Desempeño predictivo de R-INLA SPDE para el Mapeo Digital de Suelos

Franca Giannini-Kurina[1], Franco Suarez[1], Pablo Paccioretti[1,2], Raúl Macchiavelli[3], Balzarini Mónica[1,2].

[1]CONICET-UFYMA Unidad de Fitopatología y Modelización Agrícola; [2]Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba; [3] Universidad de Puerto Rico, Mayagüez

Resumen. El mapeo digital de suelos (MDS) permite describir la variabilidad espacial de una propiedad edáfica a treves de modelos de predicción espacial que explican la relación que existe entre la variable de interés y covariables sitio-especificas. Entre los modelos estadísticos más incipientes en aplicaciones de MDS está la regresión bayesiana ajustada con INLA (del inglés, *Integrated Nested Laplace Approximation*) y SPDE (del inglés, *Stochastic Partial Differential Equation*) para modelar la correlación espacial entre sitios del dominio espacial a mapear. En este trabajo, se evaluó la implementación de la regresión Bayesianas (RB) se ilustró con tres bases de datos espaciales de características contrastantes. Los resultados de la implementación con RB se compararon con otros dos algoritmos ampliamente utilizados en el MDS, Regresión Kriging (RK) y Random Forest con residuos krigeados (RF). Se evaluó el desempeño predictivo de RB comparado con RK y RF según un diseño que propone por un lado variar la configuración de variables explicativas y por otro el número de observación entrenando el modelo. Todos los predictores espaciales fueron eficientes para el mapeo. Las mejores configuraciones de variables explicativas lograron resultados exitosos en términos de errores de predicción global (<25%). Las diferencias en el desempeño predictivo entre algoritmos de predicción espacial dependieron de particularidades de los escenarios de aplicación. El aumento en la cantidad de covariables implicadas en el modelo, es decir el número de parámetros a estimar tiene un impacto diferencial para RF, algoritmo que produce mejor rendimiento comparado con RB y RK en contextos de alta dimensionalidad. Finalmente se concluye que el desempeño estadístico de RB es competitivo frente a RK y RF. Futuras líneas de investigación deberían profundizar el estudio de propagación y dimensionamiento de la incertidumbre debido a las particularidades que RB frente a los otros métodos evaluados.

1 Introducción

El mapeo digital de suelos de ciertos atributos edáficos permite describir la variabilidad espacial de una variable en estudio a través de la predicción espacial (McBratney, Mendonça Santos y Minasny, 2003; Minasny y McBratney, 2016). De esta manera muchos atributos que antes se describían a través de una media general que caracterizaba una unidad cartográfica, hoy en día se pueden describir en un continuo. Las bases teóricas del mapeo digital de suelo radican en el esquema teórico

SCORPAN que sintetiza los factores formadores de suelo clásicos y las covariables necesarias para predecir características edáficas (Florinsky, 2012). Este enfoque establece que una propiedad de suelo en un sitio dado es función de otros datos de suelo que pueden ser previamente conocidos u obtenidos por muestreo de la variable de interés en otros sitios (S), de características climáticas (C), de la acción de los organismos vivos (O), características topográficas (R), del material parental o litología (P), del tiempo o evaluación temporal (A) y del dominio espacial (N). Este modelo también incluye los efectos de otras fuentes de variación no reconocidas a priori en el formato de una componente de error aleatorio aditivo. La predicción espacial de una variable continua implica una serie de desafíos metodológicos-estadísticos (Cressie y Wikle, 2015). Por un lado, se presentan los interrogantes inherentes a los modelos predictivos, es decir: el acondicionamiento de datos, la selección de modelos, el compromiso entre la bondad de ajuste y la capacidad predictiva y la medición de incertidumbre de las predicciones (Kuhn y Johnson, 2013). Por otro lado, hay desafíos particulares asociados la necesidad de describir la variabilidad de una variable aleatoria en el dominio continuo de dos dimensiones y su incertidumbre (Schabenberger y Gotway, 2005). Para esto, es necesario entender que las mediciones con las que se trabaja no son independientes y no solo se debe recurrir a herramientas que permitan trabajar con datos correlacionados, sino que se debe utilizar esta correlación para lograr describir en el dominio de las coordenadas geográficas una variable particular.

