Universidad Nacional de Córdoba

Maestría en Estadística Aplicada

Facultad de Ciencias Agropecuarias

Facultad de Matemática, Astronomía, Física y Computación

Facultad de Ciencias Económicas

MODELOS BAYESIANOS PARA DATOS GEOESTADÍSTICOS. MAPEO DIGITAL DE SUELOS CON R-INLA

Tesista: Dra. Ing. Agr. Giannini Kurina, Franca

Director: PhD. MSc. Macchiavelli, Raúl E.

Codirectora: PhD. MSc. Balzarini, Mónica

Tesis para optar al título de Magister en Estadística Aplicada

Córdoba, 2021

MODELOS BAYESIANOS PARA DATOS GEOESTADÍSTICOS: MAPEO DIGITAL DE SUELOS CON R-INLA

Aspirante

Franca Giannini Kurina

Comisión Asesora de Tesis

Director: PhD. Raúl E., Macchiavelli

Codirectora: PhD. Mónica G. Balzarini

Tribunal

Dr. Fernando García

Dra. Analía Becker

Dr. Mariano A. Córdoba

Presentación formal académica

3 de agosto de 2021

Universidad Nacional de Córdoba

Resumen

El mapeo digital de suelos (MDS) permite describir la variabilidad espacial de una propiedad edáfica. Utiliza 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 se encuentra 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 abordaron los fundamentos estadísticos para la modelación de datos geoestadísticos en general y la modelación espacial a través de la inferencia bayesiana utilizando INLA y SPDE, en particular. 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). Finalmente 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 utilizadas para entrenar 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%). No obstante, la implementación de RB presenta algunas diferencias respecto a los otros métodos. La predicción sitio específica corresponde a una medida resumen de posición de la distribución conjunta a posteriori predicha en cada sitio. De la misma distribución de densidad se obtienen las medidas de incertidumbre de cada predicción. Estas particularidades posicionan a la RB como una buena alternativa comparada a los otros métodos evaluados en la cuantificación de la incertidumbre de los mapas creados. 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. El desempeño estadístico de RB es competitivo frente a RK y RF. Futuras líneas de investigación deberían profundizar el estudio de propagación de la incertidumbre y explorar el desempeño de RB en el mapeo de datos no normales.

PALABRAS CLAVE: Modelos predictivos, Modelos de regresión, Datos espaciales, Modelos jerárquicos Bayesianos, INLA-SPDE.

Abstract

Digital soil mapping (DSM) describes edaphic properties spatial variability. It uses spatial predictive models that explain the relationship between the variable of interest and site-specific covariates. Among the most incipient statistical methods in DSM applications is the Bayesian regression fitted with INLA (Integrated Nested Laplace Approximation) and using SPDE (Stochastic Partial Differential Equation) to model the spatial correlation between sites in the spatial domain. In this thesis, we described the statistical concepts on geostatistical data modeling in general, and particularly, spatial modeling through Bayesian inference using INLA and SPDE. The implementation of Bayesian regression (BR) was illustrated with three spatial databases with diverse characteristics. The results of the implementation with BR were compared with two most used algorithms DSM, Regression Kriging (RK) and Random Forest with residual krigeados (RF). Finally, the predictive performance of BR compared with RK and RF was evaluated according to a design that varies the number explanatory variables and the number of observations training the model. All spatial predictors were efficient for mapping. The best explanatory variable configurations achieved successful results in terms of global prediction errors (<25%). However, the BR implementation presents some differences respect to other methods. The site-specific prediction corresponds to summary measure of the joint posterior distribution predicted at each site. The uncertainty measures for each prediction are obtained from the same density distribution. These particularities position the BR as a good alternative compared to the other methods evaluated in the quantification of the uncertainty of the maps generates. Differences in predictive performance between spatial prediction algorithms depend on particularities of the application. The increase in the number of covariates involved in the model, that is, the number of parameters to be estimated, has a differential impact on RF, presenting better performance compared to BR and RK in high-dimensional contexts. BR's statistical performance is competitive against RK and RF. Further research should deepen the study of uncertainty propagation and explore the performance of BR in mapping non-normal data.

KEYWORDS: Predictive models, Spatial regression models, Hierarchical Bayesian models, INLA-SPDE.

Índice

Índice de Tablas

Índice de Figuras

Abreviaciones

AIC, criterio de información de Akaike;
AVE, promedio de varianza explicada;
BIC, criterio de información Bayesiano;
DE, desviación estándar;
EP, Error de Predicción;
ESE, Error Sitio Específico;
GPS, sistemas de posición geográfica;
IM, índice de autocorrelación espacial local de Moran;
LR, regresión lineal frecuentista;
MDS, Mapeo Digital de Suelos
ML, máxima verosimilitud;
MLM, modelos lineales mixtos;
MOS, Materia Orgánica del Suelo;
NDVI, del inglés *Normalized Difference Vegetation Index;*
INLA, del inglés *Integrated Nested Laplace Approximation.*
REML, máxima verosimilitud restringida;
RF, del inglés, *Random Forest*;
SIG, sistemas de información geográfica;
UTM, universal transversal de Mercator.

Capítulo I

Introducción General

Mapeo digital de suelos

El mapeo de suelos es una herramienta fundamental para la caracterización ambiental y la gestión de recursos naturales. Históricamente los mapas de suelo se han construido a partir de la identificación de unidades fisiográficas caracterizadas a partir de una serie de propiedades fisicoquímicas determinadas mediante muestreos de suelo a distintas profundidades. En las últimas décadas ha surgido el mapeo digital de suelos (MDS) (McBratney, Mendonça Santos y Minasny, 2003) como alternativa al mapeo tradicional para optimizar costos de obtención de mapas de numerosas propiedades de suelo. El mapeo digital hace uso de diferentes fuentes de información relacionada a la variable a mapear y de modelos estadísticos que permiten ir desde bases de datos construidas con una limitada cantidad de muestras de suelo a la predicción espacial del atributo de interés en el continuo. Los datos que alimentan estos análisis se conocen como datos espaciales, de tipo geoestadístico, ya que no solo contiene información del valor de la variable registrada, sino que también acompañan ese valor con coordenadas espaciales que lo sitúan en un dominio espacial continuo. Así, la georreferencia o posicionamiento del sitio del cual se obtuvo el dato enriquece el valor de la variable en el sitio ya que permite relacionar éste con su entorno. En el mapeo digital de suelo se supone que la variabilidad espacial no se encuentra determinada solo por la distribución en el espacio de la variable de interés, sino por otras múltiples covariables que también varían espacialmente y se relacionan con la propiedad analizada.

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). La siguiente ecuación resume las bases de este enfoque:

$$S = f(S, C, O, R, P, A, N) + \epsilon \qquad \textbf{[1]}$$

La ecuación establece que una propiedad de suelo S en un sitio dado es función $f(.)$ 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). El modelo también incluye los efectos de otras fuentes de variación no reconocidas a priori en el formato de una componente de error aleatorio aditivo(ϵ). Los efectos de cada uno de los términos de la función $f(.)$ se pueden estimar a través de información georreferenciada proveniente de distintas infraestructuras de datos espaciales, hoy ampliamente disponibles. Bajo este enfoque adquieren especial relevancia las predicciones espaciales de variables aleatorias asociadas a propiedades edáficas de interés (Milne, 2009). La predicción espacial de las variables de interés se puede realizar a distintas escalas espaciales, pero usualmente se trabaja a escalas regionales y hasta globales, al punto que existe disponible un mapeo, de atributos de suelos, para todo el planeta (Arrouays et al., 2014; Hengl et al., 2017; Yigini et al., 2018)

Numerosos mapeos de propiedades edáficas se realizan a partir de predicciones basadas en algoritmos de aprendizaje automático (Hengl et al., 2017) que si bien permiten relacionar la información auxiliar disponible para generar el output de interés de manera cuasi-automática, pueden presentar altos errores de predicción en sitos donde no hay abundante información del entorno de la predicción, por lo que aún es preciso trabajar sobre modelos estadísticos que permitan predicciones más eficientes a partir de datos muestrales.

La modelación estadística de una variable aleatoria depende al menos de dos componentes. Por un lado, un componente estructural, explicado por factores y covariables que se suponen conocidas y determinan la estructura de media o valor esperado de la variable respuesta. Por el otro, un componente asociado con la variabilidad y covariabilidad de los datos de la variable respuesta. Esta variabilidad, puede al menos en parte encontrarse espacialmente estructurada, es decir responder a un proceso estocástico de autocorrelación espacial. En el caso de una variable respuesta con distribución normal estos componentes, de la estructura de medias y de la estructura de varianza-covarianzas, se relacionan de manera aditiva. El modelo lineal puede expresarse como la suma de un componente estructural o determinístico más un componente aleatorio que se divide en un componente explicado por el proceso espacial subyacente y un término remanente o netamente residual.

Existen alternativas metodológicas de diferente naturaleza que con distinta efectividad bajo distintos escenarios de datos espaciales permiten modelar tanto la componente determinística como la componente aleatoria. Las predicciones espaciales son el resultado del ajuste de un modelo para el valor esperado de la respuesta al que se incorporan funciones que contempla la dependencia o autocorrelación entre las observaciones, que es dependiente de la distancia geográfica que separa los sitios desde los cuales se obtuvieron los datos. La construcción de un modelo predictivo permitirá predecir el valor de la variable de interés en sitios no muestreados. Luego, de obtener buenas predicciones del atributo en distintos sitios, habrá suficientes datos espaciales para producir mapas de variabilidad espacial del atributo de interés en todo el dominio espacial. El mapeo entonces no es más que la representación en el plano de los sitios de predicción utilizando herramientas cartográficas para representar las tres dimensiones del análisis Latitud, Longitud y atributo, donde generalmente se opta por denotar la tercera dimensión (atributo) a través de una escala de colores acorde.

Modelos predictivos para datos espaciales

Kuhn & Johnson (2013) establecen que la modelación predictiva es el proceso por el cual se construye una herramienta matemática o un modelo que es capaz de generar una predicción. Este proceso no implica no solo dilucidar los patrones subyacentes en los datos bajo estudio. También, se requiere hacer uso del criterio experto, dar un correcto tratamiento a los datos, identificar los predictores influyentes, validar correctamente el modelo y por último seleccionar las alternativas de mayor capacidad predictiva.

Luego de obtener los datos, como en todo análisis estadístico, se debe recurrir inicialmente a una instancia de preprocesamiento de datos que incluye la depuración y acondicionamiento a los formatos necesarios para el estudio relacional. En el caso de datos espaciales en el preprocesamiento hay que considerar los diferentes formatos de la información espacial (*raster, vectorial*) y las particularidades del sistema de referencia debido a los diferentes sistemas de proyección geográfica. Los Sistemas de Información Geográfica (SIG) (Warner y Diab, 2002) ofrecen un amplio rango de funciones para crear, integrar, transformar, visualizar y analizar de manera exploratoria datos espaciales (Longley et al., 2005). Los SIG más avanzados también disponen de funciones que generan interfase con software estadístico ampliando así la capacidad para la modelación y análisis más complejos (Bivand, 2014; Lovelace, Nowosad y Muenchow, 2019). Numerosas herramientas desarrolladas sobre el software libre R reúnen las rutinas necesarias para

el análisis estadístico espacial y manejo de datos georreferenciados (Bivand, Pebesma y Gómez-Rubio, 2013).

Las bases de datos asociadas a los SIG son usualmente multivariadas (i.e. múltiples capas de variables distribuidas regionalmente). Particularmente, los sensores remotos satelitales (Mulder et al., 2011) tanto ópticos como de radar, permiten el acceso a grandes volúmenes de información espacial sobre la superficie terrestre. No obstante, otras fuentes de datos espaciales como las derivadas de modelos digitales de elevación (Moravec et al., 2017), los provenientes de sensores proximales (Viscarra Rossel et al., 2011) y las obtenidas por muestreos georreferenciados pueden acoplarse a estos SIG (Margules y Pressey, 2000).

La etapa de selección de las variables para ajustar un modelo de regresión es otra etapa crucial en el ajuste de modelos predictivos. Las variables disponibles con potencialidad explicativa suelen estar altamente correlacionadas (colinealidad) y complicar el ajuste de modelos lineales (Dormann et al., 2013). Esta instancia en modelos de regresión espaciales y espaciotemporales adquiere especial importancia debido a los sesgos que pueden producirse en la estimación de los coeficientes de regresión de los modelos lineales debido al "confundimiento espacial" (Page et al., 2017) o dificultad en disociar los efectos de las variables predictoras respecto a los efectos aleatorios espaciales. Las técnicas de selección de variables que funcionan bien en el contexto de modelos lineales no necesariamente arrojan los mismos resultados que otras técnicas de selección con capacidad de captar estructuras no lineales como las derivadas de métodos de inteligencia artificial o aprendizaje de máquina (Elith et al., 2006; Kanevski, Timonin y Pozdnukhov, 2009a).

Luego de la depuración de datos atípicos y la selección criteriosa de variables basada en el conocimiento del problema, prosigue la etapa de ajuste del modelo predictivo. El concepto de autocorrelación espacial es crucial para el tratamiento de los datos georreferenciados: éste describe la correlación de los valores de una variable entre distintos sitios desde donde se releva en relación con la distancia que existe entre éstos (Matérn, 1986). Una vez que se ha modelado la función espacial se recurre a ésta para generar interpolaciones entre valores de la variable para generar datos en sitios donde el atributo no ha sido muestreado (Cressie, 1990; Webster y Oliver, 2007). La interpolación espacial de una variable en el continuo del dominio espacial donde ha sido observada permite obtener mapas de variabilidad espacial.

Cuando existe información auxiliar que puede ser usada como variables explicativas o covariables de la variable de interés las predicciones de ésta en sitios que no han sido muestreados pueden derivarse de modelos predictivos del tipo de modelos de regresión. Desde la estadística espacial se han realizado numerosos aportes para el ajuste de regresiones con datos espaciales tanto desde una perspectiva frecuentista (Cressie, 1990; Webster y Oliver, 2007) como desde la teoría Bayesiana (Blangiardo y Cameletti, 2015; Lindgren y Rue, 2015; Krainski et al., 2018) y desde un enfoque más computacional como son los algoritmos de aprendizaje automático (Li et al., 2011; Hengl et al., 2017).

Una vez que el modelo predictivo se pone en marcha es necesario evaluar las predicciones. Esta etapa debe ser planificada para evitar que el proceso de validación sea inadecuado. Predicción implica asignar nuevos valores de las variables respuesta u observaciones independientes a las que se utilizaron para ajustar el modelo, por lo cual estimar el error de predicción es solo posible a través de la comparación de los valores estimados por un modelo y los valores observados en sitios no utilizados en la construcción del modelo.

Entre las alternativas para estimar la precisión de la predicción existen distintos formatos de la técnica de validación cruzada o técnicas de partición del conjunto de datos en datos de calibración y datos de validación (Efron y Hastie, 2016). En este contexto es necesario identificar un grupo de observaciones sobre las que se ajustará el modelo, usualmente llamado grupo de entrenamiento, y otro grupo que se usará para validar, llamado grupo de validación. No todas las formas de muestrear para particionar el conjunto de datos para la validación ni todos los criterios de resumen del proceso de validación producen los mismos resultados (Golub, Heath y Wahba, 1979; Efron y Tibshirani, 1997; Brenning, 2012). En general cuanto menor es la proporción de datos que contiene el grupo de validación menor capacidad de extrapolación tendrá el modelo, siempre existe un compromiso entre la información disponible para el ajuste y para la validación (Kuhn y Johnson, 2013).

Cuando se modelan variables continuas una métrica usual para validar los modelos obtenidos es la Raiz del Error Cuadrático Medio de Predicción (RMSPE por el término en inglés *"Root Mean Square Prediction Error"*) (Willmott, 1981). En primer lugar, se calculan los errores de predicción como la diferencia entre los valores predichos por el modelo en el grupo de validación y los valores observados. Entonces, el RMSPE se calcula tomando la raíz cuadrada de la media de estos errores elevados al cuadrado. Suele interpretarse como la distancia promedio de los valores observados y

las predicciones del modelo y muchas veces es conveniente expresarla como una medida relativa o porcentual a la media de la población. La partición de la base de datos en grupo de entrenamiento y de validación suele realizarse de manera aleatoria y repetida, de esta manera es posible calcular no sólo el RMSPE esperado para el ajuste sino también la varianza del RMSPE. Sobre la base de estas medidas puede realizarse un análisis de sensibilidad (Hamby, 1994).

Otras métricas usadas para evaluar la bondad del ajuste de un modelo de regresión, como el R^2 o coeficiente de determinación y los criterios de información basados en verosimilitud penalizada, suelen ser usadas para evaluar modelos predictivos ya que un modelo bueno como herramienta predictiva, debería representar también un buen ajuste de los datos de entrenamiento. Sin embargo, existen riesgos de modelos de buen ajuste, pero sobreparametrizados que pueden derivar en una mala capacidad predictiva o complejizar inútilmente la interpretación.

Para el caso de datos espaciales, también es necesario evaluar del error de predicción referido al sitio específico en el que se hizo la predicción, también conocido como *error puntual de la predicción*. Es importante resaltar que siempre debe existir un compromiso entre el ajuste, la predicción y la interpretación a la hora de evaluar un modelo predictivo. No solo se debe prestar atención al desempeño predictivo del valor promedio o esperanza en cada sitio sino al dimensionamiento de la incertidumbre asociada a la predicción. Trabajos recientes enfatizan que las diferencias entre métodos alternativos para la predicción espacial radican en el dimensionamiento de la incertidumbre y no en la medición del valor promedio de predicción (Giannini-Kurina et al., 2019; Koo et al., 2020).

Estadística espacial

La modelación de procesos espaciales se desarrolla a través de la teoría de campos aleatorios (Besag, 1981; Cressie, 1990). Un campo aleatorio en el dominio espacial se suele modelar como la suma de un componente estructural o determinístico (tendencia o media), un proceso aleatorio de autocorrelación espacial y un proceso aleatorio de errores independientes (Burrough y McDonnell, 1998). El proceso de autocorrelación espacial refiere a la correlación de un mismo atributo entre dos sitios, si ésta es positiva implica que dos puntos cercanos tienen valores similares. Es decir, los valores entre dos sitios son dependientes del modelo estocástico subyacente. Se asume comúnmente que los procesos de autocorrelación espacial son estacionarios de segundo orden,

esto implica que existe una media constante y una función de covariancia que depende únicamente de la distancia entre observaciones.

