COMPARACIÓN DE MÉTODOS DE INTERPOLACIÓN ESPACIAL EN AGRICULTURA DE PRECISIÓN

Alemán, Alejandro

Facultad de Ciencias Agrarias de la Universidad Nacional del Litoral.

Director: Dr. Carlos Agustin Alesso

Codirector: Ing. Agr. (MSc) Ricardo J. M. Melchiori

Área: Ingeniería

Palabras claves: Agricultura de precisión, Interpolación espacial, geoestadística, kriging, random forest, idw.

INTRODUCCIÓN

El crecimiento de la población mundial demanda producir más alimentos sobre los espacios ya cultivados (Ray et al., 2012), . Tecnologías como la agricultura de precisión (AP) permiten utilizar de manera eficiente los insumos reduciendo al mínimo el impacto ambiental para , de manera de no afectar la producción a largo plazo (Radočaj et al., 2021).

La AP requiere de técnicas de interpolación que permitan mapear atributos de suelo a partir de muestreos de suelo georeferenciados o mediciones realizadas directamente a campo mediante sensores (Heuvelink & Webster, 2022). Por lo tanto, el estudio comparativo de las técnicas de interpolación aplicadas a los datos utilizados en AP puede ayudar a identificar los métodos más apropiados para maximizar la precisión de las estimaciones, la utilidad de la información generada y el tiempo de procesamiento.

Los métodos de interpolación espacial buscan estimar atributos en sitios no observados $z(x_0)$ a partir de la combinación lineal de observaciones vecinas $z(x_i)$ (Ec. 1):

$$\hat{z}(x_0) = \sum_{i=1}^{n} \lambda_i z(x_i) \quad (1)$$

donde λ_i es el peso asignado al punto muestreado y n es el número de observaciones que se utilizan para estimar (Webster & Oliver, 2007). Si bien existen varias técnicas de interpolación, las más empleadas en AP son el inverso de la distancia (IDW, del inglés inverse distance weighting) y kriging (Oliver, 2010).

Los últimos años se ha incrementado por el uso de algoritmos de aprendizaje automático o “machine learning” (ML) aplicados a la estadística espacial y mapeo de suelos digital (Hengl et al., 2018; Wadoux et al., 2020). Un algoritmo muy usado es “random forests” (RF) (Breiman, 2001) que permite representar relaciones complejas más allá de los patrones lineales particionando el espacio de la

Título del proyecto: Experimentación a campo en la era de la Agricultura Digital: diseño y análisis de experimentos a escala de lote. Instrumento: PEIC I+D
Año convocatoria: 2021
Organismo financiador: ASACTEI
Director/a: Alesso, Carlos Agustín

variable objetivo mediante árboles de decisión recursivos en el espacio de los predictores. Dado que RF ignora la localización espacial y por lo tanto la correlación espacial de los datos que no está en las covariables, Hengl et al. (2018), y posteriormente Sekulić et al. (2020) propusieron su representación mediante la creación de covariables en función de la distancias a los vecinos introduciendo el algoritmo Random Forest Spatial Interpolation (RFSI).

OBJETIVOS

- Comparar la performance general de los métodos de interpolación.
- Caracterizar los conjuntos de datos según fuente de datos (e.g. mapa rendimiento, etc), propiedades estadísticas (distribución) y características espaciales (densidad, arreglo y estructura espacial).
- Evaluar el efecto de estas características en la calidad de las predicciones de cada método.

METODOLOGÍA

Recopilación de datos

Se reunieron 125 conjuntos de datos de monitores de rendimiento (YM), mediciones de conductividad eléctrica aparente (Eca) del suelo a profundidades de 0-30 y 0-90 cm, Determinaciones de profundidad al horizonte petrocálcico (HPdepth), monitores de rendimiento de biomasa de maíz para ensilado (SYM) y elevación (Elv) de los lotes , los cuales se exploraron visualmente mediante el software QGIS para verificar la consistencia de las ubicaciones espaciales y de los atributos clave de los datos. Se digitalizaron los perímetros de cada lote para calcular la superficie de interés. Se identificaron los valores outliers dentro de la distribución de cada conjunto de datos para su posterior filtrado.