Los algoritmos alternativos para la predicción espacial provienen de diferentes enfoques estadísticos. Recientemente ha surgido una alternativa moderna para la predicción espacial que es la regresión bayesiana ajustada con INLA usando SPDE para modelar la correlación espacial (Lindgren y Rue, 2015; Krainski et al., 2018; Gómez-Rubio, 2020). El auge de implementaciones de esta técnica se ha dado en las ciencias ambientales por lo que su desempeño en el mapeo digital de resulta prometedor (Cameletti et al., 2013; Blangiardo y Cameletti, 2015; Huang et al., 2017). Bajo el paradigma bayesiano se considera que los parámetros del modelo de regresión son variables aleatorias y consecuentemente se asumen distribuciones de probabilidad asociadas a los parámetros. La información previa sobre la distribución de los parámetros se resume en distribuciones de probabilidad denominadas distribuciones *a priori*, a partir de las cuales se estima otra distribución de probabilidad, es decir se estima la distribución *a posteriori* de los parámetros dadas la distribución a priori y las observaciones. Calculando medidas resumen de la distribución a posteriori, como la media o el modo, se obtienen estimaciones puntuales de los parámetros que se informan juntos a intervalos de credibilidad calculados desde percentiles de la distribución a posteriori. La credibilidad de la estimación de un parámetro se interpreta como la probabilidad de que el valor estimado para el parámetro pertenezca al intervalo reportado dado los datos observados. En la predicción bayesiana la probabilidad describe grados de creencia (la creencia que tenemos de que una proposición sea verdadera), no frecuencias límite. Las distribuciones de probabilidad de los parámetros nos permiten realizar inferencia sobre el valor esperado del parámetro y la credibilidad asociada a esa estimación. Debido a que los valores predichos dependen del modelo con parámetros variables, los predichos también cuentan una distribución de probabilidad de los valores predichos a partir de las cuales se puede derivar no solo predicciones puntuales sino también medidas de incertidumbre para cada predicción lograda por el modelo (Correa Morales, Causil y Javier, 2018).

Una particularidad de INLA en la estimación de estructuras espaciales es resulta eficiente para modelar estructuras ralas, es decir con gran presencia de valores ceros, en la inversa de la matriz de variancias y covariancias (matriz de precisión). La estructura rala de la matriz de precisión se debe a la no dependencia de las variables aleatorias en la distribución multivariada conjunta (Havard Rue & Held, 2005). En R-INLA particularmente se logran matrices de precisión ralas utilizando aproximaciones por ecuaciones diferenciales parciales estocásticas (SPDE) (Lindgren, Rue y Lindström, 2011; Lindgren y Rue, 2015). Bajo este enfoque la grilla de predicción se abarca a través de una malla construida a partir de triángulos que cubren el dominio entero, cada vértice de los triángulos representa un nodo sobre los que se predice por interpolación (Blangiardo y Cameletti, 2015). Además de las ventajas computacionales que el algoritmo ofrece, permite trabajar con límites y bordes complejos (Bakka et al., 2018). En este trabajo se evalúan los modelos gaussianos latentes ajustados con INLA usando el método SPDE para modelar la correlación espacial en aplicaciones específicas del mapeo digital de suelos para variables continuas. Para esto, se comparó el desempeño estadístico en términos de predicción espacial de propiedades de suelo de modelos lineales para datos espaciales analizados con el paradigma bayesiano respecto al modelo lineal de covarianza residual para datos espaciales estimado con el enfoque frecuentista y al modelo de regresión basado en árboles de regresión.

2 Materiales y métodos

La implementación de la regresión Bayesianas (RB) se ilustró con tres bases de datos espaciales de características contrastantes. Se utilizó una base de datos desarrollada con el objetivo de mapear Materia Orgánica de Suelo de la provincia de Córdoba con más de 3000 observaciones y 25 variable explicativas (IDECOR, Piumetto et al., 2019); otra base de datos de referencia construida para mapear la concentración de metales pesados en suelo a orillas del Río Meuse en Holanda que cuenta con 150 observaciones y siete covariables explicativas (Burrough y McDonnell, 1998), por último, una base de datos también para la provincia de Córdoba utilizada para mapear la función de retención del herbicida atrazina en suelo a partir del índice de retención Kd (Giannini-Kurina et al., 2019).