Una clasificación de datos espaciales según la definición del dominio espacial subyacente es la propuesta por Cressie, 1993. Así los datos espaciales se clasifican en datos geoestadísticos, datos regionales o de área y patrones de puntos. Los datos del tipo geoestadístico emergen de mediciones sobre un dominio espacial continuo y supone que entre dos sitios existen infinitos datos. El concepto se refiere a continuidad en la estructura espacial del proceso que genera los datos, los datos pueden ser del tipo continuo o discreto indistintamente. Los datos regionales se asocian a un dominio espacial discreto haciendo referencia a cantidades por superficie o área (Schabenberger y Gotway, 2005). En la tercera categoría se definen los patrones de puntos donde los datos se presentan en lugares ubicados aleatoriamente. El dominio es aleatorio y discreto y su definición donde el interés está enfocado en describir el patrón de los agrupamientos (los agrupamientos corresponden a distintos intervalos definidos para el dominio) (Plant, 2012).

Modelación de datos espaciales

Si consideramos al valor $z(x)$ como la realización de una variable aleatoria en un sitio x, un modelo estadístico para Z se puede construir desde tres componentes:

$$z(x) = \eta(x) + f(x) + \epsilon(x) \qquad \textbf{[2]}$$

donde el primer componente $\eta(x)$, es un predictor lineal asociado a efectos fijos de variables explicativas, esos efectos están parametrizados, son desconocidos y deben ser estimados. El segundo componente $f(x)$ representa la estructura de dependencia espacial entre las observaciones. El tercer componente $\epsilon(x)$ corresponde al término del error. Un modelo matemático flexible y conveniente para $f(x)$ es el modelo Gaussiano de media cero, en el cual la función de covarianza que cuantifica la magnitud de dependencia entre dos sitios se puede expresar de la siguiente manera:

$$c(x_i, x_j) = Cov\{(x_i), (x_j)\} \qquad \textbf{[3]}$$

dados los sitios $\{x_1, \dots, x_n\}$ el valor que toma el proceso gaussiano (GP) en estos sitios $\{f(x_1), \dots f(x_n)\}$ corresponde a una distribución normal multivariada de media cero y matriz de varianzas y covarianzas de $(i,j)^{th}$entradas definidas por $c(x_i, x_j)$.

Esta formulación de modelo estadístico se puede extender a variables respuesta no normales utilizando una función de enlace $g(x)$ dada una distribución específica y media:

$$g^{-1}(\eta(x) + f(x)) \quad \textbf{[4]}$$

Estimación del modelo espacial

Métodos paramétricos basados en semivariogramas

La teoría asociada a datos geoestadísticos establece que la observación de la variable aleatoria en cada sitio es una realización de un proceso aleatorio continuo, caracterizado por una distribución de probabilidad. El proceso puede ser estacionario o no estacionario. La estacionariedad es una propiedad que surge de la repetición de un proceso estocástico en un dominio espacial y define cómo varía en función de las coordenadas. Schabenberger y Gotway, (2005) definen la estacionariedad como una propiedad que refleja la falta de importancia de las coordenadas.

El primer tipo o grado de estacionariedad es la estacionariedad en sentido estricto implica que las coordenadas no influyen en absoluto en el proceso aleatorio. El segundo tipo es lo que se denomina estacionariedad de segundo orden que implica que la media del proceso es constante a lo largo del dominio.

$$E[Z(s)] = \mu \quad \textbf{[5]}$$

pero que la covarianza entre diferentes observaciones depende de la distancia h entre las mismas, es decir:

$$Cov[Z(s), Z(s+h)] = C(h) \quad \textbf{[6]}$$

Entonces queda definida la función de covarianzas $C(h)$ que deriva en varios aspectos interesantes para la modelación espacial. La función de covarianza en un proceso estacionario de segundo orden no depende de las coordenadas absolutas y la varianza es constante en todo el proceso, es decir:

$$Cov[Z(s), Z(s+0)] = Var[Z(s)] = C(0) \quad \textbf{[7]}$$

Por lo tanto, este proceso estacionario de Segundo orden es equivalente a un proceso estocástico básico donde las observaciones que pertenecen a una misma población tienen la misma media y varianza, con la diferencia que en este caso sí están correlacionadas.

Un proceso estacionario de segundo orden solo permite funciones de covarianzas que describan una autocorrelación en función a la distancia del tipo positiva es decir que las similaridades o correlaciones entre observaciones cercanas sean mayores que las similaridades o correlaciones entre observaciones alejadas.

Una herramienta clásica para modelar autocorrelación espacial positiva en un proceso estacionario de segundo orden es el semivariograma o función de semivarianza. Para estimar la semivarianza deben agruparse los sitios de acuerdo con una medida denominada *lag,* que especifica la distancia entre sitios. El tamaño de muestra con el que se estima cada semivarianza y la distribución de los sitios en el plano determinará el conjunto de *lags* a evaluar. Para su identificación se construye a partir de las observaciones lo que se conoce como semivariograma empírico denotado por la siguiente ecuación.

$$\hat{\gamma}(h) = \frac{1}{2m(h)} \sum_{i=1}^{m(h)} [Y(x_i) - Y(x_i + h)]^2 \quad \textbf{[8]}$$

donde h es el lag o distancia entre los sitios x_i y $x_i + h$ y , $m(h)$ es el número de observaciones contenidas en el lag h.

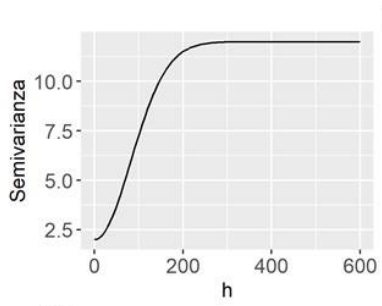
La función del semivariograma se caracteriza por tres parámetros: la varianza nugget, la varianza estructural y el rango. La varianza *nugget* refiere a la ordenada al origen del semivariograma y representa la suma de errores aleatorios no espaciales, o errores asociados con la variabilidad espacial a escalas más finas de las medidas en la grilla de datos espaciales con la que se esté trabajando (Schlesinger et al., 1996). La varianza estructural representa la varianza de observaciones independientes, es decir observaciones que fueron tomadas a una distancia tal que la autocorrelación debida a la dependencia espacial es cero. El rango es la distancia a la cual se alcanza el máximo de la función de semivarianza, es decir la distancia a la cual dos observaciones pasan a ser independientes. El rango es también entendido como la extensión del proceso de autocorrelación espacial. Cuando las funciones que describen la semivarianza no tienen un máximo absoluto a menuda se utiliza lo que se llama un rango práctico el cual representa la distancia a la cual se obtiene el 95% de la semivarianza estructural.

Para resumir el grado de estructura espacial de una variable georreferenciada en un dominio continuo se puede usar la varianza estructural relativa (RSV por el término en inglés "Relative Structural Variance"), definida como el cociente entre la varianza estructural (C) y la suma de la varianza nugget (C_0) y la varianza estructural.

$$RSV = \left(\frac{C}{C + C_0}\right) \times 100\% \qquad \textbf{[9]}$$

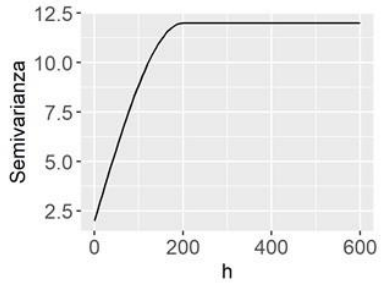
Se entiende que el RSV debe ser lo suficientemente alto para obtener una predicción espacial eficiente. Zimback (2001), establece que el grado de dependencia espacial puede ser clasificado como: $RSV \leq 25\ \%$ bajo, entre $25\ \% < RSV < 75\ \%$ medio y $RSV \geq 75\ \%$ alto.

Luego de generado el semivariograma empírico a partir de los datos, es posible ajustar un modelo de semivariograma teórico. Los modelos téoricos corresponden a una función definida positiva ya que la varianza de los datos puede tomar solo valores positivos no nulos. Entre las funciones más usadas como semivariograma teórico se encuentran funciones no lineales como el semivariograma gaussiano y el exponencial (Figura 1).



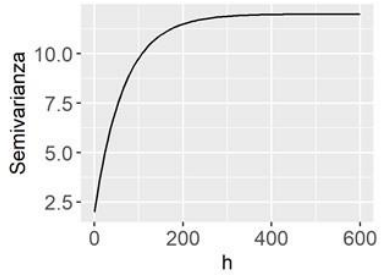
Modelo Exponencial

$$\gamma(h) = \begin{cases} C_0 & h=0 \\ C_0 + C\left\{1 - exp\left\{\frac{-3h}{R}\right\}\right\} & h \neq 0 \end{cases}$$



Modelo Esférico

$$\gamma(h) = \begin{cases} C_0 & h=0 \\ C_0 + C\left\{\frac{3}{2}\frac{h}{R} - \frac{1}{2}\left(\frac{h}{R}\right)^3\right\} & 0<h \leq R \\ C_0 + C & h > R \end{cases}$$



Modelo Gausiano

$$\gamma(h) = \begin{cases} C_0 & h=0 \\ C_0 + C\left\{1 - exp\left\{-3\left(\frac{h}{R}\right)^2\right\}\right\} & h \neq 0 \end{cases}$$

Función

Figura 1: Funciones de semivariaograma para el modelo exponencial, esférico y gaussiano. $\boldsymbol{C_0}$=2, $\boldsymbol{C}$=10 y $\boldsymbol{R}$=200.

Las funciones de semivariograma Gaussiano y Exponencial derivan de una clase flexible de funciones provenientes de una representación espectral de la función de covarianza de un proceso isotrópico llamadas funciones Matérn (Matérn, 1986). Esta familia de funciones posee una serie de propiedades convenientes para el ajuste del semivariograma empírico (Schabenberger y Gotway, 2005). Otra función utilizada para modelar semivariogramas es el modelo Esférico (Isaaks y Srivastava, 1989) que deriva de una familia de funciones indicadoras donde el valor indicador corresponde al diámetro de una esfera equivalente al rango del proceso espacial (Figura 1).

Los modelos descriptos son modelos isotrópicos es decir la función de covarianza es constante en todas las direcciones. Un proceso anisotrópico implicaría que la función o los parámetros de la función de covarianza variarán en el espacio. Para contemplar dentro de estos modelos de dependencia la presencia de estructuras espaciales que no sean isotrópicas se puede recurrir a una transformación lineal del sistema de coordenadas cuando se trate de lo que se conoce como

anisotropía geométrica. Otro tipo de proceso anisotrópico es conocido como anisotropía zonal, que puede modelarse adicionando un término dependiente de la dirección del *lag* (Goovaerts, 1999).

La estimación de los semivariogramas paramétricos se realiza frecuentemente por el método de los mínimos cuadrados ponderados (Schabenberger & Gotway, 2005). Luego de elegida la función para el modelo teórico de dependencia y asignar valores iniciales a sus parámetros (basados en la observación del semivariograma empírico), se itera hasta encontrar los estimadores de mejor ajuste en el sentido de menor error cuadrático medio. Otras alternativas de estimación son máxima verosimilitud y máxima verosimilitud restringida (Diggle, 2013). Estudios comparativos donde se han estimado los mismos procesos espaciales con distintos métodos sugieren que la estimación de los parámetros para caracterizar la variabilidad espacial es menos precisa y menos eficiente con los métodos geoestadísticos clásicos que con modelos lineales mixtos (MLM) presentando la estimación máximo verosímil mejor desempeño que la estimación máximo verosímil residual (Gili, 2013).

Ajustado el modelo de semivariograma, será posible usarlo para predicción espacial, es decir la predecir valores de la variable en sitios del dominio espacial continuo donde no existen observaciones. Para la predicción espacial a partir de semivariogramas se usan distintos tipos del método de interpolación geoestadística *kriging* (Cressie, 1990). El método kriging ordinario proporciona para cada sitio de la grilla de predicción el mejor estimador lineal insesgado. Es lineal porque el valor estimado se obtiene como un ponderado de los datos disponibles, es no sesgado porque la estimación del error medio residual es cero y se establece que es el mejor porque minimiza la varianza de los errores de estimación del modelo. Además, proporciona un error de estimación conocido como varianza de kriging, que depende del modelo de semivariograma ajustado y de la ubicación en el espacio de los datos originales.

Las interpolaciones geoestadísticas tienen ciertas ventajas respecto a otros tipos de interpolaciones espaciales que pueden aplicarse. Por un lado, las distancias sobre las que trabaja son distancias estadísticas en contraste con las distancias geométricas utilizadas en métodos no estadísticos, es decir contemplan mecanismos para aplicarse en proceso estocástico. Además, evita muestras redundantes, ponderando de forma distinta muestras que estén muy cerca entre sí y procedan de la misma región respecto a muestras que estén, por ejemplo, en lados opuestos al sitio sobre el que se quiere asignar la predicción (Webster y Oliver, 2007). Los parámetros del semivariograma son los que gobiernan la asignación de los pesos o ponderaciones que se dará a las observaciones que

rodean el sitio al cual se le asignará la predicción. Particularmente, el parámetro nugget determinará cómo se reparten estos pesos, si la varianza del error es muy alta, todas las muestras tenderán a tener el mismo peso en la interpolación. Por el contrario, si la varianza del error es pequeña, los coeficientes de ponderación serán muy distintos entre las diferentes muestras, muchas de las cuales por encontrarse cercanas en el espacio serán redundantes. El rango del semivariograma determinará cuándo el semivariograma se vuelve horizontal. Si el rango aumenta, cada punto tendrá mayor peso en la interpolación.

Los modelos lineales (Raudenbush y Bryk, 2002; West et al., 2014), su extensión desde los modelos generalizados (McCullagh y Nelder, 1989; Agresti, 2015) y los modelos aditivos generalizados (Wood, 2006), estimados por mínimos cuadrados o por métodos basados en la verosimilitud de las observaciones, son modelos estadísticos que se usan con datos espaciales para mejorar la interpolación espacial y la capacidad de predicción. En el modelo lineal se asume una relación lineal entre las covariables y la variable respuesta estructurada espacialmente. Consecuentemente en la parte aleatoria no se trata a los errores como independientes, sino que se supone algún modelo para el proceso que determina las autocorrelaciones espaciales. Las funciones de semivariograma se ajustan a partir de los residuos del modelo lineal que especifica vía la suma de efectos fijos el valor esperado de la respuesta. La estimación del modelo completo (estructura de media y de varianzas y covarianzas) puede hacerse a partir de mínimos cuadrados, máxima verosimilitud o máxima verosimilitud restringida (Gili, 2013). En el marco de modelos generalizados la estimación se hará por pseudo verosimilitud o cuasi verosimilitud restringida (Stroup, 2016).

Métodos no paramétricos y semi paramétricos

Los modelos aditivos generalizados (GAM por el término en inglés “Generalized Additive Models”) (Hastie y Tibshirani, 1990) son modelos generalizados que involucran una suma de funciones de suavizado de las covariables. Permiten especificaciones flexibles de la relación entre la variable respuesta y las covariables por medio de funciones de suavizado en lugar de establecer una parametrización detallada del proceso espacial. Para la estimación de un GAM se recurre a métodos de regresión penalizada luego de estimar el grado de suavizado a partir de los datos mediante validación cruzada o verosimilitud (Hastie y Tibshirani, 1990). Las técnicas no paramétricas y semiparamétricas de suavizado de tendencias son especialmente útiles en el contexto de datos espaciales (Higgins, 2003; Bosq, 2012). La dependencia espacial se aproxima mediante la técnica de suavizado de manera numérica mediante la maximización de algún criterio de optimalidad, como

mínimos cuadrados, condicionando el ajuste a parámetros del suavizado o de penalización, entendiendo que en cuanto menor penalidad se otorgue al criterio de optimalidad, menos suavizada será la curva de la tendencia espacial estimada. Un tipo de suavizados utilizado corresponde a los modelos de base *splines* (Wood, 2006). Estos modelos se representan como funciones polinómicas, que usan polinomios de grado *p* para ajustes locales. Cada polinomio describe un segmento de línea o una superficie planar, y se ajusta simultáneamente con el resto de los segmentos para generar un suavizado continuo. Los valores de cada segmento coinciden en lo que se denominan nodos del suavizado (Webster & Oliver, 2007). Los suavizados *splines* (Green y Silverman, 1993) utilizan todas las observaciones y estiman parámetros relacionados para lograr el suavizado, hecho que reduce la eficiencia de su implementación cuando el volumen de datos es grande. Los *splines* basados en regresión pueden ser ajustados simplemente mediante mínimos cuadrados una vez que se ha seleccionado el número de nodos. La introducción de penalizaciones en estos tipos de *splines* relaja la importancia de la elección de la cantidad y localización de los nodos o segmentación de los datos (Durbán, Lee y Ugarte, 2008).

Paradigmas para la estimación de regresiones con dependencia espacial

Enfoque Frecuentista

En la estadística frecuentista los parámetros del modelo a ajustar se consideran constantes fijas que son desconocidas y deben ser estimadas a partir de los datos. Dado estos parámetros se asignan probabilidades a las observaciones aleatorias. La probabilidad asignada a la inferencia de los parámetros refiere a una frecuencia relativa (asumiendo un tamaño grande de muestra) y se evalúa en un número infinito de repeticiones hipotéticas del proceso generador de datos.