Se realizaron transformaciones de coordenadas geográficas a coordenadas planas del sistema UTM (Universal Transverse Mercator), utilizando el meridiano 21S para los lotes del sur de Buenos Aires y 20S para los lotes de Santa Fé. Finalmente, los datos se sistematizaron agrupándolos por tipo de datos.

Análisis estructural

Una vez sistematizados los datos, se procedió a realizar un análisis estructural para caracterizar:

- Densidad espacial de muestreo calculando la distancia promedio entre todos los puntos, la densidad de muestreo y la distancia promedio al vecino más cercano (NN por sus siglas en inglés ‘nearest neighbour’).
- Distribución estadística calculando el número de observaciones n, la media, la desviación estándar (sd), el coeficiente de variación (cv), la asimetría (asim) y la curtosis (kurt).
- Estructura de correlación espacial a diferentes distancias o lags, dividiendo las observaciones en cuatro intervalos de distancia "[0,25]", "(25,50]", "(50,75]", y "(75,100]" m.

Definición de hiperparámetros, entrenamiento y validación cruzada

Para cada set de datos se entrenaron modelos alternativos con diferentes valores de hiperparámetros siguiendo un grilla de hiperparámetros, y se seleccionó la combinación que resultó en menor error de predicción estimado mediante validación por k pliegues o folds cruzada. Esta metodología consistió en dividir los datos en k pliegues o folds para realizar k iteraciones, durante las cuales se entrenó un modelo utilizando k-1 folds, mientras que el fold restante se utilizó para evaluar el rendimiento del modelo obteniendo los residuos como la diferencia entre el valor observado y el predicho. De esta manera, se aseguró que cada predicción fuera comparada con una observación no utilizada en el entrenamiento del modelo evitando el sobreajuste.

A partir de los residuos de los modelos, se procedió a obtener las siguientes métricas para seleccionar el mejor set de hiperparámetros y explorar la relación de las propiedades (e.g. tipo de dato) de los datos obtenidos con los métodos de interpolación (Veronesi & Schillaci, 2019):

- La raíz del error cuadrático medio (RMSE) (Ec. 2).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2} \quad (2)$$

- El Error Absoluto Medio (MAE) (Ec. 3).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y_i}\right| \quad (3)$$

- El Coeficiente de Correlación de Concordancia (CCC) (Ec. 4).

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+\left(\mu_x-\mu_y\right)^2} \quad (4)$$

Relación performance modelos y tipos y características de los datos

Para cada métrica se ajustaron modelos lineales para evaluar el efecto del método de interpolación y el tipo de dato en la calidad de las predicciones, así como para el tiempo de cómputo requerido. También se modeló el efecto del método y su interacción con de las interacciones entre el método y los atributos de los datos (e.g asimetría, curtosis, etc.) para cada métrica (e.g. RMSE, MAE, CCC). Posteriormente, se realizó un análisis post-ANOVA utilizando el método de selección de variables stepwise Akaike (AIC) para identificar y seleccionar las variables predictoras significativas.

En todos los casos, los efectos se analizaron con la técnica de ANOVA con partición de cuadrados tipo III. En caso de que la interacción resultara significativa, se llevó a cabo un análisis post-ANOVA utilizando el test de Tukey para realizar comparaciones múltiples entre los distintos grupos.

CONCLUSIONES

La caracterización de los conjuntos de datos según la fuente de los datos (e.g. rendimiento en grano, electroconductividad, elevación, etc.) o según sus propiedades estadísticas y características espaciales, aportó información relevante en cuanto a la relación efecto global de dichas características y la calidad de las predicciones (Figura 1).