Los resultados de la implementación con RB se compararon con otros dos algoritmos utilizados en la literatura moderna de MDS, Regresión Kriging (RK) (Hengl, Heuvelink y Rossiter, 2007) y Random Forest con residuos krigeados (RF) (Breiman, 2001; Li et al., 2011). Se evaluó el desempeño predictivo de RB comparado con RK y RF para las diferentes bases de datos de ilustración. Sobre la base de la hipótesis que el desempeño de la predicción espacial depende de otras particularidades de los escenarios de evaluación, como son el número de parámetros a estimar y el tamaño muestral, la evaluación se realizó según un diseño que propone por un lado variar la configuración de variables explicativas y por otro el número de observación entrenando el modelo. Para cada base de datos se configuró un diseño factorial de tres factores:1) Algoritmo de predicción espacial (niveles: RK, RF, RB); 2) Dimensionalidad o número de covariables (p). Se trabajó con todas las combinaciones posibles de covariables,

de 1 a 25 covariables para MOS, 1 a 7 para metales pesados y de 1 a 20 para Kd de atrazina. 3) Tamaño muestral (n) (con los niveles 3000, 1500, 500, 100 y 50 para MOS; 100, 80, 60, 40 y 20 para metales pesados y adsorción de atrazina). Para la evaluación de la capacidad predictiva en cada configuración se realizaron 30 repeticiones. El set de validación fue de tamaño constante en cada configuración y constó de un total de 50 muestras (no utilizadas para entrenar el algoritmo) seleccionadas al azar en un muestreo sin reposición. La medida de capacidad predictiva que se evaluaron sobre los grupos de validación fueron la raíz cuadrada del error cuadrático medio de predicción expresado relativo a la media general (EP (%)). Además, se informaron la proporción de aparición de variables explicativas en las 100 mejores configuraciones para cada modelo según el factor SCORPAN al que hacen referencia.

3 Resultados y discusión

El comportamiento de los diferentes algoritmos de predicción espacial en relación con el error de predicción promedio fue mejor al aumentar el tamaño muestral principalmente en la base de datos de materia orgánica del suelo donde los cambios en tamaño de muestra entre escenarios fueron mayores (Figura 1). Si se analizan las tendencias en el error global de predicción respecto a la cantidad de variables explicativas incluidas en la regresión (Figura 1) puede observarse que los errores obtenidos mediante el algoritmo RF fueron, en dos bases de datos, mayores que los obtenidos mediante BR y RK, en contextos de baja dimensionalidad (pocas covariables). Sin embargo, el incremento en el número de covariables hizo que el EP% del algoritmo RF fuese menor o igual que el EP% de los modelos RK o RB. Los resultados muestran que, en término de error de predicción global, existe un impacto positivo del aumento en el número de covariables (p) para el algoritmo RF que no se evidencia en los modelos RK y RB.

La implementación de regresión bayesiana para datos espaciales estimada por INLA utilizando SPDE para estimar la estructura de correlación espacial presenta algunas diferencias respecto a los dos métodos de regresión espacial más utilizados en mapeo digital de suelo. En lugar de establecer una red de vecindarios para definir la estructura espacial se utiliza una malla construida por triangulación. La estimación del efecto aleatorio espacial se realiza a partir de una función de Mátern que se resuelve por SPDE. Luego, la estimación de este efecto espacial se proyecta utilizando la malla sobre los sitios de predicción. Este mecanismo hace que no sea necesario calcular grandes matrices de distancias para la estimación de la correlación espacial como sí ocurre con los otros métodos, lo que implicaría una mayor eficiencia computacional. En este estudio, en términos computacionales, la conveniencia de RB respecto a RK y RL sólo fue evidente en el contexto del mapeo digital de MOS. El ajuste y la predicción para la provincia de Córdoba para los modelos RF y RL demoró 2.3 veces más que para RB. El proceso de cómputo para mapeo se realizó con un computador portátil de 8GB de RAM y un procesador Intel(R) Core(TM) m3-6Y30.

