Tarea 3 - MAT467

Alejandro Villazón G.

Todos los resultados presentados en adelante en este documento pueden ser replicados mediante los códigos desarrollados que se encuentre en el siguiente repositorio.
Considere el set de datos `anomalies` de la librería **GeoModels**. Además, considere que las coordenada s tiene una componente de latitud s_x y de longitud s_y, es decir $s = (s_x, s_y)$. La tarea puede ser resuelta considerando el software que usted encuentre más apropiado.

Parte 1: La primera parte se le pide considerar el ajuste para un modelo de la media **asumiendo independencia espacial**. Para esto, considere las siguientes propuestas de modelo:

(a) Una media cuadrática en las coordenadas, es decir, se tiene que

$$\mu_1(s) = \beta_0 + \beta_{1x}s_x + \beta_{1y}s_y + \beta_{2x}s_x^2 + \beta_{2y}s_y^2 + \beta_{xy}s_x s_y,$$

(b) Una media basada en promedios de los K_0 vecinos más cercanos, es decir, para $s_{(i)}$ el i-ésimo vecino más cercano de s se tiene que

$$\kappa(s) = \frac{1}{K_0}\sum_{i=1}^{K_0} X\left(s(i)\right),$$

$$\mu_2(s) = \beta_0 + \beta_1\kappa(s).$$

considere $K_0 = 10$.

(c) Seleccione, el mejor modelo vía K-fold CV, con $K = 10$.

Para seleccionar el mejor modelo consideraremos el que tenga menor promedio de RMSE en la validación cruzada.
Dada la independencia espacial, podemos estimar los parámetros β de ambos modelos de media μ_i, ajustando una regresión lineal simple, para esto utilizamos el paquete `statsmodels` de Python. Para ambos modelos de media, calculamos el RMSE entre la predicción y el verdadero valor en cada etapa de la validación cruzada. En la Tabla 1 se muestran los resultados promedio de la comparación, notamos que el modelo ganador es μ_2, es decir, utilizar la información de los K_0 vecinos más cercanos para cada punto de observación tiene menor error de predicción.

(d) Considere los residuos $r(s_i) = X(s_i) - \hat{\mu}(s_i)$, donde $\hat{\mu}$ es un modelo ajustado en la Parte 1. Explore la covarianza espacial usando el variograma.

Dado que μ_2 se posiciona como mejor modelo en el punto anterior, utilizamos este modelo para el cálculo de los residuos. Calculamos el variograma empírico y ajustamos los variogramas de diferentes familias de covarianza (Gaussian, Exponential, Matern, Rational,

Modelo	$\overline{\text{RMSE}}$
μ_1	0.841
μ_2	0.497

Table 1: Resultados validación cruzada.

Modelo	R^2
Matérn	0.98325
Exponential	0.97977
SuperSpherical	0.97363
Spherical	0.97108
Circular	0.96905
Rational	0.96746
Gaussian	0.93942
JBessel	0.93878

Table 2: Resultados validación cruzada.

Spherical, entre otros) estimando a la vez los parámetros del modelo.
Para seleccionar el modelo de covarianza analizamos el R^2 del ajuste, obteniendo los resultados presentados en la Tabla 2, resultando como ganador el modelo de covarianza Matérn con $\sigma^2 \approx 0.0766$, $\theta \approx 19.406$, $\nu \approx 0.237$ y $\tau^2 \approx 0.165$. El variograma empírico y el variograma del modelo ganador se presentan en la Figura 1, donde se observa un buen ajuste.

