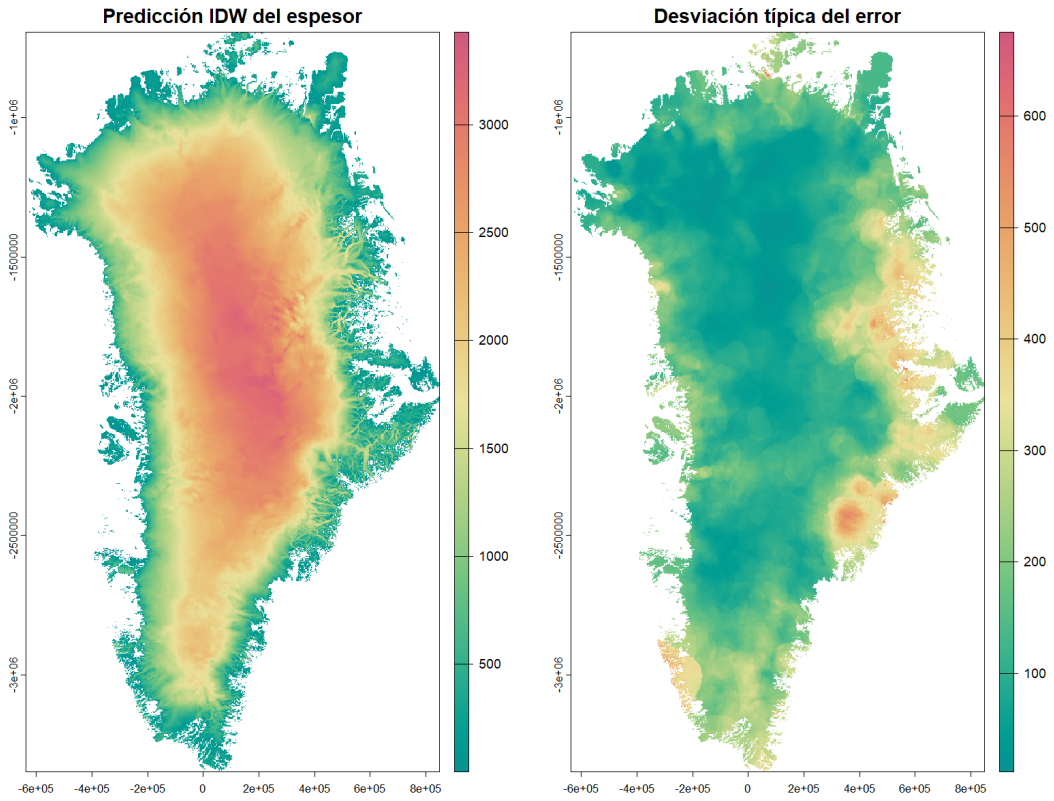


Figura 3.11: Predicciones positivas del modelo IDW que no superan el límite de desviación típica.

Mientras que el detalle mejora con respecto al MLR, no llega a la exactitud de Kriging Universal por culpa de elegir arbitrariamente una función de peso. Sin embargo, es una opción computacionalmente barata que puede incrementar la calidad de predicción en casos de dependencia espacial. Además, puede venir acompañado por una estimación de la varianza del error de predicción, que ayuda a interpretar los resultados.

En segundo lugar, el modelo BedMachine (BM) (Morlighem et al. 2017) es, en la actualidad, el mapa topológico y batimétrico más avanzado para las grandes extensiones de hielo del planeta, como la Antártida o Groenlandia. El proyecto está manejado por el NSIDC, de la Universidad de Colorado Boulder, y financiado por la NASA. Consideramos la versión más reciente del modelo (Morlighem y al 2022), representado en la Figura 3.12 tras ajustarlo a nuestra rejilla a través de una interpolación bilinear.

Veamos las diferencias más relevantes entre el procedimiento seguido por este proyecto, según su documentación (*BedMachine user guide* 2024), y el que hemos empleado nosotros para los modelos Kriging:

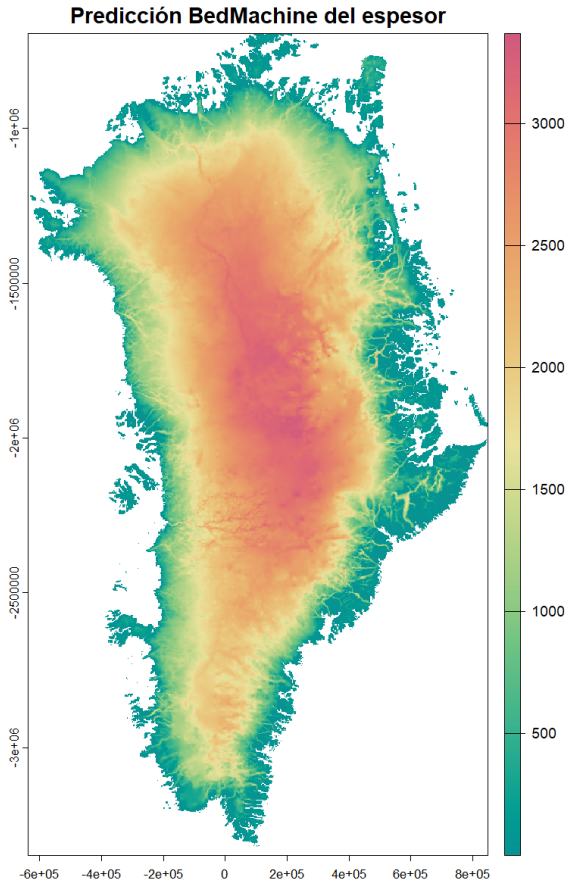


Figura 3.12: Predicción positiva del modelo BedMachine, versión 5.