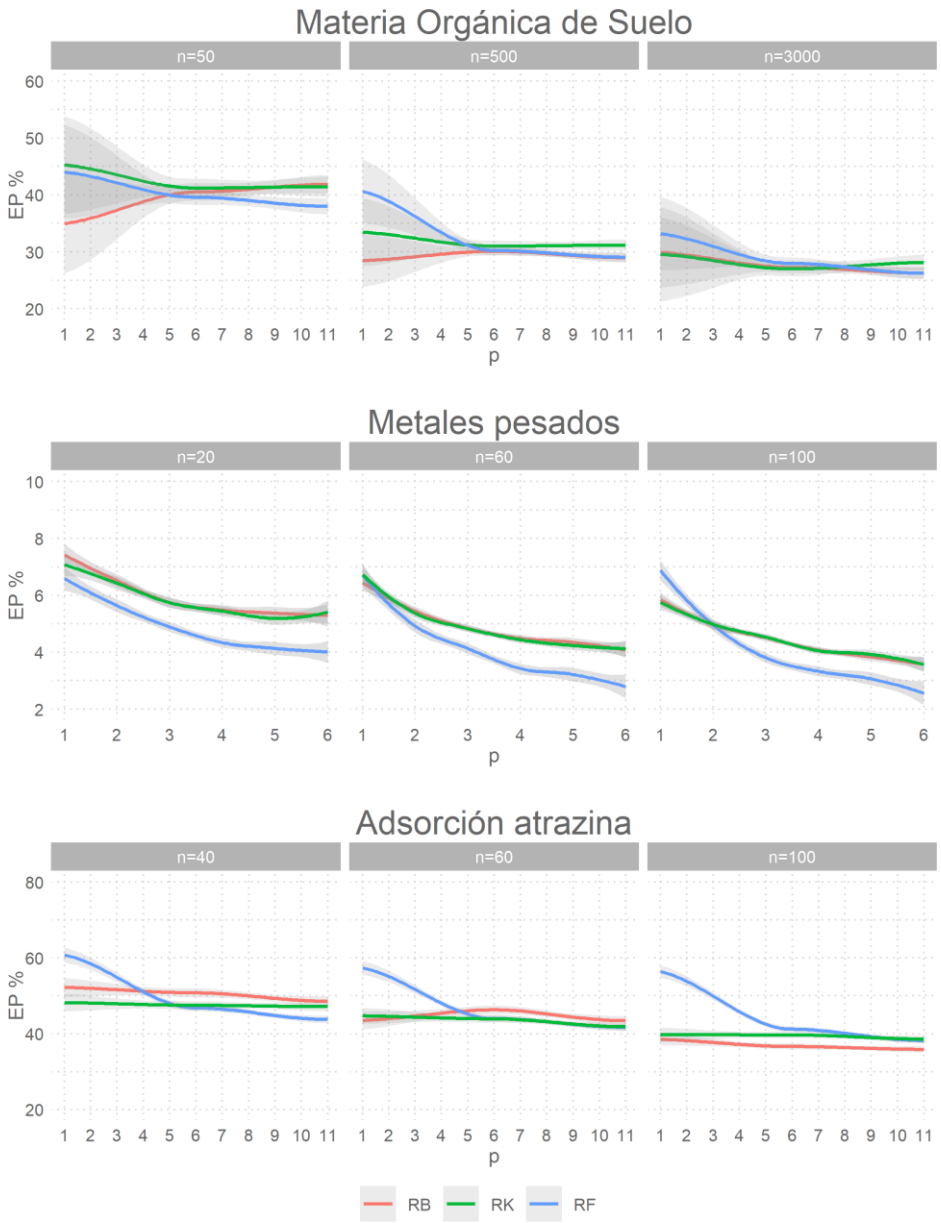

Fig. 1. Error de Predicción Global (EP%) expresado como porcentaje de la media predicha para cada sitio por tres algoritmos de predicción espacial: Regresión Bayesiana para datos espaciales (RB), Regresión Kriging (RK) y Random Forest con residuos krigeados (RF) frente a diferentes escenarios configurados según cantidad de variables explicativas (p) y tamaño muestral usado en la estimación (n).