Cuando los datos de la variable respuesta se estructuran en el espacio, i.e. con autocorrelación espacial, los modelos de regresión frecuentistas pueden estimarse desde el marco teórico de los MLM (Molenberghs, Verbeke y Demétrio, 2007) que además de asumir una relación lineal entre la variable respuesta y las covariables, asume la presencia de autocorrelación espacial en los términos de error del modelo como causante de la correlación espacial entre los datos (Raudenbush y Bryk, 2002; West et al., 2014). En estos MLM, también llamados modelos de covarianza residual, no se trata a los errores como independientes, sino que se supone algún modelo (Figura1) para el proceso que determina las autocorrelaciones espaciales en la componente aleatoria (Balzarini, Macchiavelli y Casanoves, 2004). Para variables dependientes con distribución normal, la estimación del modelo completo (efectos fijos o coeficientes de regresión y varianzas y covarianzas de los errores) puede

hacerse a partir de mínimos cuadrados ponderados, máxima verosimilitud o máxima verosimilitud restringida o residual (Schabenberger y Gotway, 2005).

Enfoque Bayesiano

La construcción de un modelo predictivo para datos espaciales también puede realizarse desde un paradigma bayesiano donde 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, i.e. 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. Por lo tanto, el paradigma bayesiano se basa en los siguientes postulados: 1) la probabilidad describe grados de creencia (la creencia que tenemos de que una proposición sea verdadera), no frecuencias límite. Como tal se pueden realizar afirmaciones probabilísticas acerca de distintas cosas y no solo datos sujetos a variabilidad aleatoria, consecuentemente podemos hacer afirmaciones probabilísticas sobre los parámetros, 2) las distribuciones de probabilidad de los parámetros nos permiten realizar inferencia sobre el valor esperado del parámetro y la credibilidad asociada a esa estimación, y 3) como los valores predichos dependen del modelo con parámetros variables, los predichos también cuentan una distribución de probabilidad. Las distribuciones de los valores predichos permiten derivar no solo predicciones puntuales sino también medidas de incertidumbre para cada predicción lograda por el modelo (Correa Morales, Causil y Javier, 2018).

La estadística bayesiana ha tenido un gran desarrollo en los últimos años debida a la mejora de los algoritmos computacionales que permiten estimar de manera rápida las distribuciones *a posteriori*. Un aspecto dificultoso en la estimación bayesiana se basa en cómo definir la distribución *a priori* encargada de resumir el conocimiento previo. Aunque este último aspecto puede solucionarse desde la práctica usando distribuciones *a priori* bastante generales o poco informativas como pueden ser distribuciones normales con varianzas muy grandes.

Los métodos de simulación por cadenas de Markov Monte Carlo (MCMC) (Besag et al., 1995), han permitido resolver modelos complejos sin la necesidad de imponer estructuras que lo simplifiquen (Schabenberger y Gotway, 2005). Por ello, han sido usados para la estimación de modelos con datos espaciales (Best, Richardson y Thomson, 2005; Reich y Fuentes, 2007). El método MCMC conlleva alta demanda computacional. Rue et al (2009) propusieron una alternativa para aproximar la distribución a posteriori en contextos de datos espaciales a partir de aproximaciones basadas en el algoritmo INLA (por término en inglés "Integrated Nested Laplace Approximation") lo que ha permitido simplificar las estimaciones (Wang, Ryan y Faraway, 2018).

Una particularidad de INLA en la estimación de estructuras espaciales 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 se construye una malla 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).

Sobre la base de las aproximaciones por INLA y la implementación de la alternativa en el lenguaje de programación R (R-INLA) se han popularizado las aplicaciones de la inferencia bayesiana espacial y espaciotemporal (Cameletti et al., 2013).

Enfoque de aprendizaje automático

El término aprendizaje de máquina o aprendizaje automático corresponde a una rama de la inteligencia artificial. Por definición este tipo de métodos no realizan supuestos sobre la estructura de los datos. Bajo el enfoque de la estadística computacional esa afirmación no es determinante, en este sentido el término minería de datos contempla una combinación de herramientas de estadística y computación (Witten et al., 2016). Se hace referencia a aquellos algoritmos usualmente basados en procesos computacionales intensos que "aprenden" automáticamente de los datos intentando minimizar la intervención humana. Algunos métodos de aprendizaje automático se basan en algoritmos heurísticos de particiones recurrentes de los datos y

evaluaciones de estas hasta identificar la mejor partición para explicar el comportamiento de la variable respuesta como es el caso de los árboles de clasificación y regresión (algoritmos CART) (Breiman, 2001) o algoritmos de redes neuronales, como máquinas de vectores de soporte (Kohonen, 1982) y mapas autoorganizados (Van Hulle, 2012). Estas herramientas se utilizan con fines de predicción y clasificación y son potencialmente útiles para interpretar relaciones lineales y no lineales bajo multicolinealidad. También pueden ser combinadas mediante métodos de muestreo y remuestreo de manera de obtener múltiples predicciones para finalmente ensamblarlas con el fin de minimizar el error de predicción obtenido a partir del modelo ensamblado. La eficiencia de las aplicaciones de aprendizaje automático en contextos predictivos ha mostrado resultados comparables y en algunos casos superiores a las regresiones lineales múltiples (Demetriou, 2017). Entre los algoritmos de este tipo usados con fines predictivos, se tienen las regresiones por bosques aleatorios (RF por el término en inglés "Random Forest") y de árboles de regresión generalizados (GB por el término en inglés "Generalized Boosting") (Efron y Hastie, 2016). Si bien estos algoritmos se han utilizado con frecuencia en el ajuste de modelos predictivos (Kanevski et al., 2009b), la incorporación de la estructura espacial en datos georreferenciados es reciente. En otros ámbitos las máquinas de soporte vectorial y las redes neuronales denotan alto desempeño en contextos predictivos de alta dimensionalidad, aunque en el contexto del mapeo digital de suelo existe una amplia adopción de las regresiones por bosques aleatorios.

RF es un ensamble de múltiples árboles de regresión (Breiman, 2001). Los árboles de regresión emplean particiones binarias recursivas en base a una selección de covariables X para crear grupos homogéneos respecto a la variabilidad de Y. Las particiones continúan hasta que el agregado de otro agrupamiento deja de implicar un decaimiento de la heterogeneidad dentro de los grupos. Una explicación gráfica se puede ver en la Figura 2.

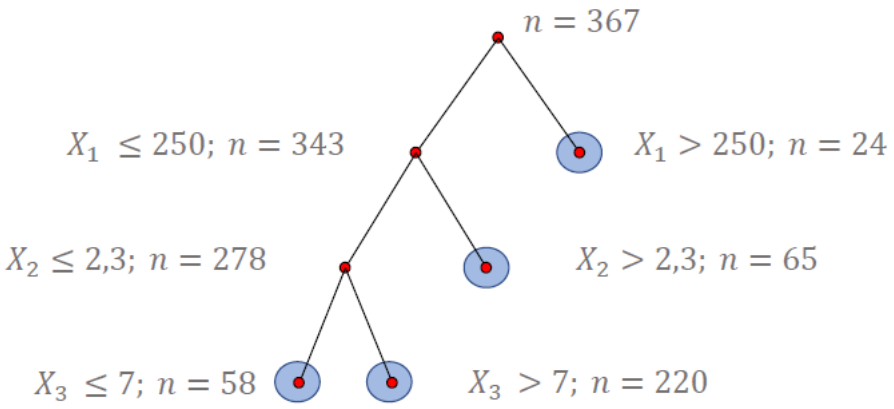


Figura 2: Esquema del resultado de un árbol de regresión.

En el ejemplo Y depende de X_1, X_2 *y* X_3 . X_1 fue seleccionada para la primera partición por lo que se trata de la variable que aporta en mayor medida a explicar Y. Los tres nodos generan 4 agrupamientos relativamente homogéneos en valores de Y para los cuales la estimación puntual corresponde a la media dentro del grupo.

La predicción en un sitio no observado s_0 dado $X = x$ es una media ponderada de las observaciones en los sitios medidos $Y_i,\ i = 1, \dots, n$, $\hat{Y}(s_0) = \sum_{i=1}^{n} w_i(x, \theta)\, Y_i$. Donde $w_i(x, \theta)$ es el peso dado por una constante si x_i es parte de la partición y θ es el indicador del nodo al que pertenece la realización de esa covariable.

Entonces, el algoritmo RF (Breiman, 2001) ensambla múltiples árboles de este tipo formados a partir de muestras Bootstrap. Se realiza una selección aleatoria de covariables para cada muestra Bootstrap, el número de covariables que conforman cada árbol de regresión está definido por el parámetro *mtry*. Luego, el promedio de las predicciones realizadas por todos los árboles se convierte en la predicción puntual y resulta superador a la predicción generada por un único árbol.

Una propuesta para incorporar la estructura espacial es utilizar las coordenadas o matrices de distancias generadas a partir de las mismas como una covariable más en la construcción del modelo (Hengl et al., 2017, 2018). Otra propuesta, es modelar el término remanente del ajuste del algoritmo de aprendizaje automático con una función de autocorrelación espacial (Li et al., 2011) de manera análoga a los modelos del enfoque frecuentistas.

Hipótesis y Objetivos

En el contexto del MDS, los modelos de regresión que consideran la variabilidad espacial subyacente resultan más eficaces que aquellos que tratan los datos como independientes. No obstante, modelos de regresión para datos espaciales derivados desde distintos paradigmas analítico podrían mostrar desempeños similares respecto a los mapas digitales derivados.

La comparación de los enfoques de predicción espacial podría depender de otras particularidades de los escenarios de evaluación como son el número de parámetros a estimar (expresado en la cantidad de covariables utilizadas) y el tamaño muestral.

En contexto de baja dimensionalidad, la predicción Bayesiana podría superar el desempeño de predictores basados en árboles de regresión.

Objetivo general

Implementar y evaluar el desempeño estadístico de predicciones espaciales bayesianas resueltas por aproximación integradas anidadas de Laplace con R-INLA usando ecuaciones diferenciales estocásticas parciales (SPDE) en el contexto del mapeo digital de suelos.

Objetivos específicos

Revisar conceptualmente la predicción bayesiana lograda a partir de modelos de regresión jerárquico estimados INLA y SPDE.

Ilustrar la modelación estadística de procesos estructurados espacialmente en el contexto del mapeo digital de suelo.

Comparar 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.

Capitulo II

Inferencia Bayesiana

Bajo el enfoque bayesiano los parámetros de un modelo se consideran desconocidos y son variables aleatorias que pueden estimarse a partir de la estimación de la distribución a posteriori conjunta $\pi(\theta \mid \mathbf{y})$, es decir la distribución conjunta de los parámetros θ condicional a los datos observados y. La estimación de la distribución a posteriori es el principal objetivo de la inferencia bayesiana y se obtiene a partir del teorema de Bayes:

$$\pi(\theta \mid \mathbf{y}) = \frac{\pi(\mathbf{y} \mid \theta)\pi(\theta)}{\pi(\mathbf{y})} \quad \textbf{[10]}$$

donde $\pi(\mathbf{y} \mid \theta)$ es la función de verosimilitud para los parámetros. $\pi(\theta)$ es lo que denominamos distribución a priori de los parámetros desconocidos y $\pi(\mathbf{y})$ es la verosimilitud marginal. Este ítem resulta difícil de computar y se puede expresar como la siguiente integral:

$$\int_{\Theta} \pi(\mathbf{y} \mid \theta)\pi(\theta)d\theta \quad \textbf{[11]}$$

Frecuentemente $\pi(\theta \mid \mathbf{y})$ es multivariada y solo cuenta con una expresión cerrada o conocida en pocos casos particulares debido a las dificultades en el cálculo de $\pi(y)$. En la práctica la distribución a $\pi(\theta \mid \mathbf{y})$ se estima sin computar la verosimilitud marginal por lo cual la aplicación de Bayes se expresa como:

$$\pi(\theta \mid \mathbf{y}) \propto \pi(\mathbf{y} \mid \theta)\pi(\theta) \quad \textbf{[12]}$$

De esta manera la distribución a posteriori se puede estimar integrando a uno el producto de la verosimilitud y la distribución a priori de los parámetros. La verosimilitud del modelo describe los datos generados por un proceso dado los parámetros. La distribución a priori resume el conocimiento previo acerca de la distribución de los parámetros a estimar. No siempre contamos con información previa para definir o proponer distribuciones a priori con capacidad informativa, en ese caso se proponen lo que se denominan distribuciones a priori vagas o no informativas.

Una vez estimada la distribución a posteriori conjunta se pueden obtener medidas de resumen para los diferentes parámetros de interés en el modelo y para el proceso multidimensional en sí mismo. Especialmente para θ que alberga los parámetros que describen la relación entra la variable respuesta y las variables explicativas. Si imaginamos el proceso como la variación de una propiedad específica del suelo explicada por una serie de variables descriptoras de los procesos formadores sobre el perfil, al obtener una estimación de $\pi(\theta \mid \mathbf{y})$ seremos capaces no solo de conocer la variabilidad de la propiedad en estudio sino cómo contribuye cada variable explicativa a la misma.

Ahora bien, la forma de la distribución a posteriori solo está disponible en forma cerrada para algunos modelos, por lo cual casi siempre es necesaria la estimación por aproximación. Existen modelos que cuentan con lo que se denomina distribuciones a priori conjugadas, en referencia a aquellos en los que la distribución a priori está en la misma familia que la a posteriori. Esta condición facilita la estimación.

Por ejemplo, para un conjunto de observaciones $\{y_i\}_{i=1}^{n}$ distribuidas que siguen una distribución Gaussiana:

$$y_i \mid \mu, \tau \sim N(\mu, \tau),\ i = 1, \dots, n \qquad \textbf{[13]}$$

Con μ desconocido y τ conocido. Entonces la distribución a priori de μ puede ser también Normal con media μ_0 y parámetro de precisión τ_0:

$$\mu \sim N(\mu_0, \tau_0) \qquad \textbf{[14]}$$

La distribución asumida para μ dada los datos observados en Y es $N(\mu_1, \tau_1)$ con

$$\mu_1 = \mu_0 \frac{\tau_0}{\tau_0 + \tau n} + \overline{y} \frac{\tau n}{\tau_0 + \tau n} \qquad \textbf{[15]}$$

$$\tau_1 = \tau_0 + n\tau \qquad \textbf{[16]}$$

La media a posteriori es un compromiso entre la media a priori μ_0 y la media de las observaciones $\overline{y}$. A medida que el número de observaciones crece μ_1 se acerca a $\overline{y}$. En cambio, si n es pequeño la información en la distribución priori adquiere mayor relevancia. Del mismo modo, la precisión τ_1 a posteriori es función de la precisión a priori τ_0 y la función de verosimilitud de la precisión, la cual tiende al infinito a medida que crece n, mientras que la varianza de μ tiende a cero. Este compromiso fundamenta la potencialidad de la inferencia bayesiana frente a contextos de estudios observacionales donde la cantidad de observaciones es escasa, pero abunda la información auxiliar sobre el proceso que genera los datos como es el caso del MDS.

Cuando la distribución a posteriori no presenta una forma cerrada, es necesario recurrir a otros métodos para estimarla o, alternativamente, extraer muestras de ella. Aquí los métodos computacionales adquieren especial relevancia y se dedican a estimar las expresiones integrales que aparecen en la inferencia bayesiana. Por ejemplo, la media posterior del parámetro θ_i que toma valores en el espacio Θ_i, se calcula como:

$$\int_{\Theta_i} \theta_i \, \pi(\theta_i \mid \mathbf{y}) d\theta_i \qquad \textbf{[17]}$$

donde $\pi(\theta_i \mid \mathbf{y})$es la distribución marginal a posteriori de un parámetro θ_i. Entre las opciones que tenemos para resolver este tipo de integrales están las aproximación numérica y aproximación por Laplace. Los métodos clásicos por Monte Carlo que toman muestras de densidades conocidas (hasta una constante) pueden usarse para obtener muestras de la distribución a posteriori. Sin embargo, la mayoría de estos métodos no funcionarán bien en espacios de alta dimensión.

Las estimaciones puntuales para los parámetros son factibles de computar maximizando el producto entre la función de verosimilitud y la distribución a priori y se puede lograr de manera efectiva utilizando diferentes métodos como Newton-Raphson (Gelman et al. 2013).

Para muestrear la distribución a posteriori conjunta existen métodos computacionales por Cadenas Markovianas de Monte Carlo (MCMC). Estos métodos se basan en de una cadena de Markov para la construcción de la distribución a posteriori. Por lo tanto, al tomar muestras repetidas de esta cadena de Markov, se obtienen muestras de la distribución a posteriori conjunta luego de varias iteraciones por algoritmos como Metropolis-Hastings y Gibbs sampling (Brooks et al. 2011). La convergencia se logra luego de suficientes iteraciones y se define a través de diversos criterios.

Estimación vía INLA

Otro enfoque para lograr la inferencia bayesiana ha sido propuesto por Havard Rue, Martino y Chopin (2009), quienes a trevés de método computacionales se concentran en estimar las distribuciones a posteriori marginales para cada uno de los parámetros en el modelo. Es decir, en lugar de estimar la distribución a posteriori conjunta $\pi(\theta \mid \mathbf{y})$ (distribución multivariada de alta multidimensionalidad), INLA (por su término en inglés *"Integrated Nested Laplace Approximation"*) estima las distribuciones marginales a posteriori de cada parámetro del modelo $\pi(\theta_i \mid \mathbf{y})$.

Los modelos que INLA puede ajustar están restringidos aquellos modelos que pueden expresarse como un Campo Aleatorio Markoviano Gaussiano Latente (GMRF) por razones computacionales

(Havard Rue & Held, 2005). No obstante, los GMRF cubren un amplio rango de modelos dando lugar a numerosas aplicaciones en análisis de datos espaciales (Cameletti et al., 2013; Huang et al., 2017; Moraga et al., 2017; Poggio et al., 2016).