- **Espesor:** BedMachine usa principalmente las mismas mediciones del espesor de hielo que nosotros al tomar todas las expediciones Greenland Platform 3 de la NASA entre 1993 y 2017.
- **Covariables:** Mientras que coincidimos en el uso de covariables como Balance de masas, Altitud o Velocidad del hielo, nosotros hemos incluido a mayores Anomalía Gravitatoria y Distancia a la costa.
- **Método:** Nosotros utilizamos Kriging para predecir el espesor del hielo en toda Groenlandia. Sin embargo, BedMachine sólo aplica este procedimiento en el interior y únicamente con los datos de entre 1993 y 2016, mientras que para los datos del 2017 usa un método de difusión procedente de la dinámica de fluidos. Además, cerca de la costa se deriva el espesor a través de la ecuación de conservación de la masa y la velocidad del hielo. Veamos las regiones exactas en la Figura 3.13, donde también incluiremos las zonas donde se usó la covariable Altitud para determinar la elevación en áreas sin hielo.

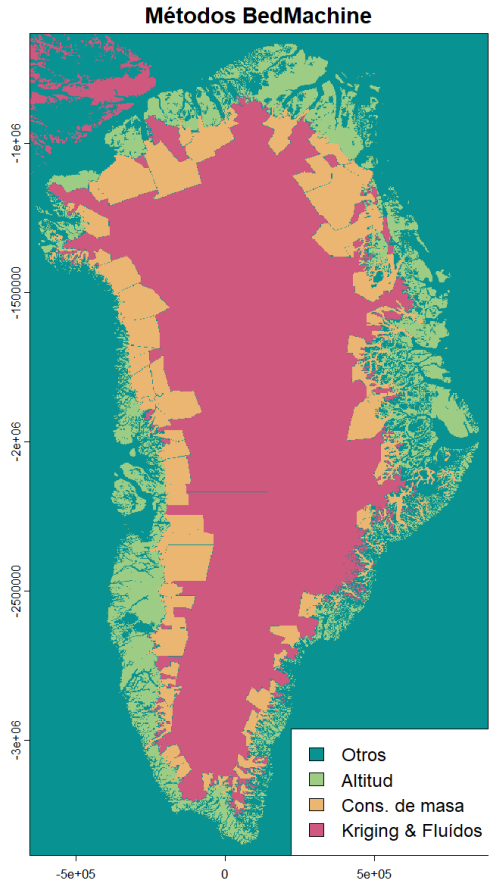


Figura 3.13: Métodos más relevantes usados por BedMachine, versión 5.

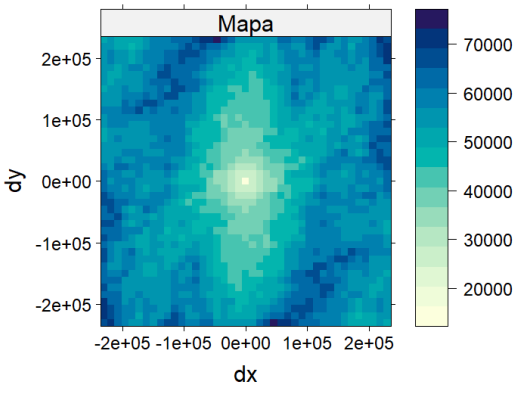
El tercer modelo que compararemos es UK entrenado con otros datos (UK_{197X}), tomados con radar desde aviones en los años setenta y digitalizados recientemente (Karlsson et al. 2024). Esto nos aportará una visión de la diferencia que puede suponer cambiar los datos de entrenamiento a otros con más error y menor número de observaciones. También es posible, como algunos de estos datos son de hace más de medio siglo, que en el tiempo transcurrido se haya modificado en alguna medida la capa de hielo. Veamos, por ejemplo, mediciones del vuelo número 10 del año 1974:

CBD	lon_degrees	lat_degrees	srf_elev_m.asl	bed_elev_m.asl
89	-46.503	67.072	1908	NA
90	-46.497	67.053	1892	60
91	-46.487	67.038	1900	NA
92	-46.482	67.022	1914	108
93	-46.477	67.007	1912	NA
94	-46.467	66.988	1911	105

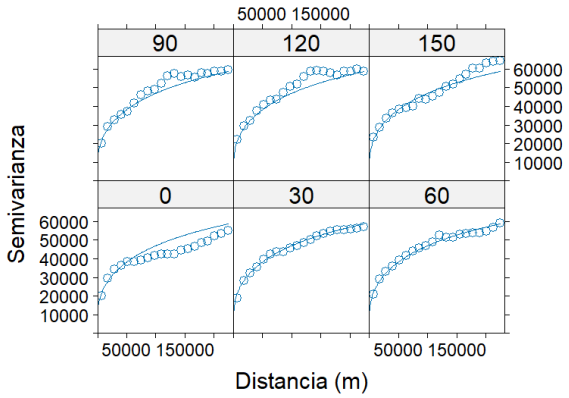
Las medidas de Espesor las tenemos que calcular manualmente como la diferencia entre las columnas tercera y cuarta, que expresan la altitud sobre el nivel del mar de la superficie y del continente debajo del hielo. Nótese que en algunos casos aparecen mediciones vacías en forma de *NaN*, por lo que eliminaremos estas observaciones además de los 7 valores atípicos globales que detectamos a través de la función `performance::check_outliers`. La matriz de observaciones final consta de 26 524 observaciones, con una estructura idéntica a la construida con los datos de CReSIS:

Espesor	Altitud	AnomGrav	BalMasa	DistCosta	VelHielo	x	y
219	986.6743	82.46817	-303.2745	58.27979	5.602218	130621.1	-881318.3
144	1032.1473	81.44789	-246.6447	58.52967	1.984312	130621.1	-882890.8
332	1066.3364	78.79760	-194.1033	58.82048	3.448803	130621.1	-884463.4
636	1097.9789	74.04370	-149.9691	59.15177	15.590328	130621.1	-886036.0
1006	1127.0452	70.60751	-126.6524	61.04703	13.973682	129048.5	-887608.6
350	164.8858	140.16660	-1475.8657	2.44681	227.135925	196669.1	-889181.1

Consideramos la matriz de observaciones entera como datos de entrenamiento, ya que para compararlo utilizaremos los datos de evaluación procedentes de CReSIS. Predeciremos con UK, por lo que seguimos todos los pasos ya detallados, comprobando que cumple un poco peor las hipótesis de regresión pero que se sigue obteniendo un ajuste bueno del semivariograma teórico, visible en la Figura 3.14b, sobre direcciones del mapa del semivariograma muestral representada en la Figura 3.14a.



(a) Mapa del semivariograma muestral.



(b) Modelo estable ajustado sobre el mapa del semivariograma muestral, en incrementos de 30° en sentido horario comenzando en orientación Norte.

Figura 3.14: Semivariograma teórico estable con $\alpha = 0{,}5$ ajustado al experimental considerado UK sobre los datos de entrenamiento de 1971-1979.

Con esto, podemos realizar la predicción. A diferencia de cuando usamos los datos de CReSIS, aplicaremos UK en todo el territorio en lugar de un entorno local, motivados por el menor número de puntos presentes en los datos de entrenamiento de 1971-1979, que resulta en la Figura 3.15.

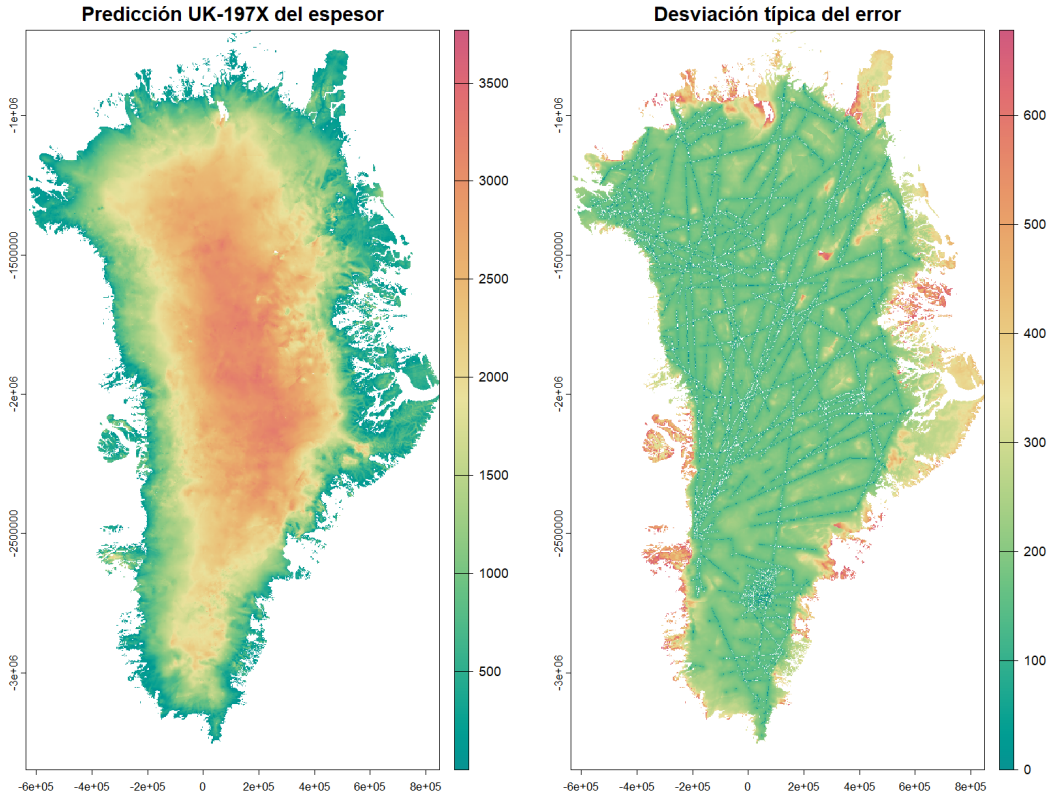


Figura 3.15: Predicciones positivas de KU entrenado sobre los datos de 1971-1979 que no superan el límite de desviación típica.