Para la evaluación estadística de las diferencias en términos de EP (%) entre modelos se ajustó, para cada base de datos, un ANAVA clásico con efectos de: algoritmo, cantidad de covariables, tamaño muestral y las correspondientes interacciones dobles más la interacción triple. Las magnitudes en los errores de predicción de las tres bases de datos difirieron estadísticamente. Para todas las bases de datos, el impacto del incremento en n no interactuó con el Algoritmo, es decir los tres algoritmos responden de igual manera al incremento del tamaño muestral. No ocurre lo mismo respecto al aumento en el número de predictoras en donde el término de interacción entre el Algoritmo RF y p fue estadísticamente significativo en todas las bases de datos. De igual manera el efecto significativo de la interacción triple entre el Algoritmo RF, p y n en las tres bases de datos demuestra el impacto diferencial de este algoritmo ante aumentos en tamaño muestral número de variables predictoras. En cambio, la influencia de p y n en RB y RK fue la misma en las tres bases de datos.

Las mejores configuraciones (Tabla 1) dentro de todas las evaluadas lograron errores resultados exitosos en términos de errores de predicción global (<25%). Siendo el ln(Kda) la variable con mayor error de predicción, seguida por MOS y por último el ln(Zinc). Las diferencias en el desempeño predictivo entre los casos de estudio se corresponden con la variabilidad que presenta la variable respuesta, a mayor variabilidad de la variable respuesta mayor EP%. Finalmente, los desempeños las mejores configuraciones obtenidos para cada caso de estudio presentaron pocas diferencias. Los factores SCORPAN preponderantes en los procesos que determinaron la variabilidad espacial de cada variable modelada fueron contundentes en todas las configuraciones.

Tabla 1. Error de Predicción de las mejores configuraciones de variables explicativas para tres algoritmos de predicción espacial Regresión Bayesiana para datos espaciales (RB), Regresión Kriging (RK) y Random Forest con residuos krigeados (RF).

Base de datos	Algoritmo	EP	*p*	Factor SCORPAN *			
				S	C	O	R
Materia Orgánica de Suelo n=3000	RB	16.4	11	1	0.99	0.71	1
	RK	17	11	1	0.96	0.78	1
	RF	15.1	16	1	0.99	0.7	1
Metales pesados n=100	RB	2.31	4	1	-	0.62	0.89
	RK	2.24	2	1	-	0.54	0.84
	RF	1.4	6	1	-	0.66	0.81
Adsorción Atrazina n=100	RB	20.4	6	0.99	0.8	-	0.43
	RK	20.4	6	1	0.97	-	0.5
	RF	23	11	1	0.85	-	0.44

*En cada una de las configuraciones se evaluó la presencia de una o más variables explicativas clasificadas según el factor SCORPAN al que pertenecen (suelo (S), clima (C), acción de los organismos vivos (O) y relieve (R)). La proporción de participación se calcula como cantidad de configuraciones donde el factor estuvo presente sobre un total de 100 configuraciones las cuales presentaron los mejores desempeños predictivos.

A modo de ejemplo se presenta en la Figura 2 el mapeo de la distribución espacial predicha de ln(Kd) y las medidas de incertidumbre para cada algoritmo. La combinación de variables utilizadas en la implementación de cada método correspondió a aquellas configuraciones que minimizaron el EP %. Si bien las predicciones resultan similares para los tres algoritmos, RF genera predicciones menos variables con CV de los valores predichos entre 10% y 20% mayores que RB y RK. Resulta discutible la comparación de las medidas de incertidumbre informadas, ya que tanto la estimación como los supuestos en cada enfoque son diferentes. No obstante, se puede observar que los desvíos estándares de las predicciones con RK son mayores a RB y RF. Los algoritmos RK y RF suponen que las estimaciones de la parte fija del modelo son correctas ya que se informa el DE obtenido a partir de la varianza kriging estimada sobre los residuos del modelo (Hengl, Heuvelink y Rossiter, 2007). RB en cambio informa la desviación estándar de la distribución a posteriori predicha para cada sitio cuya variabilidad depende de la incertidumbre asociada a las distribuciones marginales de parámetros e hiperparámetros marginales que dan origen a la distribución a posteriori (Gelman, 2004).