Para un vector de variables aleatorias $\mathbf{y} = (y_1, \dots, y_n)$ de distribución perteneciente a la familia exponencial, la media μ_i para la observación y_i se encuentra asociada a un predictor lineal η_i mediante una función de enlace conveniente. El predictor lineal puede incluir efectos que contribuyen a la media condicional de diversos tipos de covariables y aquellos efectos que aportan una variabilidad no controlada también de diversas fuentes. Entonces el vector correspondiente a los efectos latentes del modelo se denota como x incluyendo los coeficientes para de las covariables. Además, la distribución de $\mathbf{y}$ puede depender de otro vector de hiperparámetros θ_1. La distribución de los efectos latentes $\mathbf{x}$ se asume un GMRF que tendrá una media igual a cero y una matriz de precisión $Q(\theta_2)$ siendo θ_2 un vector de hiperparámetros, entonces el vector de todos los hiperparámetros del modelo se denota $\theta = (\theta_1, \theta_2)$. A su vez las observaciones se asumen independientes dados $\mathbf{x}$ y θ. Entonces la verosimilitud puede escribirse como:

$$\pi(\mathbf{y}|\mathbf{x}, \theta) = \prod_{i \in \mathcal{I}} \pi\,(y_i|\eta_i, \theta). \qquad \textbf{[18]}$$

Donde η_i es el predictor lineal latente que es una combinación lineal de $\mathbf{x}$. Los índices para los valores observados de $\mathbf{y}$ se encuentran en $\mathcal{I}$ es importante aclarar que pueden existir valores faltantes en $\mathbf{y}$.

La metodología por INLA consiste entonces en aproximar las distribuciones marginales de los parámetros $\mathbf{x}$ e hiperparámetros θ del modelo, aprovechando las propiedades computacionales propias de los GMRF y el método de aproximación de integrales por Laplace. La distribución a posteriori conjunta de $\mathbf{x}$ y θ dado $\boldsymbol{y}$ se puede expresar como

$$\begin{aligned}(\mathbf{x}, \theta|\mathbf{y}) &\propto \pi(\theta)\pi(\mathbf{x}|\theta) \prod_{i \in \mathcal{I}} \pi\,(y_i|x_i, \theta) \\ &\propto \pi(\theta)|\mathbf{Q}(\theta)|^{1/2} \exp\{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{Q}(\theta)\mathbf{x} + \sum_{i \in \mathcal{I}} \log\left(\pi(y_i|x_i, \theta)\right)\}\end{aligned} \qquad \textbf{[19]}$$

Donde $\mathbf{Q}(\theta)$ representa la matriz de precisión para los efectos latentes y $|\mathbf{Q}(\theta)|$ su determinante. Además, $x_i = \eta_i$ cuando $i \in \mathcal{I}$. Dado la estructura de los GMRF $\mathbf{Q}(\theta)$ se presenta frecuentemente como una matriz rala.

El cómputo de las distribuciones marginales de los efectos latentes y los hiperparámetros se realiza considerando las siguientes expresiones:

$$\pi(x_i|\mathbf{y}) = \int \pi(x_i|\boldsymbol{\theta},\mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta},$$
$$\pi(\theta_j|\mathbf{y}) = \int \pi(\theta|\mathbf{y})d\theta_{-j} \quad \textbf{[20]}$$

En ambas expresiones la integración se realiza sobre el espacio de los hiperparámetros por lo que es necesaria una buena aproximación de la distribución a posteriori de estos. Rue et al. (2009) proponen aproximar $\pi(\theta|\mathbf{y})$ denotada por $\tilde{\pi}(\theta|\mathbf{y})$y utilizarla para aproximar las distribuciones a posteriori marginales de los parámetros latentes x_i como:

$$\tilde{\pi}(x_i|\mathbf{y}) = \sum_k \tilde{\pi}\,(x_i|\theta_k,\mathbf{y}) \times \tilde{\pi}(\theta_k|\mathbf{y}) \times \Delta_k \quad \textbf{[21]}$$

Donde Δ_k son ponderaciones o pesos asociadas con un vector de valores para los hiperparámetros θ_k en una grilla. INLA obtendrá estos puntos de integración remplazando la moda a posteriori de θ_k o utilizando un diseño de composición central centrado en la moda a posteriori (Gómez-Rubio, 2020). Luego, existen diferentes formas de aproximar $\tilde{\pi}(\theta_k|\mathbf{y})$ que pueden encontrarse en Håvard Rue et al. (2009).

Modelación Espacial vía INLA-SPDE

La metodología INLA encuentra implementación en una librería específica denominada INLA (Finn Lindgren & Rue, 2015a) desarrollada en el software R, y presentando sintaxis similar con librerías clásicas como *glm* y *gam,* lo que facilita su implementación. Esta herramienta cuenta con una gran gama de opciones para proponer efectos aleatorios lo que la hacen especialmente atractiva en el contexto del MDS. Los modelos aleatorios en INLA se definen utilizando una distribución normal multivariada de media 0 y matriz de precisión que se puede expresar genéricamente como $\tau\Sigma$, donde τ es un parámetro de precisión genérico y Σ es la matriz que define la estructura de los efectos aleatorios del modelo que a su vez pueden depender de otros parámetros. Suponiendo una estructura espacial, en un contexto de predicción espacial estos otros parámetros que determinarían Σ serían análogos a los parámetros que definen una función de semivarianza clásica (nugget, rango, varianza estructural). De nuevo, cuando los efectos aleatorios son GMRF la estructura de Σ es lo suficientemente rala y lo simplifica en términos computacionales. Ahora abordemos la alternativa para expresar un efecto aleatorio espacial como un GMRF.

Supongamos que llamamos s a un sitio específico en un área de estudio definida como D y definimos $U(s)$ sea el efecto aleatorio espacial en ese sitio. Decimos que $U(s)$ es un proceso estocástico con $s \in D$ donde $D \subset R^d$. Por ejemplo, para el dominio espacial definido en este trabajo, D es la provincia de Córdoba, territorio sobre el cual se han realizado una serie de medidas en sitios georreferenciados (s) donde d=2, es decir queda definido en dos dimensiones (Latitud y Longitud). Denotamos la medición en cada sitio $u(s_i), i = 1,2, \dots, n$ como una realización de $U(s)$. Comúnmente se asume que $u(s)$ tiene una distribución Normal multivariada. Si $U(s)$ se asume como un proceso continuo en D entonces lo catalogamos como un campo gaussiano (por su término en inglés *Gauassian Field,* GF) (Besag, 1977) que implica que es posible obtener realizaciones en cualquier sitio dentro de la región de estudio.

Para completar la especificación de la distribución de $u(s)$ debemos definir su media y varianza. Una opción es definir una función basada en una distancia Euclídea entre las observaciones, asumiendo que existe igual correlación para dos pares de puntos separados por la misma distancia h. Para extender la aproximación a variables aleatorias no normales se asume la función de verosimilitud de los datos condicional a los efectos aleatorios no observados, también definidos como un como campo gaussiano GF (Diggle et al., 2003). Mediciones una variable aleatoria y_i (en este caso una variable edáfica) observada en diferentes sitios se asumen como realizaciones de un campo gaussiano que se puede expresar como:

$$\begin{aligned} y_i|\beta, u_i, \mathbf{F}_i, \phi &\sim f(y_i|\mu_i, \phi) \\ \mathbf{u}|\theta &\sim GF(0, \Sigma) \end{aligned} \qquad \textbf{[22]}$$

Donde $\mu_i = h\left(\mathbf{F}_i^T\beta + u_i\right)$. $\mathbf{F}_i$ es la matriz que contiene las covariables asociadas a los coeficientes β, u es el vector de efectos aleatorios de sitios y la función $h_i(.)$ lleva el predictor lineal definido por $\mathbf{F}_i^T\beta + u_i$ a $E(y_i) = \mu_i$. Luego θ son los parámetros de (co)varianza y ϕ un parámetro de dispersión de la distribución que describe la variable aleatoria medida $f(.)$ la cual debe ser parte de la familia exponencial.

Tal vez la función de correlación más utilizada para proponer estructuras espaciales de efectos aleatorios sea la función de correlación de Matérn, definida para un proceso estacionario e isotrópico como:

$$Cor_M(U(s_i), U(s_j)) = \frac{2^{1-\nu}}{\Gamma(\nu)}(\kappa \parallel s_i - s_j \parallel)^\nu K_\nu(\kappa \parallel s_i - s_j \parallel) \qquad \textbf{[23]}$$

Aquí se presentan dos parámetros, un parámetro de escala $k > 0$ y un parámetro de suavizado $v > 0$. La distancia utilizada entre par de observaciones s_i y s_j es la distancia euclídea en la expresión anterior definida como $\|.\|$. K_v es una función de modificada Bessel de segundo orden (Weniger & Cížek, 1990). Entonces la función de covarianza de Matérn es:

$$\sigma_u^2 Cor_M\left(U(s_i), U(s_j)\right) \quad \textbf{[24]}$$

donde σ_u^2 es la varianza marginal del proceso. Ahora siendo $u(s)$ una realización de $U(s)$ en los n sitios observados entonces podemos expresar la distribución conjunta de la matriz de varianzas y covarianzas como:

$$\Sigma_{i,j} = \sigma_u^2 Cor_M\left(u(s_i), u(s_j)\right) \quad \textbf{[25]}$$

De nuevo es frecuente asumir que $U(.)$ tiene una media igual a cero y entonces queda definida una distribución multivariada para $u(s)$. Para un mayor desarrollo de Matérn como un GF ver Krainski et al. (2018).

Lindgren, Rue y Lindström (2011b) propusieron un nuevo enfoque para representar la función de covarianza de Matérn como un GF, es decir un campo markoviano gaussiano aleatorio, GMRF (Håårard Rue & Tjelmeland, 2002) (por su término en inglés, "*Gaussian Markov Random Fields"*). Esta representación es posible a través de la solución de ecuaciones diferenciales estocásticas parciales (SPDE por su término en inglés, "*Stochastic Partial Differential Equation"*) utilizando una aproximación por el método de los elementos finitos (Zienkiewicz et al., 1977) (solo posible para algunos valores de v, ver ecuación de correlación de Matérn). Lindgren, Rue y Lindström (2011b) desarrollaron esta solución eligiendo ciertas funciones de base que preservan la estructura dispersa de la matriz de precisión resultante para el campo aleatorio en un conjunto de nodos de una malla. El principal beneficio de la representación GMRF del GF es que permite el cálculo explícito de la matriz de precisión ($\Sigma^{-1} = Q$) que a su vez cuenta con la ventaja de, como se dijo, una estructura rala con excelentes propiedades computacionales para un cálculo factible de implementarse vía R-INLA (Krainski et al., 2018). Se presentan aquí dos resultados resultados elementales para comprender la aproximación por SPDE basados en la síntesis de Krainski et al. (2018) para más detalle se recomienda recurrir a el apéndice de Lindgren et al. (2011).

El primer resultado es una extensión de Besag, (1981) y establece que un GF que parte de una función de covarianza general de Matérn Ec. [18] obtenida cuando $v > 0$ es una solución de una

SPDE. Se analiza este resultado sobre un ejemplo considerando un cuadrado latice en dos dimensiones

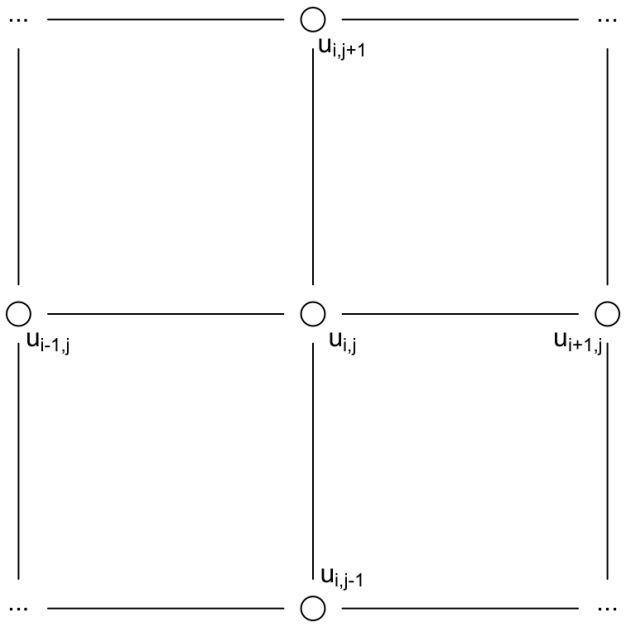


Figura 3: Representación de sitios en un cuadrado látice de dos dimensiones para estimar un proceso espacial (Krainski et al., 2018).

Aquí la esperanza condicional en el sitio i, j es:

$$E(u_{i,j}|u_{-i,j}) = \frac{1}{a}(u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1}) \quad \textbf{[26]}$$

Y la varianza $Var(u_{i,j}|u_{-i,j}) = \frac{1}{a}$ con $|a| > 4$. Aquí la matriz de precisión representada matricialmente es:

$$\begin{vmatrix} -1 & \\ a & -1 \end{vmatrix} \quad \textbf{[27]}$$

Como se muestra, solo el cuadrante superior derecho (en Σ) y con una a como elemento central. Entonces un GF $U(s)$ con una función de covarianza de Matérn se puede expresar como una solución de las siguientes funciones SPDE:

$$(\kappa^2 - \Delta)^{\alpha/2} u(s) = W(s), \quad \textbf{[28]}$$

$$\text{Con: } s \in R^d,\ \alpha = \nu + d/2,\ \kappa > 0,\ \nu > 0$$

donde Δ es el operador Laplaciano definido como $\sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$ y d la dimensión del dominio espacial. $W(s)$ es el ruido blanco espacial que es un proceso estocástico gaussiano de varianza unitaria. La función de varianza marginal σ^2 se puede expresar como una SPDE a través de:

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)^{d/2}\kappa^{2\nu}\tau^2} \qquad \textbf{[29]}$$

Esta para dos dimensiones parametrizada con $\nu = \frac{1}{2}$ se corresponde a una función de covarianza del tipo exponencial. Lo que define un $\alpha = \frac{3}{2}$ (Moraga, 2019).

Lindgren et al., (2011) demuestran que las parametrizaciones de $\nu = 1$ y $\nu = 2$ la representación del GMRF es una convolución de la matriz anterior:

$$\begin{array}{|ccc} & 1 & \\ -2a & & 2 \\ 4+a^2 & -2a & 1 \end{array} \qquad \textbf{[30]}$$

$$\nu = 1$$

$$\begin{array}{|cccc} & -1 & & \\ 3a & & -3 & \\ -3(a^2+3) & 6a & & -3 \\ a(a^2+12) & -3(a^2+3) & 3a & -1 \end{array} \qquad \textbf{[31]}$$

$$\nu = 2$$

Como se puede observar a medida que el parámetro de suavizado crece la matriz de precisión en el GRF se vuelve menos rala o más densa. Además, mayores densidades en la matriz se dan a medida que las distribuciones condicionales dependen de un vecindario mayor.

El segundo resultado resuelve el caso de grillas irregulares ya que las grillas regulares rara vez se obtienen en la práctica. Esta solución se basa en un método ampliamente utilizado en matemática aplicada e ingeniería para resolver ecuaciones diferenciales, el método de los elementos finitos (Zienkiewicz et al., 1977).

El dominio en estudio puede dividirse en una serie de triángulos sin solaparse, que pueden ser irregulares, en donde dos triángulos comparten cómo máximo un lado. Los vértices de cada triángulo se denominan nodos. Aquí la solución para el SPDE depende de las funciones base utilizadas, que han sido cuidadosamente seleccionadas de manera tal que se preserve la estructura laxa en la matriz de precisión. La aproximación se define de la siguiente manera:

$$u(s) = \sum_{k=1}^{m} \psi_k(s) w_k \qquad \textbf{[32]}$$

donde ψ_k es una serie de funciones de base, w_k coeficiente de ponderaciones. $k = 1, \dots, m$ con m nodos o vértices en una malla. w_k tiene media cero y distribución normal.

La definición de un modelo de regresión espacial bayesiana estimado vía SPDE es:

$$\begin{gathered} y|\beta_0, u, \sigma_e^2 \ \sim \ N(\beta_0 + \mathbf{A}u, \sigma_e^2) \\ u \sim GF(0, \Sigma) \end{gathered} \qquad \textbf{[33]}$$

Dado n observaciones y_i en los sitios s_i, β_0 es el intercepto, $\mathbf{A}$ es la matriz de proyección y u el campo gaussiano espacial. La matriz A cumple la función de asociar el GMRF definido a partir de los nodos de la malla a los sitios donde se ha observado o medido el dato.

Capítulo III

Modelos de predicción espacial en mapeo digital de suelo. Ilustraciones

En este capítulo se aplican modelos de predicción espacial, estimados desde distintos enfoques, en el MDS. Se comparan los resultados de distintos predictores espaciales sobre distintas bases de datos conformadas a partir de múltiples propiedades de suelo para un conjunto *n* de sitios distribuidos espacialmente a diferentes escalas.

Materiales y Métodos

La implementación se ilustró con tres bases de datos espaciales de características contrastantes:

Materia orgánica del suelo

La base de datos de materia orgánica del suelo (MOS) pertenece a la infraestructura de datos espaciales de la provincia de Córdoba, (IDECOR) (Piumetto et al., 2019). Las muestras de MOS provienen de la combinación de muestreos realizados por la Secretaría de Agricultura de la Provincia, el Instituto de Tecnología Agropecuaria, la Facultad de Cs. Agropecuarias de la Universidad Nacional de Córdoba, el Instituto Multidisciplinario de Biología Vegetal de CONICET-UNC, la Universidad Nacional de Río Cuarto y tres empresas privadas. En la etapa de sistematización y preprocesamiento, se estandarizaron los datos a una profundidad de 0-20 cm mediante funciones de suavizado *splines* de dimensión de orden 1 a través del paquete GSIF (Reuter y Hengl, 2012).

La base de datos combinada conforma una muestra de 3260 sitios distribuidos en Córdoba (Figura 4). Cada sitio fue caracterizado por covariables ambientales que describen otras propiedades de suelos obtenidas a partir de mapas de suelos preexistentes, datos de cobertura y uso del suelo, datos de la dinámica de la vegetación a partir de información remota satelital, datos de mapas litológicos y datos extraídos desde modelos digitales de elevación (MDE), como atributos topográficos primarios y secundarios (Tabla 1). La distribución espacial de las covariables usadas se puede consultar en la web de IDECOR mientras que la distribución de valores de la variable dependiente MOS se presentan en la Figura 1 cubriendo un área total de alrededor de 165000 km2.