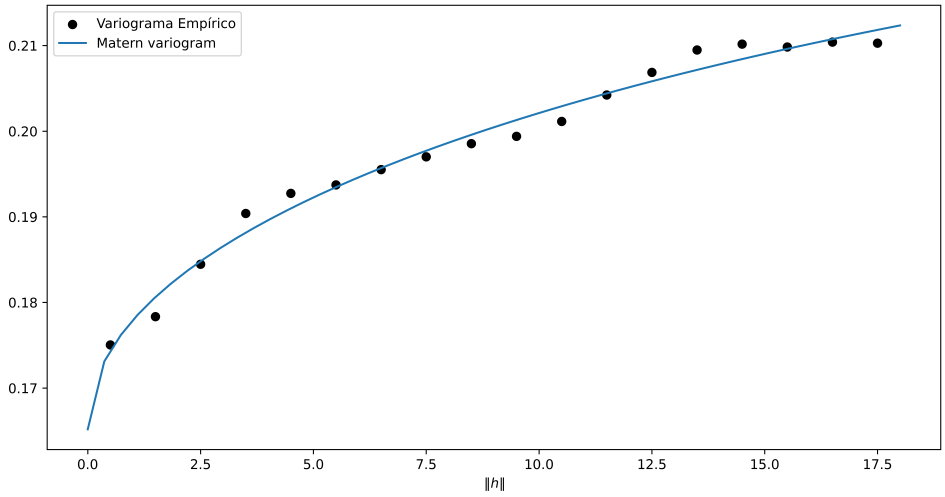

Figure 1: Variograma empírico y mejor modelo de covarianza.

Modelo	$\overline{\text{RMSE}}$
μ_1	0.546
μ_2	0.496

Table 3: Resultados validación cruzada con dependencia espacial.

Parte 2: En esta parte consideraremos la dependencia espacial. Para esto:

(a) Ajuste los modelos de media $\mu_1(s), \mu_2(s)$ considerando un modelos de covarianza (o variograma) a su elección.

Dado que el modelo Matérn se posiciona como mejor modelo de covarianza en la parte anterior, utilizaremos este modelo para ajustar los modelos de media.

(b) Seleccione, el mejor modelo vía K-fold CV, con $K = 10$.

Similar a la Parte 1, seleccionaremos el modelo minimizando el RMSE promedio en las etapas de la validación cruzada. Para ajustar los modelos y realizar las predicciones mediante Kriging Universal, haremos uso de las funciones del paquete `GSTools` de Python, las cuales en particular nos permiten entregar funciones de media personalizadas.

(c) ¿Cambian sus conclusiones?

Los resultados de la validación cruzada son presentados en la Tabla 3, donde observamos que considerando dependencia espacial μ_2 sigue siendo el mejor modelo de media. Además, podemos observar que el RMSE promedio disminuyó en el caso de μ_1 y el RMSE de μ_2 se mantuvo casi igual, por lo tanto, al considerar dependencia espacial se reduce la brecha de RMSE entre los modelos de media, pero no se logra disminuir en gran medida el error del mejor modelo de media.

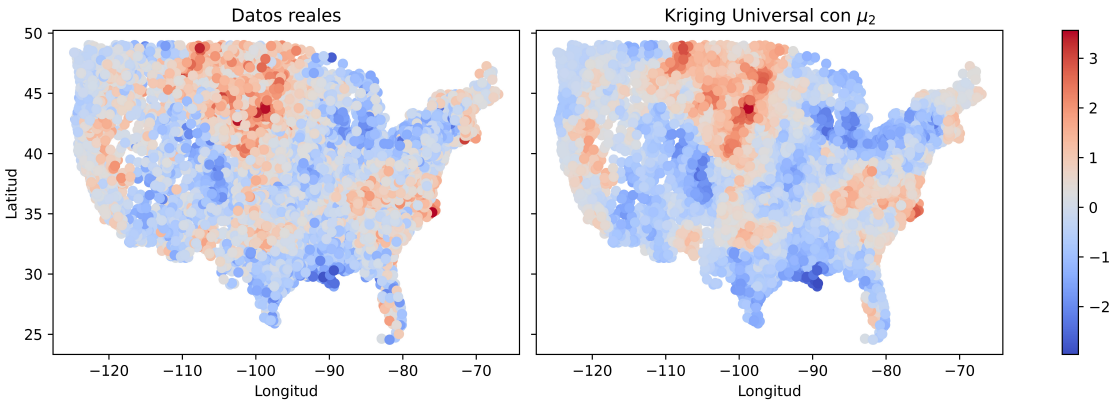

Figure 2: Datos vs Kriging Universal (μ_2).

(d) Realice una predicción espacial considerando Kriging Universal.

En la Figura 2 observamos los resultados de la predicción espacial utilizando Kriging Universal con el modelo μ_2. Observamos que al utilizar este modelo, la imagen de los datos se 'suaviza' y se logra capturar los valores extremos en las diferentes zonas, lo cual se puede apreciar por el pintado de las imágenes. Esto tiene sentido, pues estamos utilizando como media la información de los vecinos cercanos, que a priori deberían tener valores similares.