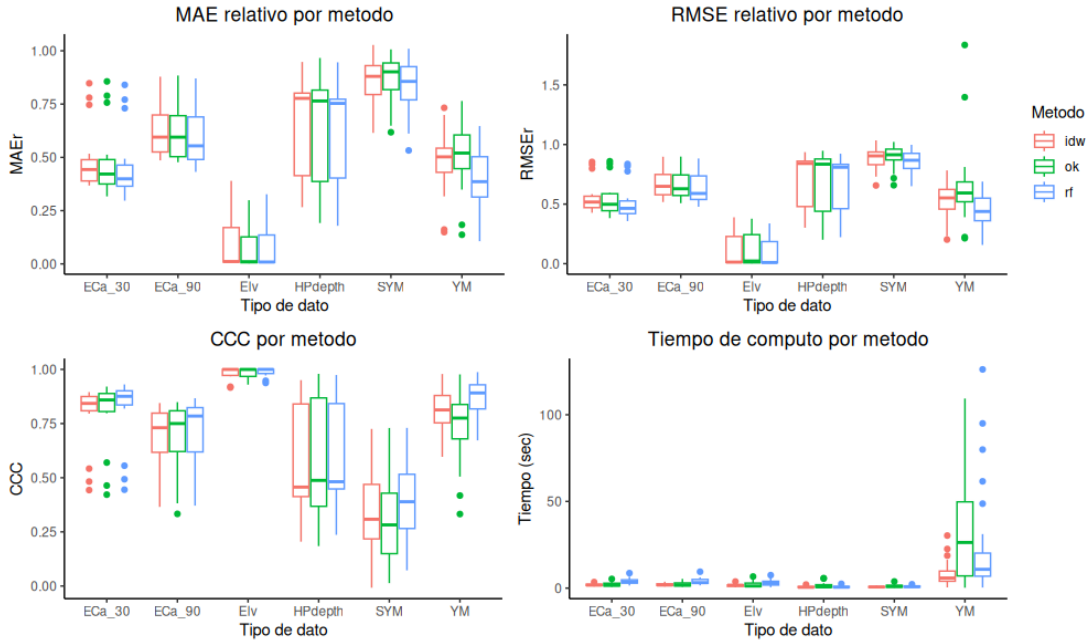

Figura 1. Distribución de los resultados de cada métrica por método y tipo de dato.

No obstante, al contrario de lo esperado, la interacción con los métodos de interpolación fue nula sugiriendo que estos factores no necesariamente pueden orientar la elección del método de interpolación.
Al comparar las metodologías de interpolación espacial para predecir atributos de suelo o cultivo en puntos no muestreados, la calidad de las predicciones está determinada por la estructura de los datos, por lo que a mayor estructura espacial hay menores errores de predicción independientemente del método utilizado. De esto se desprende que los métodos que incorporan estructura espacial no necesariamente requieren de una mayor densidad de muestreo. No obstante, el conocimiento de antemano de dicha estructura espacial no es tarea sencilla, requiriendo muestreos preliminares para diseñar los muestreos.
Por otro lado, tal como era esperado, se observó el efecto de los patrones de muestreo en la calidad de las predicciones, independientemente del método de interpolación escogido. Se recomienda entonces utilizar el método que mejor se ajuste a la capacidad de cómputo disponible de acuerdo a la estructura espacial de los datos en cuanto a la cantidad y distribución de los puntos o si se quiere incluir al análisis covariables que den cuenta de la estructura espacial de los datos o estimación de errores de predicción.

BIBLIOGRAFÍA BÁSICA

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.
Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ, 6, e5518.
Heuvelink, G. B. M., & Webster, R. (2022). Spatial statistics and soil mapping: A blossoming partnership under pressure. Spatial Statistics, 50, 100639.
Oliver, M. A. (2010). Geostatistical Applications for Precision Agriculture. Springer Netherlands.
Radočaj, D., Jurišić, M., Gašparović, M., Plaščak, I., & Antonić, O. (2021). Cropland Suitability Assessment Using Satellite-Based Biophysical Vegetation Properties and Machine Learning. Agronomy, 11(8), 1620.
Ray, D. K., Ramankutty, N., Mueller, N. D., West, P. C., & Foley, J. A. (2012). Recent patterns of crop yield growth and stagnation. Nature Communications, 3(1), 1293.
Sekulić, A., Kilibarda, M., Heuvelink, G. B. M., Nikolić, M., & Bajat, B. (2020). Random Forest Spatial Interpolation. Remote Sensing, 12(10), Article 10.
Veronesi, F., & Schillaci, C. (2019). Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation. Ecological Indicators, 101, 1032–1044.
Wadoux, A. M.-C., Minasny, B., & McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews, 210, 103359.
Webster, R., & Oliver, M. A. (2007). Geostatistics for environmental scientists (2nd ed). John Wiley & sons.