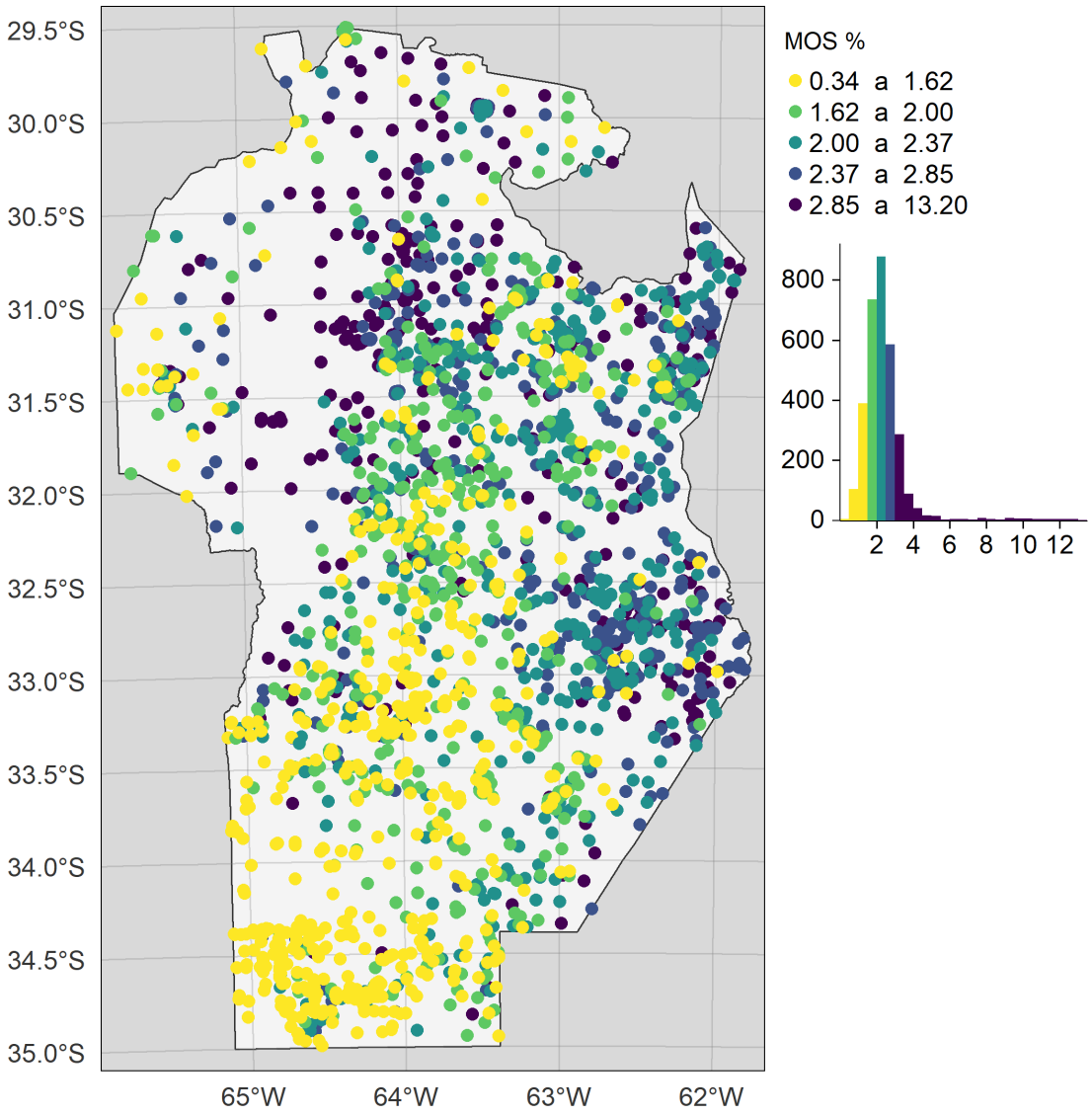


Figura 4: Materia orgánica del suelo (MOS) en 3260sitios de muestreo de Córdoba, Argentina. Los valores de MOS se encuentran expresados en unidad de porcentaje p/p (IDECOR, 2019).

Tabla 1. Potenciales variables explicativas de MOS.

Factor scorpan	Variable	Código	Fuente
Ubicación geográfica	x y		Coordenadas geográficas (m) sistema (Sistema WGS 84)
Clima	Precipitación media anual	pp_med_anua	Producto satelital *World Clim version 2*
	Temperatura máxima anual	t_max_med	
	Temperatura media anual	t_med_anua	
	Temperatura mínima anual	t_min_med	
	Radiación solar media	rad_solar	Producto satelital *TerraClimate*
	Déficit hídrico (promedio 2001-2020)	def_hidric	
	Evapotranspiración media mensual acum (promedio 2001-2018)	evapo_medi	Producto satelital *MOD16*
Acción de los organismos vivos	NDVI mediana, serie 2001-2020	ndvi_media	Producto MOD13Q1 V6 (Teich et al. 2019)
	SWATI (ESPI). Tendencia NDVI 2001-2018	ssw_espi_v	
Suelo	Índice Productividad de suelo Profundidad efectiva Arcilla (Mapa MDS)	ip_med	Cartas de suelo
Topografía	Altura (m.s.n.m.)	altura_med	Modelo digital de Elevación Instituto Geográfico Nacional y productos derivados
	Pendiente (%)	slope_rp	
	Índice de humedad topográfico (TWI)	twi	
	Depresiones cerradas	flow_accum	
	Zona de captación	catch_area	
	Zona de captación modificada	mod_catch	
	Pendiente de captación	catch_slop	
	LS-Factor	ls_factor	
	Profundidad del Valle (Valley Depth)	valley_dep	

Metales pesados

La segunda base de datos a usar corresponde a un muestreo de metales pesados en suelo realizado en la orilla este del Río Meuse en Holanda (Figura 5) (Burrough y McDonnell, 1998). Cada sitio está caracterizado por las covariables de suelo y paisaje enumeradas en la Tabla 2. La variable respuesta en la modelación predictiva será el log de la concentración de Zn (cinc) medida en ppm (Figura 5),

y el resto de las variables continuas se usará como predictoras. La grilla de muestreo tiene una dimensión aproximada de 15 x 15 m y el área de estudio es de alrededor de 34000 m^2.

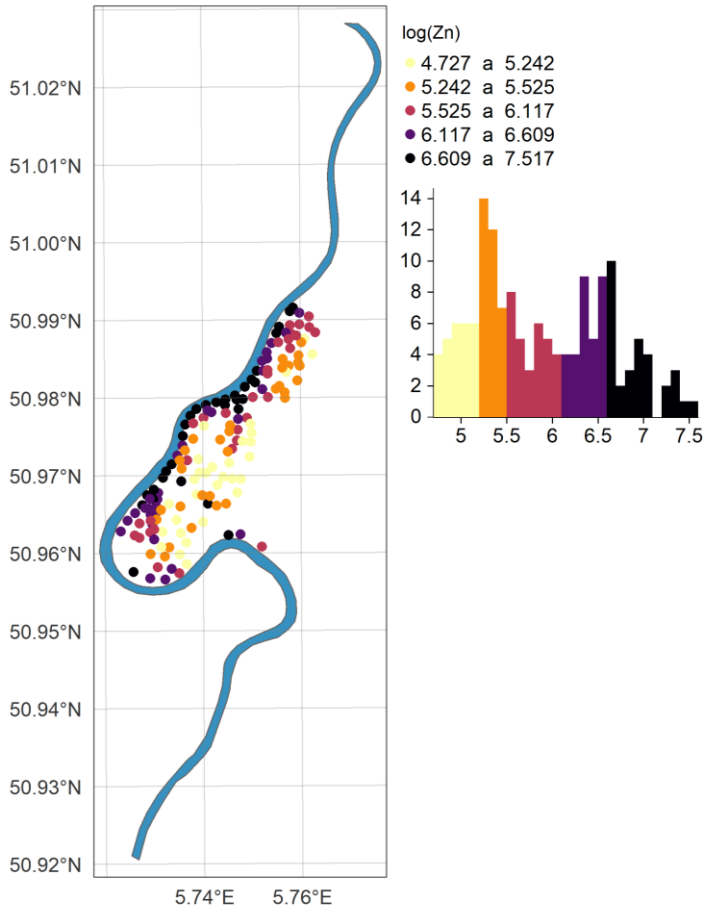


Figura 5: Metales pesados en n=153 sitios de muestreo a orillas del Río Meuse (Burrough y McDonnell, 1998).

Tabla 2: Potenciales variables explicativas ln(Zn)

Factor scorpan	Variable	Descripción
Ubicación geográfica	x	Coordenadas geográficas (m) sistema RDM (Dutch Topographical map coordinates)
	y	
Suelo	Cd	Concentración de metales en suelo (ppm)
	Cu	
	Pb	
	Zn	
	MOS	Materia Orgánica del Suelo %
Topografía	elev	m.s.n.m.
	Distancia al río	Distancia en metros obtenida desde el muestreo

Adsorción atrazina

La tercera base de datos a usar para la ilustración del mapeo digital se construyó a partir de determinaciones de laboratorio del coeficiente de adsorción (Kd) del herbicida atrazina en suelo. Este índice expresa la relación entre la concentración de una molécula aplicada al suelo que es retenida en la fase sólida y la concentración de ésta en fase acuosa (Bailey y White, 1970). La unidad se expresa en unidades de volumen por masa; mientras mayor es el Kd, mayor es la adsorción de la molécula al suelo. Se evalúa aquí una muestra de tamaño n=156 de log de coeficientes Kd para la molécula del principio activo del herbicida atrazina. (Figura 6). Las covariables usadas son propiedades edafoclimáticas de los sitios de muestreo (Tabla 3) (Giannini-Kurina et al., 2020).

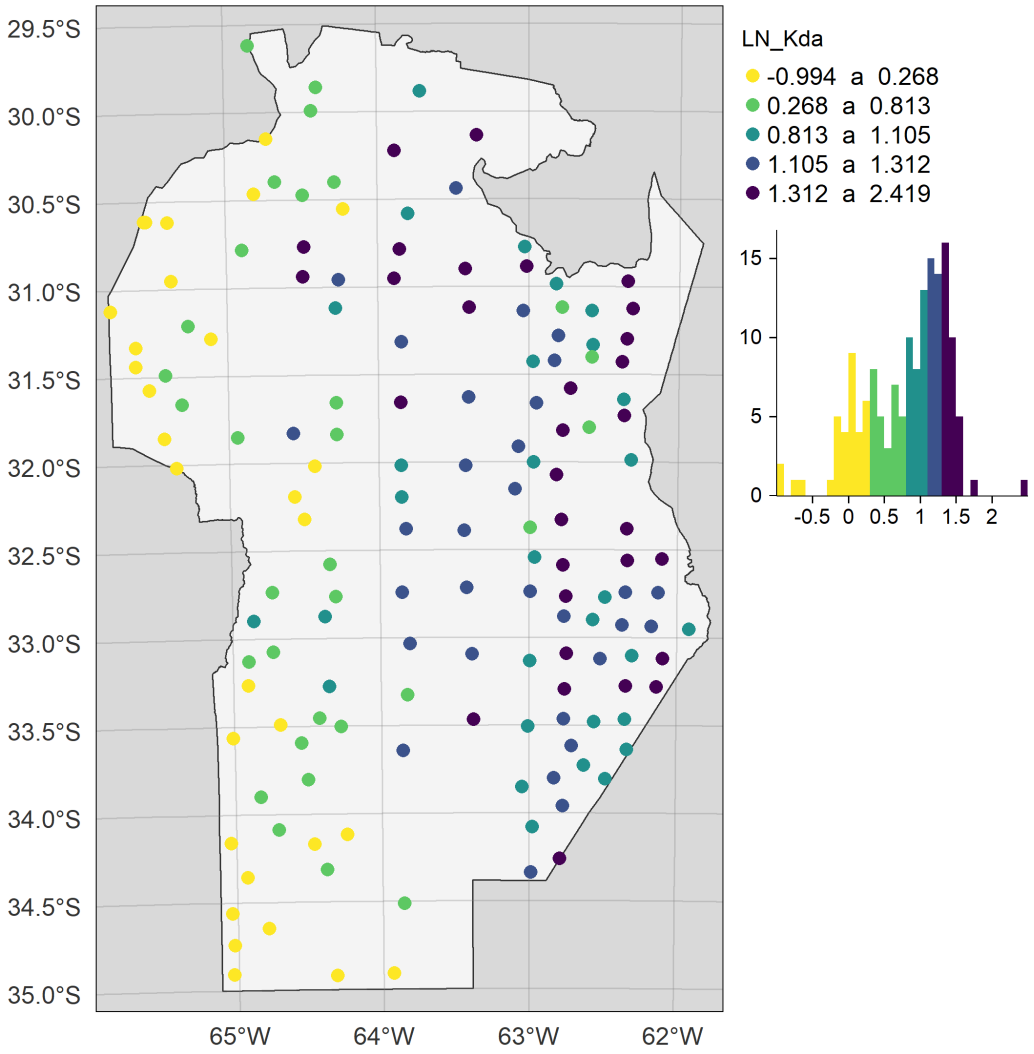


Figura 6: Coeficiente de adsorción de atrazina en n=156 sitios muestreados en Córdoba, Argentina (Giannini-Kurina et al., 2019).

Tabla 3. Potenciales variables explicativas índice de absorción Atrazina.

Factor scorpan	Variable	Unidades	Descripción
Ubicación geográfica	X UTM20	m	Sistema de coordenadas Universal Transversal de Mercator zona 20.
	Y UTM20	m	
Suelo	pH	-	pH en agua 1:2,5 (suelo:agua)
	CE	dS m−1	Conductividad Eléctrica en agua 1:2,5 (suelo:agua)
	COT	g kg-1	Carbono Orgánico Total por combustión húmeda por 1N K2Cr2O7, método Walkley y Black (Sparks, Helmke y Page, 1996)
	NT	% p:p	Nitrógeno Total, método Kjeldahl (Sparks et al., 1996)
	Mn	mg kg−1	Magnesio extractable por Mehlic-3 (Mehlich, 1984)
	Cu	mg kg−1	Cobre extractable por Mehlic-3 (Mehlich, 1984)
	Zn	mg kg−1	Zinc extractable por Mehlic-3 (Mehlich, 1984)
	CC	%	Capacidad de Campo, 300 kPa en olla de presión (Klute, 1986)
	Arena	%	Contenido de arena, método de pipeta de Robinson (Sparks et al., 1996)
	Limo	%	Contenido de limo, método de pipeta de Robinson (Sparks et al., 1996)
	Arcilla	%	Contenido de arcilla, método de pipeta de Robinson (Sparks et al., 1996)
	Al(Ox)	%	Óxidos de Aluminio (Loeppert, Inskeep y Sparks, 1996)
	Fe(Ox)	%	Óxidos de Hierro (Loeppert et al., 1996)
	P	ppm	Fósforo extractable por Bray y Kurtz, medición por colorimetría (Sparks et al., 1996)
	K	ppm	Potasio intercambiable, medición por fotometría de llama (Sparks et al., 1996)
	Ca	ppm	Calcio intercambiable, complexometría (Sparks et al., 1996)
	Na	ppm	Sodio intercambiable, fotometría de llama (Sparks et al., 1996)
	Mg	ppm	Magnesio intercambiable, complexometría (Sparks et al., 1996)
	CIC	Cmol kg−1	Capacidad de Intercambio Catiónico (Sparks et al., 1996)
Topografía	Elevación	m.s.n.m	Elevación, Modelo Digital de Elevación, STRM (Farr et al., 2007)
	Pendiente	%	Pendiente derivada de DEM STRM (Farr et al., 2007)
Clima	pp	mm	Precipitaciones acumuladas anual, BIOCLIM (Booth et al., 2014)
	Tm	°C	Temperatura media anual, BIOCLIM (Booth et al., 2014)
	TvsPP	°C mm-1	Cociente entre Tm y pp, indicador de balance hídrico

Análisis Estadístico

Caracterización y análisis exploratorio de las bases de datos de ilustración

Para cada una de las tres bases de datos se analizaron las estructuras de correlación a partir de una matriz de correlaciones presentado como un gráfico de correlaciones (Wei et al., 2017). Se caracterizo la distribución de cada variable respuesta y la autocorrelación espacial a través del Índice de Moran global y el Índice de Geary, cuyas expresiones son:

$$IMG = \frac{n}{(n-1)S^2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(Z_{(s_i)} - \bar{Z}\right)\left(Z_{(s_j)} - \bar{Z}\right) \quad \textbf{[34]}$$

$$GI = \frac{(n-1)\sum_i \sum_j w_{ij}\left(Z(s_i) - Z(s_j)\right)^2}{2w..\sum_i (Z(s_i) - \bar{z})^2} \quad \textbf{[35]}$$

donde w_{ij} es el peso de conectividad espacial obtenido a partir de una red de vecindarios previamente establecida. Para la red de vecindarios se usó una distancia máxima de 10 Km para la base MOS, 30 m para el Meuse, 30 Km para set de Kd. En cada variable se caracterizó el grado de estructuración espacial a partir del ajuste de un modelo de semivariograma desde donde se derivó una medida de la varianza estructural relativa (RSV, Eq [9]).

Algoritmos de predicción espacial

Los algoritmos para la predicción espacial implementados en este trabajo son: 1) Regresión lineal múltiple con errores correlacionados espacialmente (“Regression Krigging”, RK); 2) Regresión por bosques aleatorios con errores correlacionados espacialmente ("Random Forest ", RF); 3) Regresión Bayesiana (RB) ajustados con estructuras espaciales SPDE (“Integrated Nested Laplace Approximation con SPDE”, INLA-SPDE). La selección de RK y RF para comparar frente a RB está fundada no solo en los marcos estadísticos conceptuales contrastantes de cada algoritmo sino que se trata de los dos algoritmos más utilizados en el área de estudio (Lamichhane, Kumar y Wilson, 2019). La implementación de las rutinas de análisis se realizó en el software R (R Core Team, 2020).