3.2.2. Comparación

Comparamos ahora todos los modelos introducidos a través de la predicción en los datos de evaluación, que consta de $0{,}3M = 15\ 000$ observaciones de CReSIS. Para medir la precisión con la que logran predecir usamos la raíz del error cuadrático medio (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)},$$

para N observaciones, siendo y_j la observación j y $\hat{y}_j$ la predicción del modelo, y el coeficiente de determinación R^2:

$$R^2 = 1 - \frac{\sum_{j=1}^{N} (y_j - \hat{y}_j)}{\sum_{j=1}^{N} (y_j - \bar{y})},$$

siendo $\bar{y} = \frac{1}{N} \sum_{j=1}^{N} y_j$. Un mejor modelo viene acompañado de un RMSE más bajo, al igual que un R^2 más cercano a 1. Tras extraer las predicciones de los mapas correspondientes a los datos de evaluación y aplicar las métricas, escribimos los resultados en el Cuadro 3.1:

Modelo	$RMSE$	R^2
UK_{197X}	436,82	0,708
MLR	282,70	0,878
IDW	156,75	0,962
OK	145,38	0,968
UK	133,26	0,973
BM	103,43	0,984

Cuadro 3.1: Comparación de la precisión entre modelos, con las métricas raíz del error cuadrático medio (RMSE) y coeficiente de determinación (R^2).

Capítulo 4

Conclusiones

Con todos los cálculos hechos, este capítulo lo dedicamos a la discusión del trabajo realizado. Recordemos que nuestro objetivo es estimar el volumen y área total de la capa de hielo sobre Groenlandia y, como consecuencia del proceso, comparar distintos métodos para interpretar las ventajas de cambiar a un modelo más o menos complejo.

A esta segunda cuestión ya podemos responder con los datos del Cuadro 3.1. Están ordenados en orden creciente de complejidad, con la excepción de UK_{197X}, que consideraremos como un caso aparte. Vemos como se reduce progresivamente el RMSE mientras que el R^2 crece cada vez más. Por tanto, este caso refleja satisfactoriamente lo que de manera ideal pensamos que debería suceder: dado un orden de modelos de creciente complejidad, cada una mejorará en precisión respecto a la anterior. Naturalmente, esto vendrá a coste de más consideraciones y mayor tiempo de computación.

El modelo BM logra una diferencia notable con respecto a los demás, razón por la que se reconoce como el mejor modelo actual. Sin embargo, para un público menos exigente el modelo UK puede ser una opción satisfactoria, incluso más atractiva ya que es de implementación mucho más sencilla. OK no es una buena elección para este caso por el semivariograma que obtuvimos, y UK será claramente superior sin requerir mucho más esfuerzo ni tiempo de computación. En el siguiente escalón descendiente tenemos a IDW, un modelo muy flexible que, al igual que Kriging, puede incorporar una regresión previa o no, con su ventaja siendo un funcionamiento sencillo de entender. Por último tenemos a MLR, cuya ventaja más grande es el tiempo de cómputo, notablemente menor que el del resto de modelos.

Como caso aparte tenemos a UK_{197X}, donde no sólo reducimos el número de datos de entrenamiento sino que los cambiamos por completo a otros con una incertidumbre posicional más elevada (Karlsson et al. 2024). Por esta razón, su predicción resulta muy inestable, con cambios bruscos, por lo que su precisión es mucho menor.

Calculemos por fin en el Cuadro 4.1 las estimaciones del volumen y área que hemos estado buscando todo este tiempo. Dados los mapas de predicciones, el cálculo es prácticamente trivial utilizando el método comentado en la introducción de este trabajo, utilizando las celdas en donde el espesor predicho es positivo. Añadiremos un paso, sin embargo, donde cambiamos sus unidades de medida para incrementar su interpretabilidad. Para el área, en lugar de expresarlo en millones de kilómetros cuadrados, lo escribiremos como un porcentaje respecto al área total de Groenlandia según nuestra región de estudio. Para el volumen, en lugar de expresarlo en millones de kilómetros cúbicos o de gigatoneladas, lo convertiremos a su incremento equivalente del nivel del mar (SLE), por medio de que 361.8 Gt de hielo es equivalente a un incremento global del nivel del mar de 1 mm (*Calculating glacier ice volumes and sea level equivalents* 2025):

Modelo	$SLE(m)$	$Área(\,\%)$
UK_{197X}	7,650	87,99 %
MLR	7,707	89,62 %
IDW	7,700	89,48 %
OK	7,819	98,14 %
UK	7,690	89,06 %
BM	7,477	89,61 %

Cuadro 4.1: Incremento equivalente del nivel del mar (SLE), en metros, y porcentaje de la superficie de Groenlandia que ocupa la capa de hielo, a partir de la predicción de cada modelo.

La mención más notable del Cuadro 4.1 es el área cubierta de hielo predicha por OK, acompañada por ser, aunque no por mucho, el modelo con el valor más alto de SLE. Atribuimos este resultado absurdo a la mala elección del modelo, ya que su semivariograma muestral nos sugería incluir una tendencia externa por su elevado ratio de crecimiento. Es también curioso que el modelo más fiable, BM, tenga el SLE notablemente más pequeño, que se podría relacionar con su mejor definición en las zonas de costa por las ecuaciones de conservación de la masa.