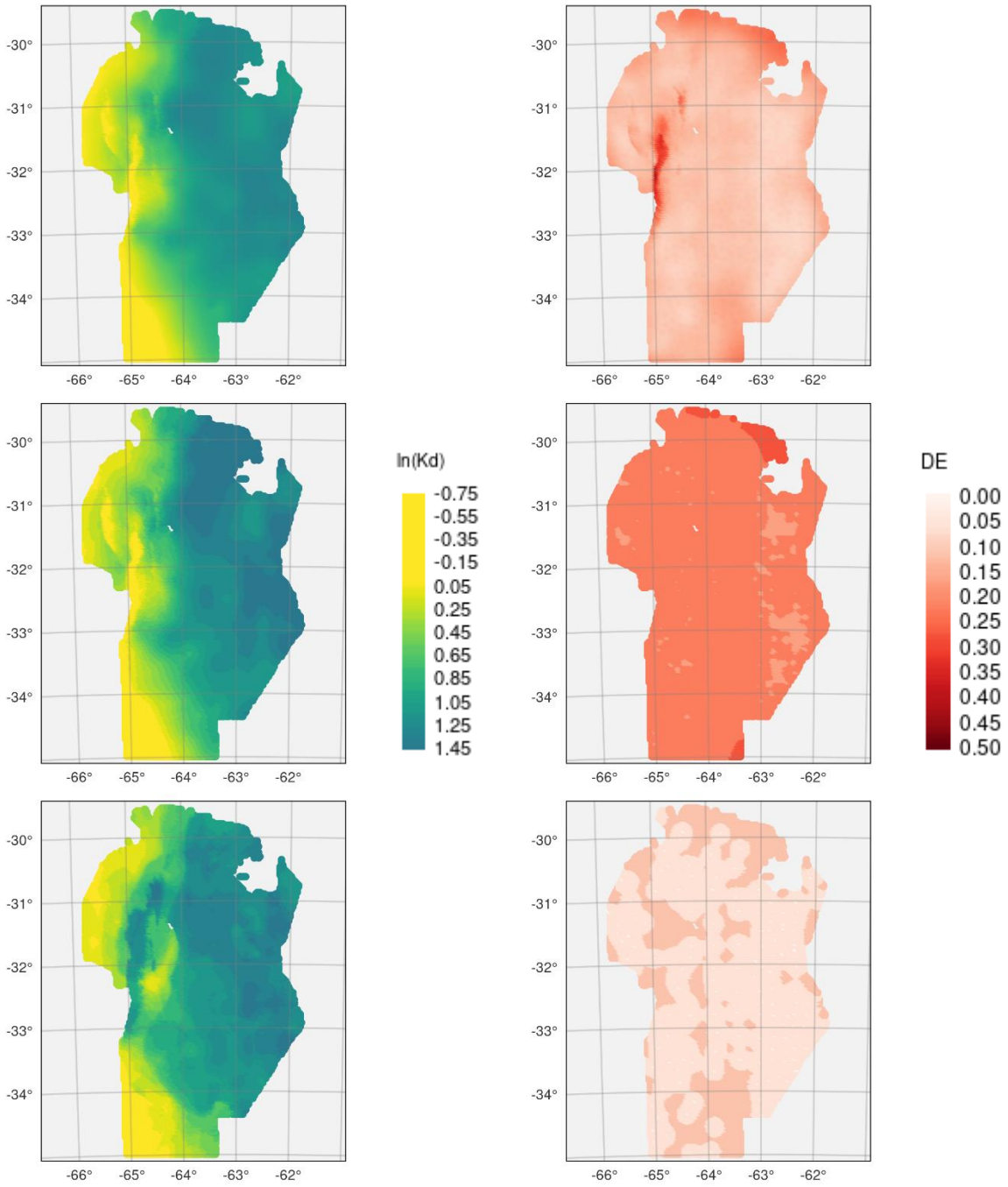

Fig. 2. Mapeo digital de ln(Kd) en base a tres algoritmos de predicción espacial. De arriba hacia abajo Regresión bayesiana, Regresión Kriging y Random Forest. En la columna de la derecha se muestran las predicciones puntuales y en la columna de la derecha las medidas de incertidumbre de la predicción.