Las rutinas iterativas de gran costo computacional se corrieron en el computador Mendieta del Centro de Cómputos de Alto Desempeño (CCAD) de la Universidad Nacional de Córdoba.

Para la (RK) se utilizó la librería clásica *"lm"*, luego los residuos de este modelo se modelan mediante un ajuste automático para identificar la estructura de semivarianza remanente. Para identificar y ajustar el modelo espacial sobre los residuos se utilizaron funciones del paquete *gstat* (Gräler, Pebesma y Heuvelink, 2016).

Los ajustes de modelos de regresión lineal se realizaron utilizando la siguiente fórmula:

$$Y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_{si} \qquad \textbf{[36]}$$

$$e_{si} = w_i e_i$$

donde Y_i es el valor de la variable respuesta (Contenido de MOS, log de la concentración de Zn o log de Kd de atrazina) en un sitio dado *i*, β_0 es la ordenada al origen; β_j es el coeficiente de regresión asociado con cada variable explicativa x_{ij}. e_{si} es el término de error aleatorio. El residuo del modelo de regresión lineal de efectos fijos e_i a su vez se afectado por un coeficiente w_i proveniente de una matriz de pesos que se obtiene de una función de semivarianza ajustada sobre los residuos del modelo de efectos fijos. A continuación se presenta la notación matricial según Hengl, Heuvelink y Stein (2004), donde se establece el método y se especifica que las predicciones se hacen de manera separada, por un lado para los efectos fijos o tendencias y luego para los residuos.

$$\hat{z}(s_0) = q_0^T\hat{\beta} + \lambda_0^T e \qquad \textbf{[37]}$$

Donde $\hat{z}(s_0)$ es el valor predicho q_0 es el vector correspondiente a las $p+1$ covariables predictoras en el sitio de predicción, $\hat{\beta}$ es el vector de $p+1$ parámetros estimados de los efectos fijos del modelo, λ_0 es el vector de los n pesos kriging obtenidos a partir del modelo propuesto de matriz de varianza y covarianza de los residuos de $q_0^T\hat{\beta}$. De la misma maneta se extiende esta relación aditiva a la varianza de las predicciones, donde se suma la varianza residual de los efectos fijos y la varianza kriging sobre los residuos.

El ajuste del algoritmo de aprendizaje automático, RF , se realizó con kriging sobre los residuos (Li et al., 2011). Para evitar el sobreajuste, el modelo espacial es ajustado sobre una muestra de los datos, los cuales son diferentes a los datos de entrenamiento. El ajuste de estos modelos se realizó en el software R, especificando el modelo *randomForest* (Liaw y Wiener, 2002) en la función *train* del paquete *caret* (Kuhn, 2015). El algoritmo fue optimizado dentro de la misma función, utilizando validación cruzada por *k-fold* con k=10. Para identificar y ajustar el modelo espacial sobre los residuos se utilizaron funciones del paquete *gstat* (Gräler et al., 2016).

Finalmente, se ajustó un modelo jerárquico bayesiano o regresión bayesiana para datos espaciales (RB):

$$
\begin{aligned}
Y_i &\sim N(\eta_i, \sigma_e^2) \\
\eta_i &= \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \xi(s_i) \\
\xi(s_i) &\sim \mathrm{N}(0, \Sigma_{n\times n})
\end{aligned}
\qquad \textbf{[38]}
$$

donde Y_i es el contenido de la variable respuesta en el sitio i; β_0 es la ordenada al origen; β_j es el coeficiente de regresión asociado a las covariables x_{ij}; x_{ij} la valuación de x_j en el sitio i y $\xi(s_i)$ el efecto aleatorio de sitio que se asume una realización de un proceso gaussiano latente $\xi(s_i) \sim \mathrm{MVN}(0, \Sigma)$. Siendo Σ la matriz de varianza y covarianza de los efectos de sitio definidos por la función de covariación espacial de Matérn (Matérn, 1986), obtenida mediante aproximación por Laplace (INLA) aproximada por suavizado mediante el método de ecuaciones diferenciales estocásticas (SPDE). La estimación de la inversa de Σ (matriz de precisión) se realizó usando R-INLA. La distribución a posteriori de los valores predichos se obtiene para cada nodo de una malla de predicción, la mediana de la distribución representa el rendimiento esperado para los niveles de insumos especificados y el desvío estándar de la distribución de los valores predichos en cada sitio provee una medida de incertidumbre de la predicción. El modelo especial bayesiano se ajustó utilizando el paquete *INLA* (Lindgren y Rue, 2015) del software R. Para la definición de la malla se utilizó una herramienta de INLA denominada *meshbuilder* generada utilizando el paquete *Shiny* (Krainski et al., 2018) que permite especificar los parámetros de manera interactiva y grafica la malla.

Resultados y Discusión

Caracterización y análisis exploratorio de las bases de datos de ilustración

Las tres bases de datos seleccionadas para la implementación de algoritmos de predicción corresponden a tres escenarios contrastantes en el contexto del MDS. La base de datos de MOS cuenta con un gran número de observaciones y de covariables con capacidad explicativa, es decir tanto en tamaño (n) como en dimensión (p) se trata de la base de datos más robusta entre las utilizadas. Además, es importante notar que MOS es una variable ampliamente estudiad por lo cual los factores SCORPAN que determinan su variabilidad se conocen de estudios previos.

La distribución de metales pesados en suelos también es un interrogante que se ha abordado con anterioridad mediante técnicas de MDS no solo por las múltiples implementaciones por ser un conjunto de datos de referencia sino porque los procesos de contaminación ambiental por metales pesados son casos que adquieren especial relevancia para la sociedad. Es la base de datos que cuenta con menor n y p. Aunque no se cuenta con gran cantidad covariables, existe conocimiento previo respecto a la fuente de contaminación y a la dinámica en suelo de los metales como Zn lo que deviene en facilidades a la hora de modelar su distribución espacial. Además, la escala que abarca es distinta respecto a las otras dos bases de datos siendo la de mayor resolución.

La base de datos de coeficientes de adsorción de Atrazina es tal vez aquella que presenta mayores dificultades a la hora de proponer un modelo. Por un lado, se trata de la variable respuesta menos estudiada, además, es un parámetro edáfico que resume interacciones entre otras propiedades de suelo y una sustancia aplicada al suelo, por lo que podemos decir que es un orden de complejidad mayor ya que es factible que los factores formadores de suelo estén incidiendo no de manera directa sobre la variabilidad de Kd sino a través de otras propiedades edáficas más “básicas” o intrínsecas del suelo. Por otro lado, es la base de datos de menor resolución y la que cuenta con la relación entre n y p menos robusta.

Estructura de correlación entre variables

A continuación, se presentan las matrices de correlación entre variables para cada una de las bases de datos. La matriz de correlación para MOS de los Suelos de Córdoba indica que existen altas correlaciones entre algunas variables, por ejemplo, las variables derivadas de modelos digitales de elevación presentan altas correlaciones entre sí. Además, variables climáticas como las temperaturas están correlacionadas con la altura (Figura 7).

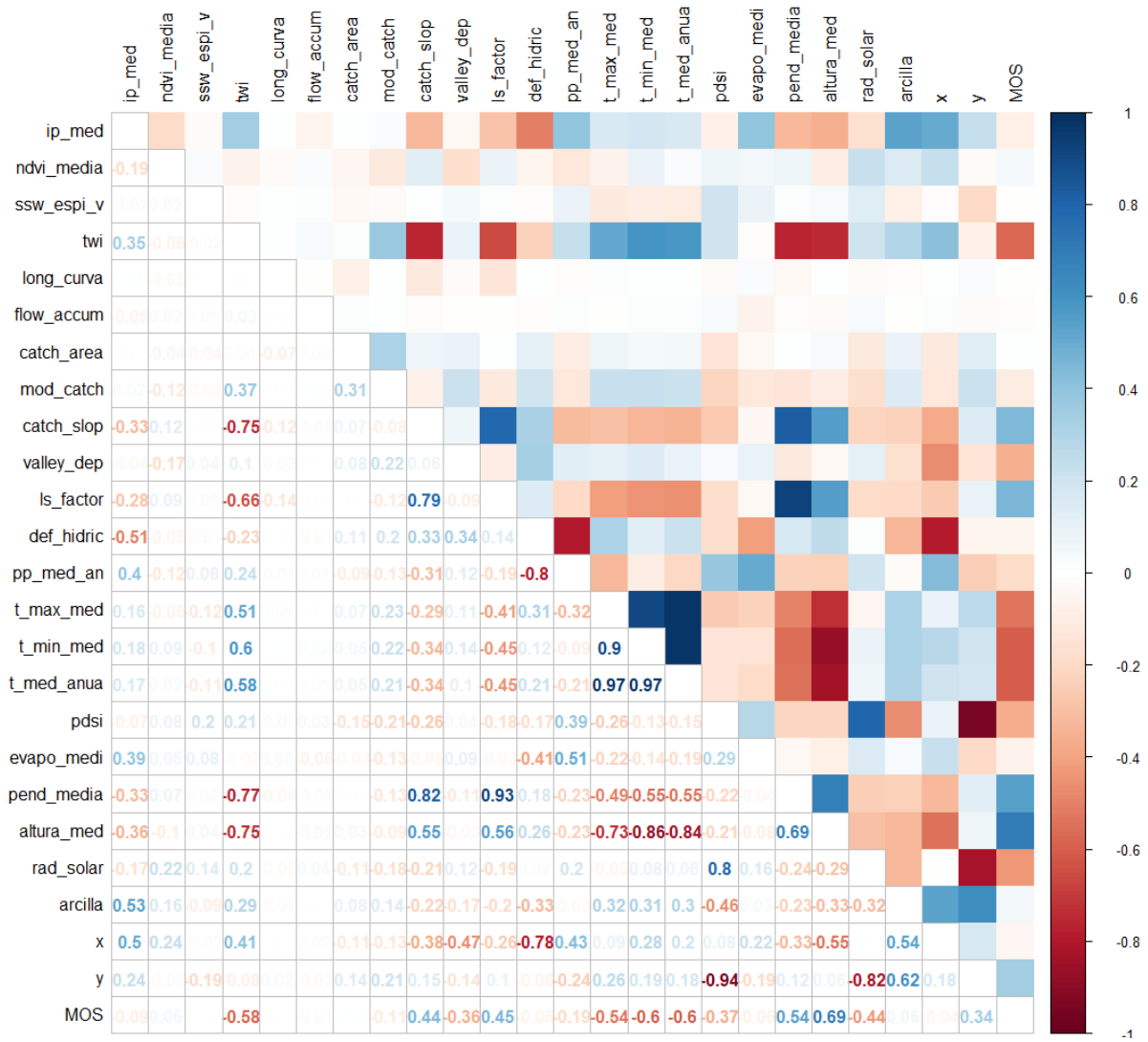


Figura 7: Correlograma para base de datos MOS de Córdoba. En la diagonal inferior de la matriz pueden observarse los coeficientes de correlación de Pearson, mientras que en la diagonal superior se esquematiza con escala de colores la magnitud y el sentido de la correlación.

La base de datos de Metales pesados presenta correlaciones entre variables asociadas a la distancia al río. Las variables relativas a las concentraciones de distintos metales se encuentran correlacionadas y, consecuentemente, las concentraciones de otros metales son potenciales variables explicativas de la variabilidad de zinc.

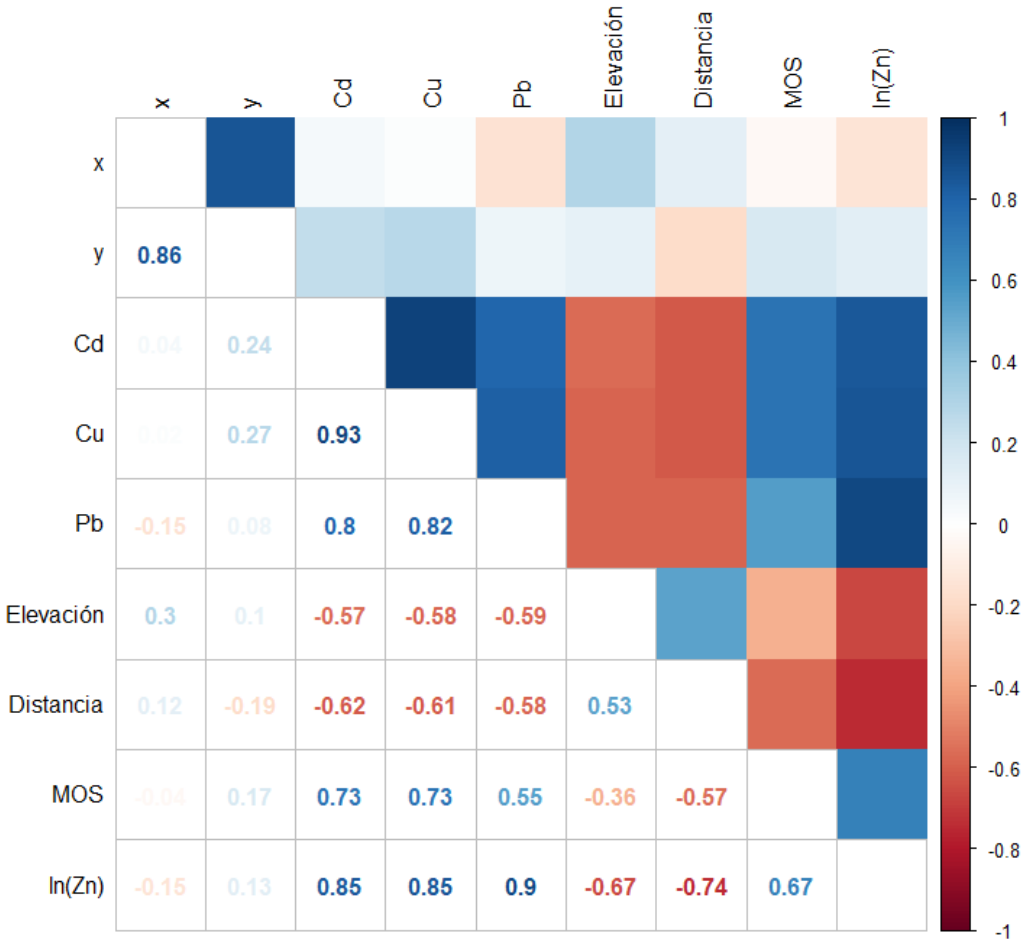


Figura 8: Correlograma para base de datos metales pesados río Meuse.

Para la base de datos de Adsorción de Atrazina se observan correlaciones entre elevación y variables texturales. Las variables de suelo que tienen que ver con la acción de organismos vivos como Carbono Orgánico del Suelo correlacionan con NT.

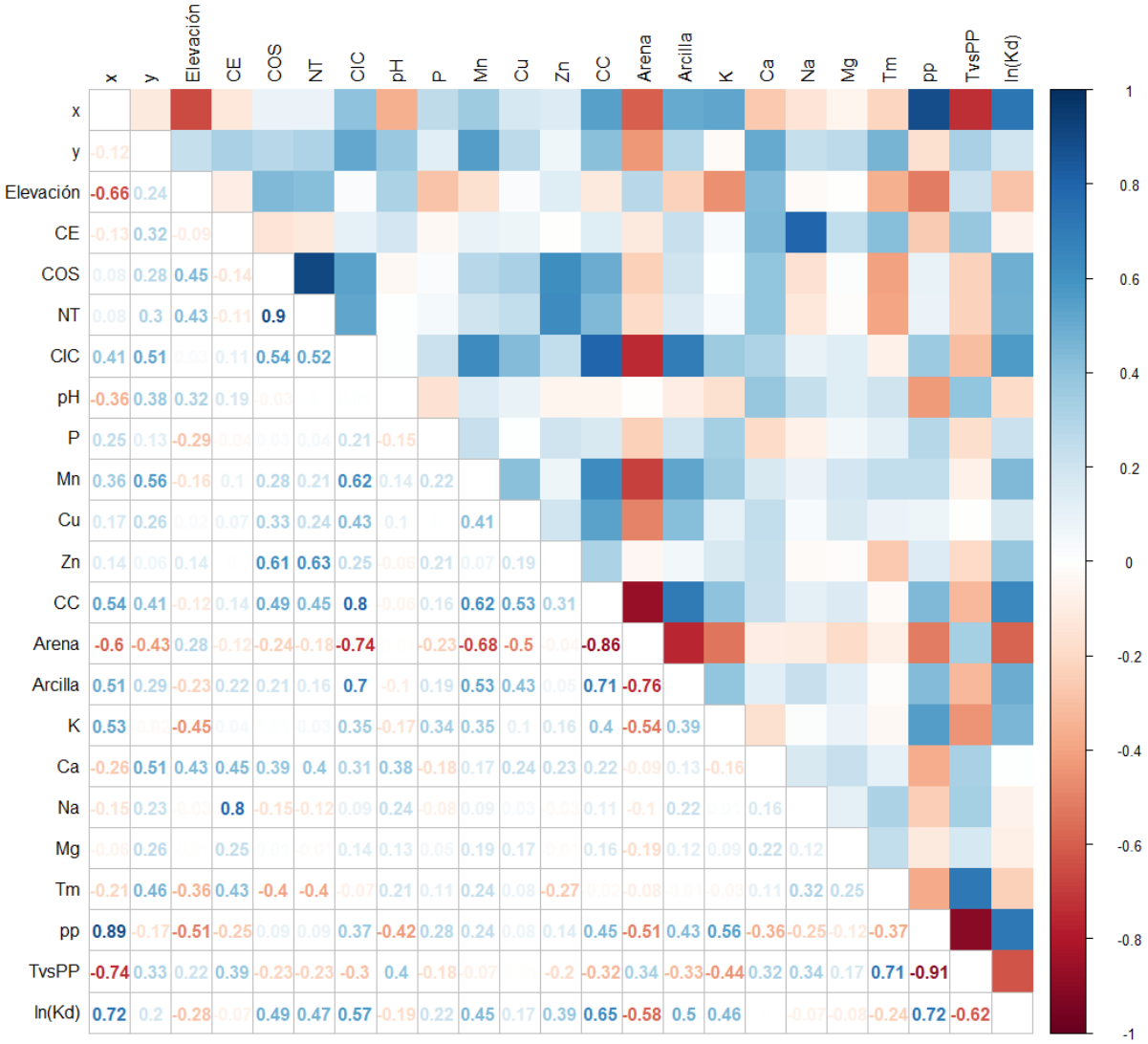


Figura 9: Correlograma para base de datos Coeficiente de Adsorción Atrazina.