Como continuación a este trabajo podríamos añadir más modelos con las que comparar, modelizar la tendencia con regresión múltiple no lineal para incluir el efecto de interacciones entre covariables o simplemente transformar, eliminar o añadir variables para mejorar la predicción.

Anexo I

Figuras

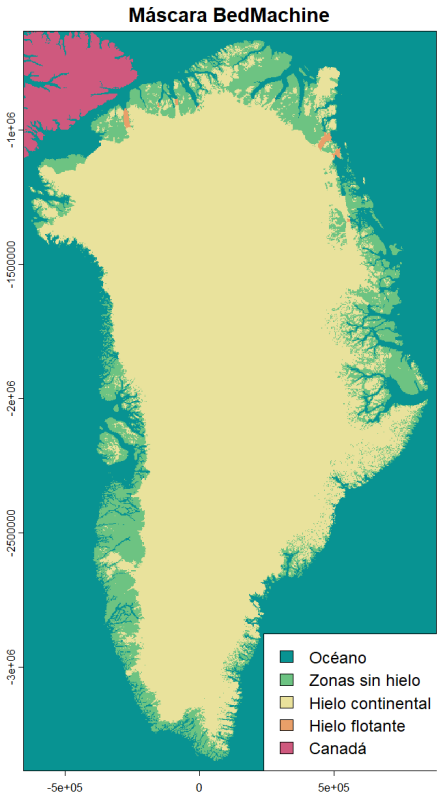


Figura I.1: Máscara utilizada por el modelo BedMachine.

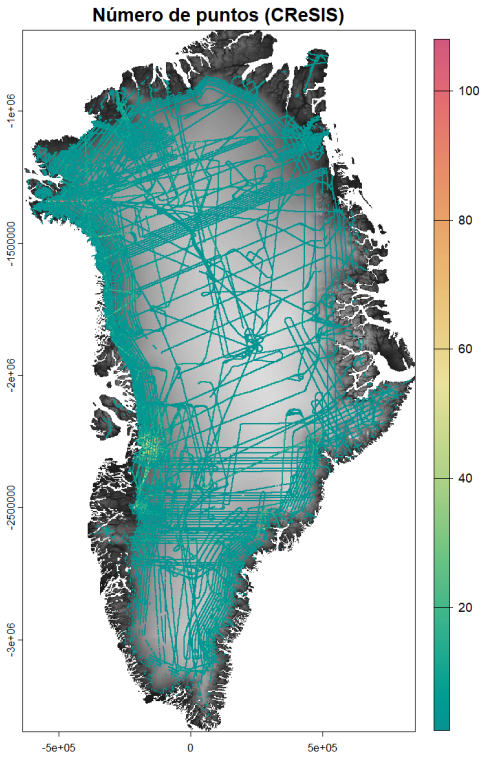


Figura I.2: Densidad de los datos CReSIS representada en número de puntos promediados en cada celda de la rejilla.

Anexo II

Código de R

```r
#» No incluiremos las funciones para generar las imágenes por su tamaño
#» Evitamos repetir secciones similares entre sí para acortar

options(digits = 8, max.print = 100) # formato de consola

# install.packages("terra")
library(terra)
# install.packages("performance")
library(performance)
# install.packages("gstat")
library(gstat)
```

II.1. Procesado de covariables

```r
#> Importamos los datos y creamos los rasters
Altitud.frast <- rast("./features/GIMP/gimpdem_90m_v01.1.tif")
crs(Altitud.frast) <- crs("EPSG:3413")

gravity <- read.table("./features/gravity_anomaly/grid_bouguer.txt")
gravity_lonlat.rast <- rast(gravity, type = "xyz", crs = "EPSG:4326")
Anomalia_gravitatoria.frast <- project(gravity_lonlat.rast, "EPSG:3413")

Balance_de_masa.frast <- rast("./features/surface_mass_balance/smb_rec-mean.nc")
```

```
crs(Balance_de_masa.frast) <- crs("EPSG:3413")

bedmachine_mask.rast <- bedmachine.rast$mask
bedmachine_mask.rast[bedmachine.rast$mask == 2] <- 1
Distancia_a_la_costa.frast <- distance(bedmachine_mask.rast, target = 1)/1e3
crs(Distancia_a_la_costa.frast) <- crs("EPSG:3413")

icevelx.rast <- rast("./features/ice_velocity/greenland_vel_mosaic250_vx.tif")
icevely.rast <- rast("./features/ice_velocity/greenland_vel_mosaic250_vy.tif")
Velocidad_del_hielo.frast <- lapp(c(icevelx.rast, icevely.rast),
fun = function(x, y){ return(sqrt(x^2+y^2)) })
crs(Velocidad_del_hielo.frast) <- crs("EPSG:3413")

#> Los organizamos en una lista por eficiencia
raw_covariables.rastlist <- mget(ls(pattern = "\\.frast$"))
names(raw_covariables.rastlist) <-
gsub(".frast", "", names(raw_covariables.rastlist))

#> Transformamos importados a la extensión y resolución del de menos resolución
name_biggest <- names(which.max(sapply(raw_covariables.rastlist,
function(rast){ prod(res(rast)) })))
resampled_covariables.rastlist <- lapply(raw_covariables.rastlist,function(rast)
resample(rast, raw_covariables.rastlist[[name_biggest]], method = "bilinear"))

#> Construimos plantilla y DistCosta a partir de este antes de ajustarlos todos
template.rast <- trim(!is.na(app(rast(resampled_covariables.rastlist),
fun = sum)))
covariables.rastlist <- lapply(resampled_covariables.rastlist, function(rast)
trim(mask(crop(rast, template.rast), template.rast, maskvalues = FALSE)) )

names(covariables.rastlist) <-
c("Altitud", "AnomGrav", "BalMasa", "DistCosta", "VelHielo")
covariables.rast <- rast(covariables.rastlist)
```