4 Comentarios finales

El desempeño predictivo de cada algoritmo depende de particularidades de los escenarios a los cuales se aplica. Aumentos en el tamaño muestral implican mayor precisión en la predicción, este comportamiento es constante entre los algoritmos y las bases de datos aquí evaluadas. No ocurre lo mismo con el número de parámetros a estimar. En este sentido el aumento en el número de variables predictoras tiene un impacto diferencial positivo para en el algoritmo basado en arboles de regresión y clasificación, obteniéndose mejores desempeños comparados con los otros algoritmos en contextos de mayor dimensionalidad. Los resultados confirman que el desempeño estadístico en términos de predicción espacial de propiedades de suelo de modelos lineales para datos espaciales analizados con el paradigma bayesiano es competitivo frente al modelo lineal de covarianza residual para datos espaciales estimado con el enfoque frecuentista y al modelo de regresión basado en árboles de regresión. En trabajos futuros se debe profundizar el estudio de propagación y dimensionamiento de la incertidumbre.

Referencias

Blangiardo, M. y Cameletti, M. (2015) *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Breiman, L. (2001) Random forests. *Machine learning*, **45**, 5–32.

Burrough, P.A. y McDonnell, R.A. (1998) *Principles of Geographical Information Systems*. Oxford University Press.

Cameletti, M., Lindgren, F., Simpson, D. y Rue, H. (2013) Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*, **97**, 109–131.

Cressie, N. y Wikle, C.K. (2015) *Statistics for spatio-temporal data*. John Wiley & Sons.

Giannini-Kurina, F., Balzarini, M., Rampoldi, E.A. y Hang, S. (2019) Site-specific data on herbicide soil retention and ancillary environmental variables. *Data in Brief*, **27**, 27–30.

Gómez-Rubio, V. (2020) *Bayesian inference with INLA*. CRC Press.

Hengl, T., Heuvelink, G.B.M. y Rossiter, D.G. (2007) About regression-kriging: From equations to case studies. *Computers and Geosciences*, **33**, 1301–1315.

Huang, J., Malone, B.P., Minasny, B., McBratney, A.B. y Triantafilis, J. (2017) Evaluating a Bayesian modelling approach (INLA-SPDE) for environmental mapping. *Science of the Total Environment*, **609**, 621–632.

Krainski, E.T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., et al. (2018) *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. Chapman and Hall/CRC.

Kuhn, M. y Johnson, K. (2013) *Applied predictive modeling*. Springer.

Li, J., Heap, A.D., Potter, A. y Daniell, J.J. (2011) Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling & Software*, **26**, 1647–1659.

Lindgren, F. y Rue, H. (2015) Bayesian Spatial Modelling with *R* - **INLA**. *Journal of Statistical Software*, **63**.

McBratney, A.B., Mendonça Santos, M.L. y Minasny, B. (2003) *On digital soil mapping*.

Minasny, B. y McBratney, A.B. (2016) Digital soil mapping: A brief history and some lessons.

Geoderma, **264**.

Piumetto, M., Córdoba, M., Morales, H., Fuentes, L., Álvarez, P., Carranza, J.P., et al. (2019) *Mapeo de Materia Orgánica del Suelo*. Infraestructura de Datos Espaciales Córdoba, IDECOR.https://idecor.cba.gov.ar/wp-content/uploads/2020/12/Informe-Final-MO_1.pdf.

Schabenberger, O. y Gotway, C.A. (2005) *Statistical methods for spatial data analysis*. CRC press.