La multicolinealidad (Kuhn y Johnson, 2013) en el contexto de modelos de regresión predictivos, usualmente se trabaja con la eliminación de variables redundantes si éstas quedan en el modelo luego de implementar procesos de selección de variables y disminución de la dimensionalidad por redundancia no implica pérdida significativa de la capacidad predictiva. En los modelos basados ensamble de árboles no presentan los mismos inconvenientes al trabajar con numerosas variables correlacionadas, por lo que para responder a un objetivo predictivo la selección de covariable no es determinante (Duffy y Helmbold, 2002; Efron y Hastie, 2016). Además, pueden trabajar con relaciones no lineales entre las covariables y la variable respuesta e interacciones complejas (Segal, 2004; Breiman et al., 2017). En la Figura 10 se muestran las relaciones parciales (Friedman, 2001) de las covariables que contribuyeron en mayor medida a explicar la variabilidad de cada variables respuestas utilizando el algoritmo RF.

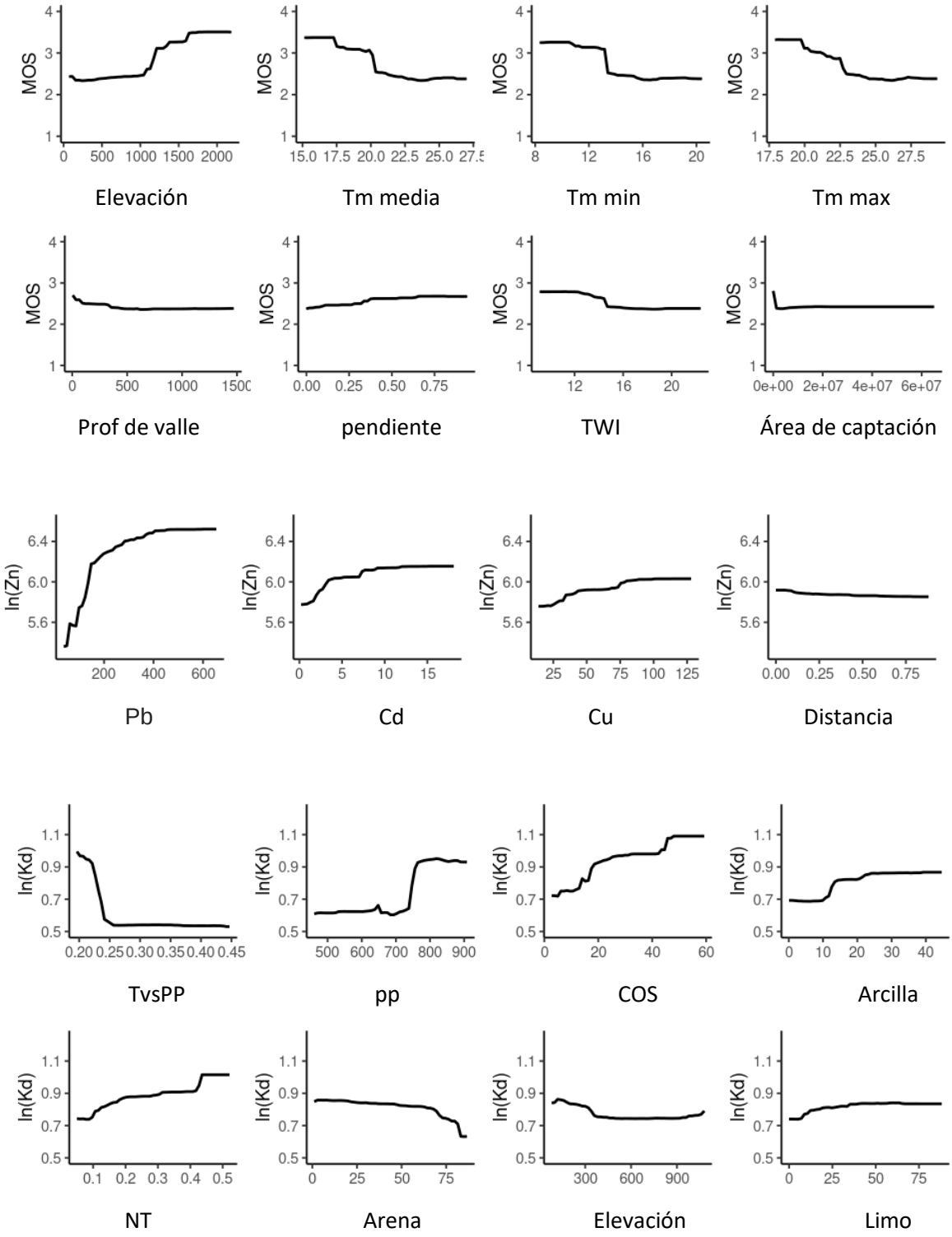


Figura 10: Relaciones parciales entre las principales covariables en explicar cada variable respuesta MOS, ln(Zn) y ln(Kd) a partir de regresiones por RF.

Los diagramas de dependencia o relaciones parciales de la relación parcial promedio entre el conjunto de predictores y la variable respuesta muestran que estas relaciones no son

necesariamente lineales. También se puede observar una concordancia con las correlaciones que se muestran en las Figura 7 a 9, tanto en magnitud como sentido.

Estructura espacial de las variables respuesta

Los índices de autocorrelación espacial fueron significativos ($p<0.0001$) tanto para el índice de Moran Global (IMG MOS=0.75; IMG ln(Zn)=0.60; IMG ln(Kd)= 0.74), como el índice de Geary (IG MOS=0.29; IG ln(Zn)=0.34; IG ln(Kd)= 0.25) en todas las bases de datos. El grado de estructuración espacial (RSV Ec [9]) se pone de manifiesto en la Figura 11.

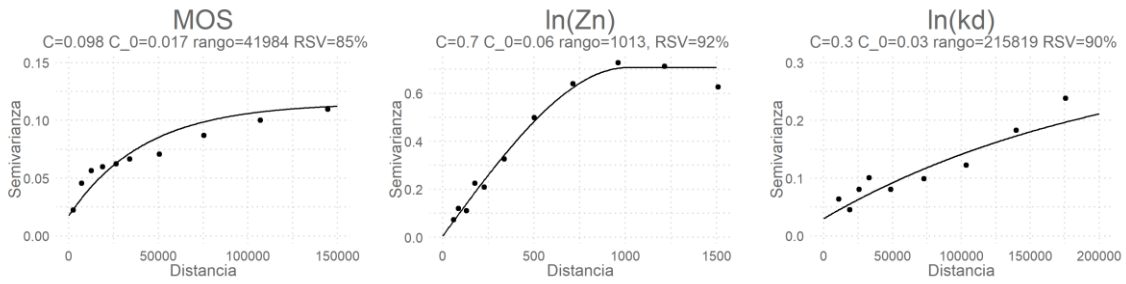


Figura 11: Semivariogramas empíricos y ajustados para MOS, ln(Zn) y ln(Kd).

Modelo de predicción espacial

Ajuste

El modelo (combinación de covariables) ajustado en cada base de datos se definió para cada algoritmo siguiendo protocolos/herramientas comunes de cada paradigma analítico. Así, se realizó una selección de variables de acuerdo al criterio DIC (por su término en inglés *"Deviance Information Criterion")* para RB (Wang et al., 2018) y una selección en base a la minimización de AIC (por su término en inglés *"Akaike Information Criterion")* en RK (West et al., 2014), mientras que la totalidad de las variables disponibles fueron usadas para ajustar el algoritmo RF .

La forma de comunicar la contribución de cada regresora a la descripción de la variabilidad de la variable respuesta también fue distinta para cada algoritmo. Para RB se presentan las medidas resúmenes de las distribuciones a posteriori marginales de los coeficientes del modelo de regresión ajustado, para RK se muestran las estimaciones de los coeficientes de regresión y para RF se muestra la importancia relativa a través de gráficos de relaciones parciales entre cada covariable y la variable respuesta. En la Tabla 4 se presentan los resultados de los modelos ajustados a partir de las covariables incluidas dentro de las regresiones lineales en RB y RK.

Se puede apreciar la similitud entre las medidas resúmenes de las distribuciones marginales a posteriori de los parámetros de RB con las estimaciones de los coeficientes de regresión en RK. Se

informan, para ambas regresiones lineales, las estimaciones de los hiperparámetros relativos a la estructura de covarianza espacial subyacente.

Tabla 4: Modelos de regresión de los algoritmos de predicción espacial: Regresión Bayesiana para datos espaciales (RB), Regresión Kriging (RK).

	RB			RK		
	Media	DE		β	EE	
			Materia Orgánica de Suelo			
Intercepto	10.683	2.492	Intercepto	7.876	1.655	*
IP	0.003	0.001	IP	0.004	0.001	*
Pendiente	-0.913	0.360	Pendiente	3.380	0.308	*
Deficit Hídrico	-0.074	0.008	Deficit Hídrico	-0.061	0.006	*
pp	-0.005	0.001	pp	-0.004	0.000	*
Tmax	-0.322	0.084	Tmax	-0.064	0.034	
Tmin	0.556	0.077	Tmin	0.362	0.040	*
Elevación	0.0040	0.0000	Elevación	0.0032	0.0002	*
			Arcilla	0.016	0.006	*
			ESPI	0.000	0.000	*
Sill	0.270	0.034	Sill	0.331		
Rango	30568.8	4189.9	Rango	17669		
			Metales pesados			
Intercepto	4.832	15.812	Intercepto	6.70839	0.05918	*
Distancia	-2.244	0.362	Distancia	-1.82799	0.24522	*
Frecuencia 2	1.435	15.811	Frecuencia 2	-0.54154	0.08594	*
Frecuencia 3	1.427	15.812	Frecuencia 3	-0.55558	0.10579	*
Suelo 2	-0.212	0.109	Suelo 2	-0.43338	0.09393	*
Suelo 3	-0.085	0.17	Suelo 3	-0.06599	0.1618	
Sill	0.199	0.043	Sill	0.220		
Rango	400.5	102.9	Rango	367.2		
			Adsorción Atrazina			
Intercepto	-1.419	1.954	Intercepto	-4.220	1.104	*
Elevación	-0.001	0.000	Elevación	-0.0004	0.0002	*
COS	0.037	0.004	COS	0.041	0.009	*
pp	0.003	0.002	pp	0.004	0.001	*
TvsPP	-0.803	2.898	TvsPP	3.225	1.418	*
Arcilla	0.006	0.003	NT	1.192	0.877	
			pH	0.108	0.044	*
			P	-0.001	0.001	
			Cu	-0.104	0.029	*
			Arena	-0.004	0.002	*
Sill	0.153	0.104	Sill	0.079		
Rango	342166	20231	Rango	70515		

* Efecto estadisticamente significativo para un α=0.05

Las mallas de predicción usadas para predicción del efecto espacial en RB se muestran en la Figura 12 junto con la proyección del efecto sitio sobre la grilla de predicción. Esta herramienta permite

evaluar el suavizado que se realiza en la dimensión de las coordenadas de sitio (Miller, Glennie y Seaton, 2020).

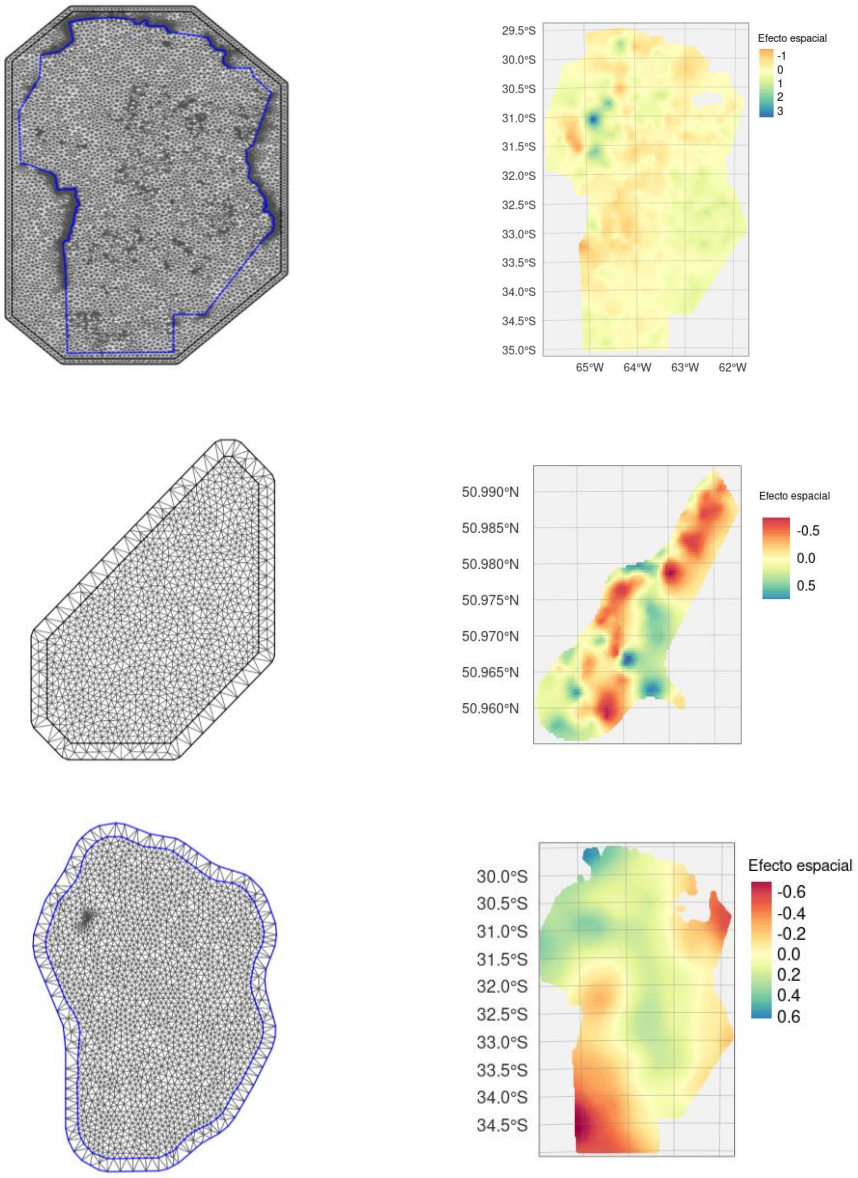


Figura 12: Estimación del efecto aleatorio de sitio para la Regresión Bayesiana (RB). Mallas de predicción (izq). Proyección del efecto espacial sobre la grilla de predicción (der.).

Una de las mayores ventajas de la implementación de la RB en el contexto del MDS es la obtención directa de las medidas de incertidumbre sitio específicas que se obtienen a partir de las distribuciones a posteriori predichas de cada sitio en la grilla de predicción. La Figura 13 muestra las predicciones y las medidas de credibilidad de estas predicciones para el modelo RB en cada base de datos. La predicción puntual corresponde a la media de la distribución posteriori predicha para cada sitio, el desvío estándar informado proviene de esa misma distribución a posteriori de valores

predichos y la amplitud del intervalo de credibilidad del 95% se obtuvo a partir de los percentiles 0.025 y 0.975 de la misma distribución.

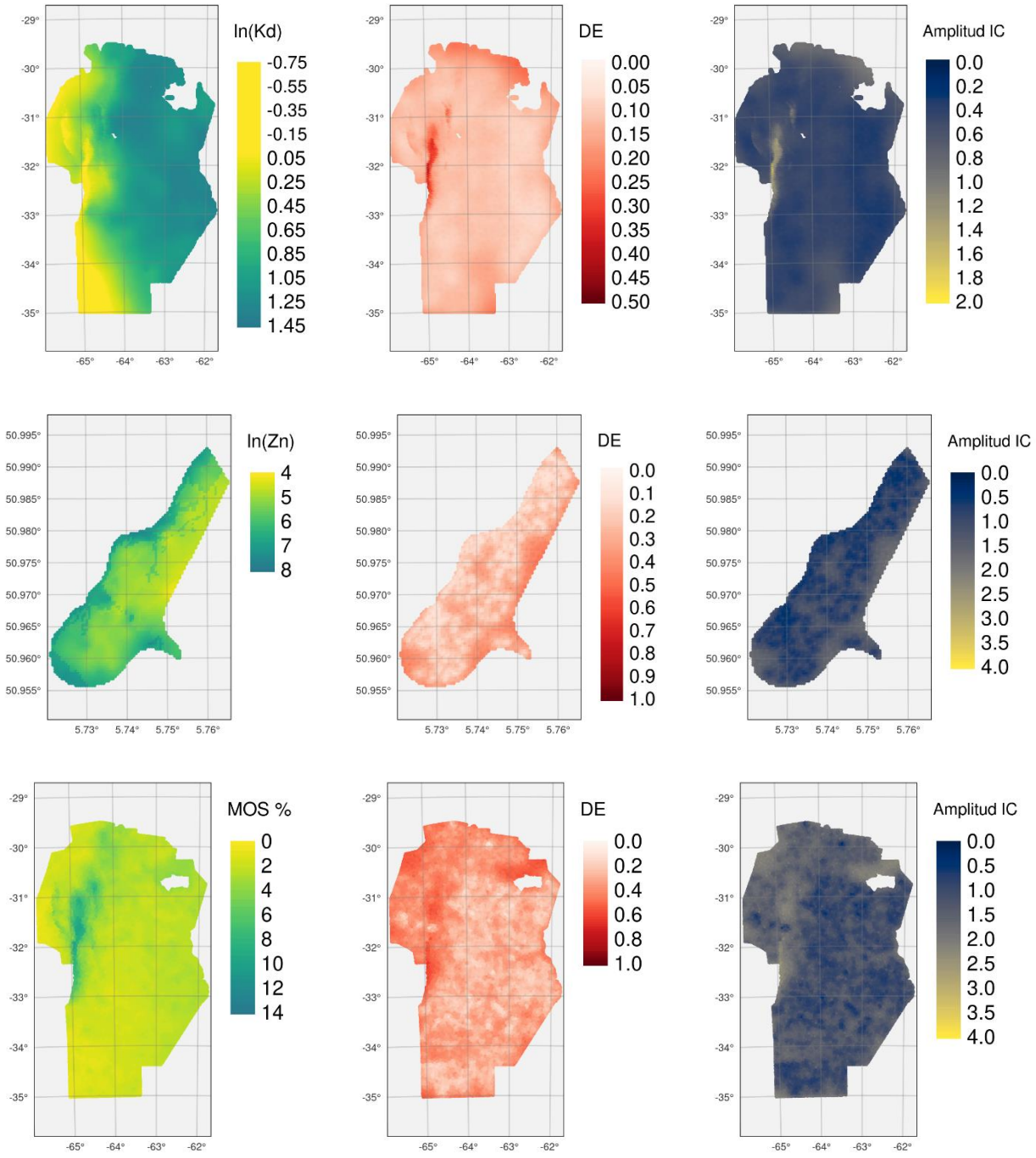


Figura 13: Predicciones y credibilidad de la predicción derivadas de la regresión bayesiana

Cuando se usa la RB para variables respuestas transformadas como ln(Zn) y ln(kda), será posible aplicar directamente la operación inversa para informar los valores en la escala real de la variable (Correa Morales et al., 2018).