II.2. Procesado del espesor de hielo

```
#> Importamos los archivos .csv a una lista y los nombramos por expedición
temp <- list.files(path = "./measurements/CReSIS/csv_files",
pattern = "\\.csv$", full.names = TRUE)
CReSIS.dflist <- lapply(temp, read.csv)
names(CReSIS.dflist) <- gsub("_Greenland|.csv", "", substring(temp, 40))

#> Reducir por porcentaje contribuído
reduced_CReSIS.dflist <-
CReSIS.dflist[which(prop.table(sapply(CReSIS.dflist, nrow)) > .024)]

#> Para cada expedición, transformamos las columnas que nos interesan a SpatVect
#> omitiendo NAs y  en el crs correcto antes de proyectar al que usaremos.
CReSIS.vectlist <- lapply(reduced_CReSIS.dflist, function(dataframe){
        project(vect(na.omit(dataframe[c("LON", "LAT", "THICK")]),
        geom = c("LON", "LAT"), crs = "EPSG:4326"), "EPSG:3413") })

    #> Quitamos valores atípicos
CReSIS_outliers <-performance::check_outliers(vect(CReSIS.vectlist)$Thick)
CReSIS_outliers # tiene 7 valores atípicos
split_index <- split(!CReSIS_outliers,
                     unlist(sapply(1:length(CReSIS.vectlist), function(k)
                       rep(k, length(CReSIS.vectlist[[k]]))))
CReSIS.vectlist <- lapply(
  setNames(1:length(CReSIS.vectlist), names(CReSIS.vectlist)), function(k)
    CReSIS.vectlist[[k]][split_index[[k]]])

#> Quitamos espesor <= 0
CReSIS_mod.vect <- vect(CReSIS.vectlist)[vect(CReSIS.vectlist)$THICK > 0,]

#> Los ajustamos a la rejilla (Altitud == template.rast)
CReSIS_mean.vect <- {
        rast <- rasterize(CReSIS_mod.vect, covariables.rast$Altitud,
        field = "THICK", fun = mean)
        vect <- as.points(mask(rast, covariables.rast$Altitud, maskvalues = NA))
        names(vect) <- "THICK"; vect
}
```

```
#> Creamos la matriz de observaciones
CReSIS_data.df <-
as.data.frame(extract(covariables.rast, CReSIS_mean.vect,
ID = FALSE, na.rm = TRUE, bind = TRUE, xy = TRUE))
names(CReSIS_data.df)[1] <- "Espesor"
```

II.3. Estimación de semivariogramas

```
#> Elección de los datos con los que seguir
data.df <- CReSIS_data.df

#> Distribución en conjuntos de datos de entrenamiento y comprobación
proportion <- c(train = .7, test = .3)
n <- 5e4; n/nrow(data.df)

{set.seed(321)
datasample <- data.df[sample(1:nrow(data.df), min(n,nrow(data.df)), replace=F),]
assignation <- sample(cut(seq(nrow(datasample)),
nrow(datasample)*cumsum(c(0, proportion)),
labels = names(proportion) ))
split_datasample <- split(datasample, assignation)
TrainingData.df <- split_datasample$train
TestingData.df <- split_datasample$test}

#> Hipótesis con performance
performance::check_model(lm(
formula= "Espesor ~ Altitud + AnomGrav + BalMasa + DistCosta + log(VelHielo)",
data = TrainingData.df), size_title = 24, base_size = 20)

##### Ordinary Kriging
formula = "Espesor ~ 1"; model = "OK"
psill <- NA; range <- 3e4; nugget <- 1e4; vmodel <- "Lin"; anis <- c(0, .8)

##### Universal Kriging
formula = "Espesor ~ Altitud + AnomGrav + BalMasa + DistCosta + log(VelHielo)"
model = "UK"
```

```
psill <- NA; range <- 1e5; nugget <- 1e4; vmodel <- "Exc"; anis <- c(0, 1)

cutoff <- max(TrainingData.df$DistCosta)*1e3/2
nbins <- 20

#> Semivariograma muestral
variogram <- variogram(as.formula(formula), ~x+y, data = TrainingData.df,
cutoff = cutoff, width = cutoff/nbins)

variogram_map <- variogram(gstat(id = "Mapa", formula = as.formula(formula),
locations = ~x+y, data = TrainingData.df),
map = T, cutoff = cutoff, width = cutoff/nbins)

#> Ajustar modelo teórico
theoretical_variogram <- fit.variogram(variogram, debug.level = 2, fit.method=2,
vgm(psill = psill, vmodel, range = range,
nugget = nugget, anis = anis))
anisotropy_variogram <- variogram(as.formula(formula), ~x+y,
data = TrainingData.df, cutoff = cutoff,
width = cutoff/nbins,
alpha = rev(seq(0, 180, by = 30))[-1])

assign(paste0(model, ".thvar"), theoretical_variogram)
```

II.4. Predicción de modelos

```
#> Construcción y predicción de modelos Kriging
nmin <- 1; nmax <- 100; maxdist <- 1e5
OK.model <- gstat(id = "OK", formula = as.formula("Espesor ~ 1"),
data = TrainingData.df, locations = ~x+y, model = OK.thvar,
nmin = nmin, nmax = nmax, maxdist = maxdist, set =list(gls=1))
OK.pred <- interpolate(covariables.rast, OK.model, cores = 4, cpkgs = "gstat",
debug.level = 2, na.rm = TRUE, index = 3:4)

nmin <- 1; nmax <- 100; maxdist <- 2e5
UK.model <- gstat(id = "UK", formula = as.formula(
```

```
"Espesor ~ Altitud + AnomGrav + BalMasa + DistCosta + log(VelHielo)"),
data = TrainingData.df, locations = ~x+y, model = UK.thvar, nmin = nmin,
nmax = nmax, maxdist = maxdist, set = list(gls = 1))
UK.pred <- interpolate(covariables.rast, UK.model, cores = 4, cpkgs = "gstat",
debug.level = 2, na.rm = TRUE, index = 3:4)

#> Construcción y predicción de otros modelos
MLR.model <- lm(formula = as.formula(
"Espesor ~ Altitud + AnomGrav + BalMasa + DistCosta + log(VelHielo)"),
data = TrainingData.df)
MLR.pred <- predict(covariables.rast, MLR.model, na.rm = TRUE)

IDW.model <- gstat(id = "IDW", formula = as.formula(
"Espesor ~ Altitud + AnomGrav + BalMasa + DistCosta + log(VelHielo)"),
data = TrainingData.df, locations = ~x+y, nmin = nmin, nmax = nmax,
maxdist = maxdist, set = list(gls = 1, idp = 2))
IDW.pred <- interpolate(covariables.rast, IDW.model, debug.level = 2,
na.rm = TRUE, index = 3:4)

bedmachine.rast <- rast("./measurements/BedMachine/BedMachineGreenland-v5.nc")
BM.pred <- trim(mask(
resample(bedmachine.rast$thickness, covariables.rast$Altitud,
method = "bilinear"), covariables.rast$Altitud, maskvalues = NA))
```

II.5. Comparación de métricas

```
#> Funciones de métricas
RMSE <- function(prediction, measurements = TestingData.df$Espesor){
        return(sqrt(mean((prediction - measurements)^2)))
}
R2 <- function(prediction, measurements = TestingData.df$Espesor){
        RSS <- sum((prediction - measurements)^2)
        TSS <- sum((measurements - mean(measurements))^2)

        return(1 - RSS/TSS)
}
```

```r
#> Función de volumen y área
SLEandArea <- function(thickness.rast){
  require(terra)

  thick_pos.rast <- mask(thickness.rast, thickness.rast > 0,
                         maskvalue = FALSE)

  area.km2 <- prod(res(thickness.rast)/1e3)
  extent.percent <- (area.km2 * global(!is.na(thick_pos.rast), sum)) /
    (area.km2 * global(!is.na(thickness.rast), sum)) * 100

  volume.km3 <- as.numeric(global(area.km2 * thick_pos.rast/1e3,
                                  sum, na.rm =T))
  print("Volume (km3):"); print(volume.km3)
  Gt <- volume.km3 * 0.9167 # 0.9167 density of ice
  print("Gt:"); print(Gt)

  SLE.m <- Gt * 1/361.8 # 361.8 Gt ice = 1 mm sea level elevation

  return(list("SLE" = SLE.m/1e3, "Extent" = extent.percent))
}

#> Evaluar todos los modelos
TestingData.vect <- vect(TestingData.df[c("Espesor", "x", "y")],
geom = c("x","y"), crs = crs(covariables.rast$Altitud))
models <- c("UK_197X.pred","MLR.pred","IDW.pred","OK.pred","UK.pred","BM.pred")
Test_results.df <- cbind(sapply(mget(models), function(rast)
extract(rast[[1]], TestingData.vect, ID = FALSE)[,1] ))

#> Calcular métricas
Metrics <- cbind(apply(Test_results.df, 2, RMSE), apply(Test_results.df, 2, R2))
colnames(Metrics) <- c("RMSE", "R2")
rownames(Metrics) <- gsub(".pred", "", models)
Metrics

#> Calcular volúmenes y áreas
Calcs <- sapply(models, function(model) unlist(SLEandArea(get(model)[[1]])))
```

```
rownames(Calcs) <- c("SLE", "Extent")
colnames(Calcs) <- gsub(".pred", "", models)
t(Calcs)
```

Bibliografía

Balmino, G. et al. (2011). "Spherical harmonic modeling to ultra-high degree of Bouguer and isostatic anomalies." *Journal of Geodesy.* DOI: 10.1007/s00190-011-0533-4.

Bárdossy, András (1997). *Introduction to geostatistics.*

BedMachine user guide (2024). Date Accessed 10-10-2024. URL: https://nsidc.org/sites/default/files/documents/user-guide/idbmg4-v005-userguide.pdf.

Bivand, Roger (2008). *Applied Spatial Data Analysis with R.*

Bonvalot, S. et al. (2012). "World Gravity Map WGM2012". *Bureau Gravimétrique International.* DOI: 10.18168/bgi.23.

Calculating glacier ice volumes and sea level equivalents (2025). Date Accessed 28-01-2025. URL: https://www.antarcticglaciers.org/wp-content/plugins/antarcticglaciers-pdf/download.php?p=7657.

CReSIS RDS Radar Guide (2024). Date Accessed 25-09-2024. URL: https://gitlab.com/openpolarradar/opr/-/wikis/Radar%20Guide-rds.

Hengl, Tomislav, Gerard Heuvelink y Alfred Stein (ago. de 2003). "Comparison of kriging with external drift and regression-kriging". *Technical Note.* URL: https://www.researchgate.net/publication/228961429_Comparison_of_kriging_with_external_drift_and_regression-kriging.

Hengl, Tomislav, Gerard B.M. Heuvelink y David G. Rossiter (2007). "About regression-kriging: From equations to case studies". *Computers & Geosciences* 33.10. Spatial Analysis, págs. 1301-1315. ISSN: 0098-3004. DOI: 10.1016/j.cageo.2007.05.001.

Howat, I., A. Negrete y B. Smith (2015). "MEaSUREs Greenland Ice Mapping Project (GIMP) Digital Elevation Model. (NSIDC-0645, Version 1)". *NASA National Snow and Ice Data Center Distributed Active Archive Center.* Date Accessed 10-07-2024. Dataset ID: NSIDC-0645. DOI: 10.5067/NV34YUIXLP9W. URL: https://nsidc.org/data/nsidc-0645/versions/1.

Howat, Ian, A. Negrete y B. Smith (dic. de 2014). "The Greenland Ice Mapping Project (GIMP) land classification and surface elevation datasets". *The Cryosphere* 8, págs. 1509-1518. DOI: 10.5194/tc-8-1509-2014.

Joughin, I., B. Smith e I. Howat (nov. de 2017). "A Complete Map of Greenland Ice Velocity Derived from Satellite Data Collected over 20 Years". *Journal of Glaciology* 64, págs. 1-11. DOI: 10.1017/jog.2017.73.

Joughin, I. et al. (2016). "MEaSUREs Multi-year Greenland Ice Sheet Velocity Mosaic, Version 1". Date Accessed 10-16-2024. DOI: 10.5067/QUA5Q9SVMSJG. URL: http://nsidc.org/data/NSIDC-0670/versions/1.

Karlsson, Nanna et al. (2024). "A Newly Digitised Ice-penetrating Radar Data Set Acquired over the Greenland Ice Sheet in 1971–1979". *Earth System Science Data* 16.7. DOI: 10.5194/essd-16-3333-2024.

Lüdecke, Daniel et al. (abr. de 2021). "performance: An R Package for Assessment, Comparison and Testing of Statistical Models". *The Journal of Open Source Software* 6, pág. 3139. DOI: 10.21105/joss.03139.

Matheron, G. (1969). *Le krigeage universel*. URL: https://cg.ensmp.fr/bibliotheque/public/MATHERON_Ouvrage_00131.pdf.

Morlighem, M. y et al (2022). "IceBridge BedMachine Greenland, Version 5". *NASA National Snow and Ice Data Center Distributed Active Archive Center*. Date Accessed 10-03-2024. DOI: https://doi.org/10.5067/GMEVBWFLWA7X.

Morlighem, Mathieu et al. (sep. de 2017). "BedMachine v3: Complete Bed Topography and Ocean Bathymetry Mapping of Greenland From Multibeam Echo Sounding Combined With Mass Conservation." *Geophysical Research Letters* 44. DOI: 10.1002/2017GL074954.

NASA's Operation IceBridge (2009-2019). URL: https://icebridge.gsfc.nasa.gov/.

Noël, B. et al. (2018). "Modelling the climate and surface mass balance of polar ice sheets using RACMO2, Part 1: Greenland (1958–2016)". *The Cryosphere* 12.3. DOI: 10.5194/tc-12-811-2018.

Oliver, M.A. y R. Webster (2014). "A tutorial guide to geostatistics: Computing and modelling variograms and kriging". *CATENA* 113, págs. 56-69. ISSN: 0341-8162. DOI: 10.1016/j.catena.2013.09.006.

Open Polar Radar (2024). "CReSIS Radar Depth Sounders Data, Greenland, years 1993 to 2017." *Digital Media*. We acknowledge the use of data and/or data products from CReSIS generated with support from the University of Kansas, NASA Operation IceBridge grant NNX16AH54G, NSF grants ACI-1443054, OPP-1739003, and IIS-1838230, Lilly Endowment Incorporated, and Indiana METACyt Initiative. URL: https://data.cresis.ku.edu/data/rds/csv_good/.

gstat user manual (2025). Date Accessed 22-01-2025. URL: https://www.gstat.org/gstat.pdf.

World Geodetic System (1984). URL: https://earth-info.nga.mil/index.php?dir=wgs84&action=wgs84.