Es importante destacar que los residuos de los modelos no mostraron autocorrelación espacial significativa para ninguno de los tres algoritmos en ninguna de las bases de datos.

Mapeo digital e interpretación ambiental de las predicciones

En las Figura 14, 15 y 16 se presenta el mapeo digital realizado con RB, RK y RF para las tres bases de datos. La distribución espacial de las predicciones es similar para los algoritmos RB y RK. Observando la escala de valores, queda de manifiesto que el algoritmo RF muestra predicciones con menor variabilidad en los valores predichos comparado a RB y RK, en las tres bases de datos. También se presentan los mapas de incertidumbre, para RK y RF estas estimaciones del error provienen de la varianza kriging de los residuos del modelo, para RB se trata del desvío estándar de la distribución a posteriori predicha por el modelo.

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% menores 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 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). Para las implementaciones ilustradas en este capítulo la proporción se sitios para los cuales el valor observado estuvo incluido en los intervalos definidos por cada incertidumbre vario entre algoritmos. RB mostro mayor cantidad de sitios en los cuales el valor observado estuvo incluido dentro del intervalos de credibilidad del 95% (MOS=70%; ln(Zn)=88%, ln(Kd)=69%) de probabilidad, en comparación con los intervalos de confianza construidos con α=0.05 para los algoritmos RK (MOS=63%; ln(Zn)=85%, ln(Kd)=66%) y RF(MOS=43%; ln(Zn)=75%, ln(Kd)=46%).

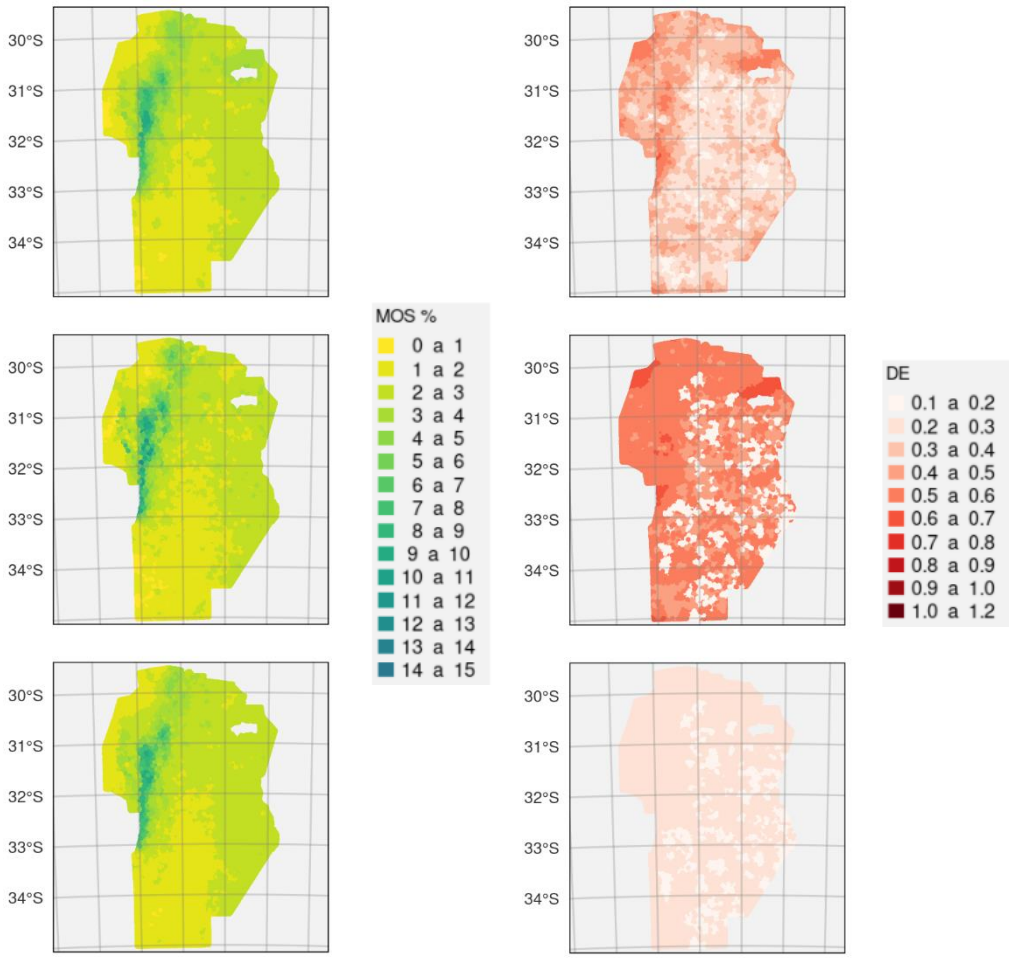


Figura 14: Mapeo digital de MOS en base a tres algoritmos de predicción espacial. De arriba hacia abajo Regresión bayesiana, Regresión Kriging y Random Forest. En la columna de la izquierda se muestran las predicciones puntuales y en la columna de la derecha las medidas de incertidumbre de la predicción.

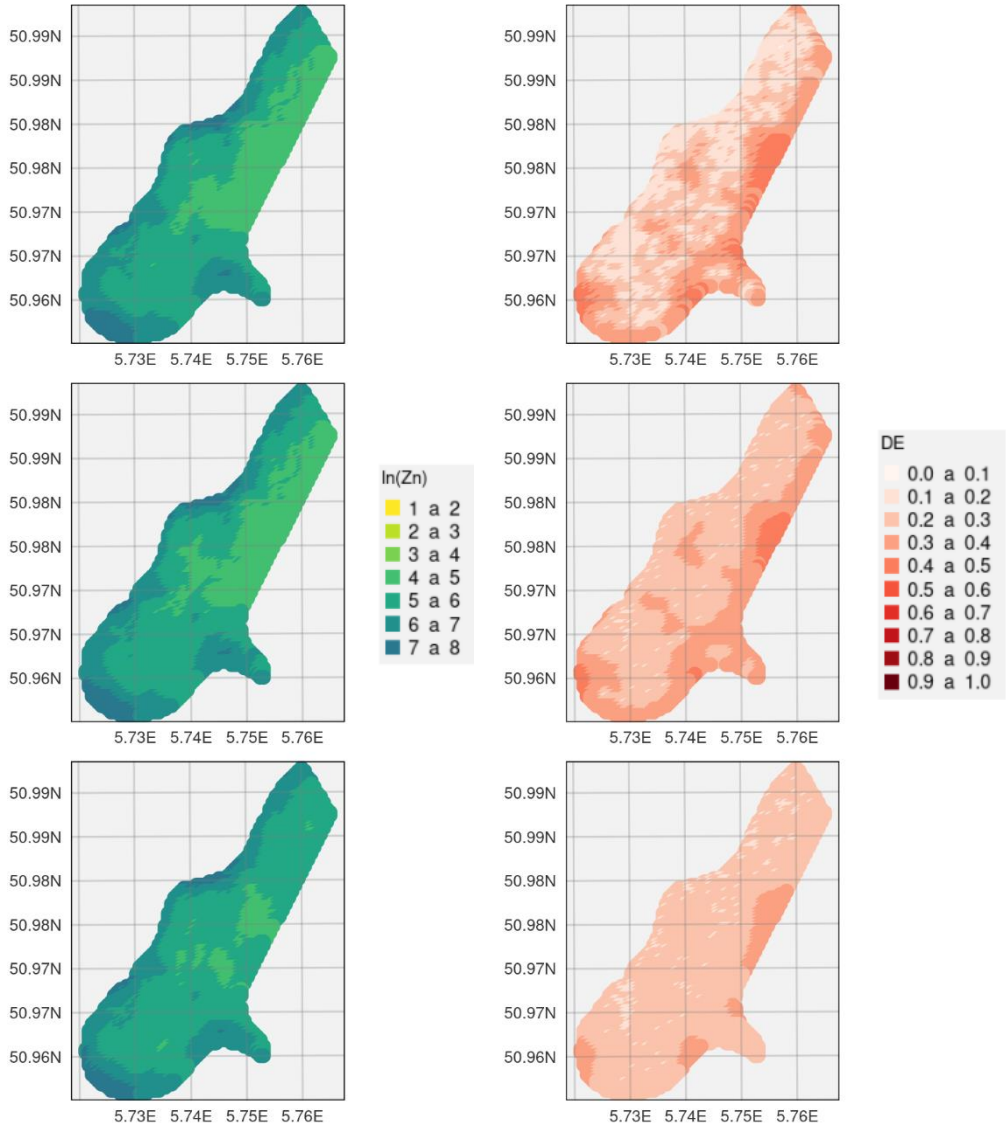


Figura 15: Mapeo digital de ln(Zn) en base a tres algoritmos de predicción espacial. De arriba hacia abajo Regresión bayesiana, Regresión Kriging y Random Forest. En la columna de la izquierda se muestran las predicciones puntuales y en la columna de la derecha las medidas de incertidumbre de la predicción.

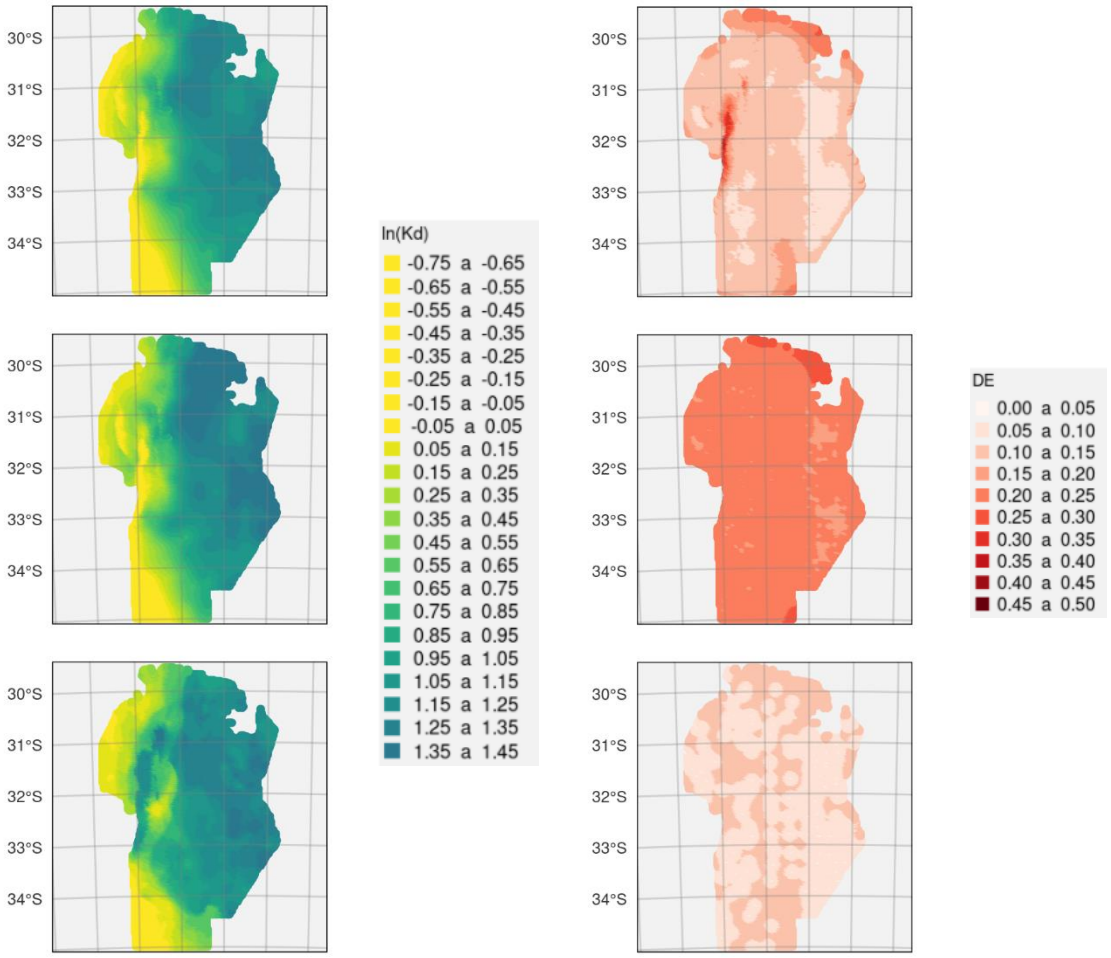


Figura 16: Mapeo digital de ln(Kd) en base a tres algoritmos de predicción espacial. De arriba hacia abajo Regresión bayesiana, Regresión Kriging y Random Forest. En la columna de la izquierda se muestran las predicciones puntuales y en la columna de la derecha las medidas de incertidumbre de la predicción.

Si bien se evidencian diferencias respecto a la implementación de los algoritmos RB, RK y RF, los resultados de las predicciones sitio específicas resultaron similares. Las principales diferencias encontradas entre métodos radican en las formas para medir la incertidumbre de las predicciones sitio específicas y la manera de incorporar la correlación espacial en las estimaciones. De esta comparación emerge la necesidad de comparar el desempeño predictivo de las alternativas implementadas bajo un diseño bajo el cual no se favorezca a alguno de los tres algoritmos evaluados. En el capítulo IV del presente trabajo se propone una evaluación del desempeño

predictivo de RB, RK y RF para las tres bases de ilustración bajo un diseño experimental con estas características.

Capítulo IV

Comparación de modelos de predicción espacial bajo distintos escenarios. Una aproximación por simulación

En este capítulo se realiza una investigación metodológica-estadística para evaluar el desempeño de modelos de predicción espacial en mapeos digitales bajo distintos escenarios en relación con el número de covariables intervinientes y los tamaños muestrales disponibles para el ajuste del modelo de predicción espacial. Los modelos son estimados desde distintos enfoques para el tratamiento de datos espaciales: regresión *kriging* (RK), *random forest* con residuos krigeados (RF), y regresión bayesiana con INLA-SPDE (RB). Se comparan los resultados obtenidos por los distintos predictores espaciales sobre escenarios definidos por simulaciones realizadas sobre las tres bases de datos de suelo descriptas en el capítulo anterior.

Diseño del estudio. Configuraciones o escenarios de comparación

El desempeño predictivo de los modelos se evaluó según un diseño que propone por un lado variar la configuración de variables explicativas y por otro el tamaño muestral. Para cada base de datos se configuró un diseño factorial de tres factores enunciados a continuación:

- Algoritmo de predicción espacial ($f(\quad)$): RK, RF, RB
- Dimensionalidad (p): número de covariables. 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.
- 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).

Criterios de comparación

Para la evaluación de la capacidad predictiva en cada configuración (combinación de los tres factores $f(\ \ ), p$ y n) se realizaron 30 repeticiones. Se utilizaron n datos para entrenar cada uno de los algoritmos $f(\ \ )$ a partir de las p de covariables definidas en cada una de las combinaciones posibles de variables. 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. Las medidas de capacidad predictiva que se evaluaron sobre los grupos de validación fueron:

Error de predicción global

Se utilizó la raíz cuadrada del error cuadrático medio de predicción expresado relativo a la media general y su medida de variabilidad de este como el desvío estándar de RMSPE.

$$\mathrm{EP} = \frac{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}}{\overline{Y}} \times 100$$

donde y_i es el valor de la variable respuesta observada en el sitio de validación i ; $\hat{y}_i$ es el valor predicho para ese sitio; n es el número de sitos en el grupo de validación ($n = 50$) e $\overline{Y}$ es el valor de medio de la variable respuesta en el set de validación.

Error sitio específico

El error de predicción puntual, al que se denomina Error sitio-específico (ESE) se expresó como un porcentaje de la media del sitio y se calculó como:

$$ESE = \frac{|y_i - \hat{y}_i|}{y_i} \times 100$$

También se calculó la proporción de sitios con errores de predicción por debajo del 10% (ESE bajo), entre el 10% y el 30% d (ESE medio) y por encima del 30% (ESE alto).

Resultados y Discusión

Desempeño de modelos predictivo en relación con p y n

Error global de predicción

El comportamiento de los diferentes algoritmos de predicción espacial en relación con el error de predicción promedio fue mejor al aumentar el tamaño muestral (*n)* principalmente en la base de datos de materia orgánica del suelo donde los cambios en tamaño de muestra entre escenarios fueron mayores (Figura 17). Si se analizan las tendencias en el error global de predicción respecto a la cantidad de variables explicativas incluidas en la regresión (Figura 17) puede observarse que (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 p para el algoritmo RF que no se evidencia en los modelos RK y RB.

Para evaluar 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, las correspondientes interacciones dobles y la interacción triple (Tabla 6 en Anexos).

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 $f(\ \ )$, 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.

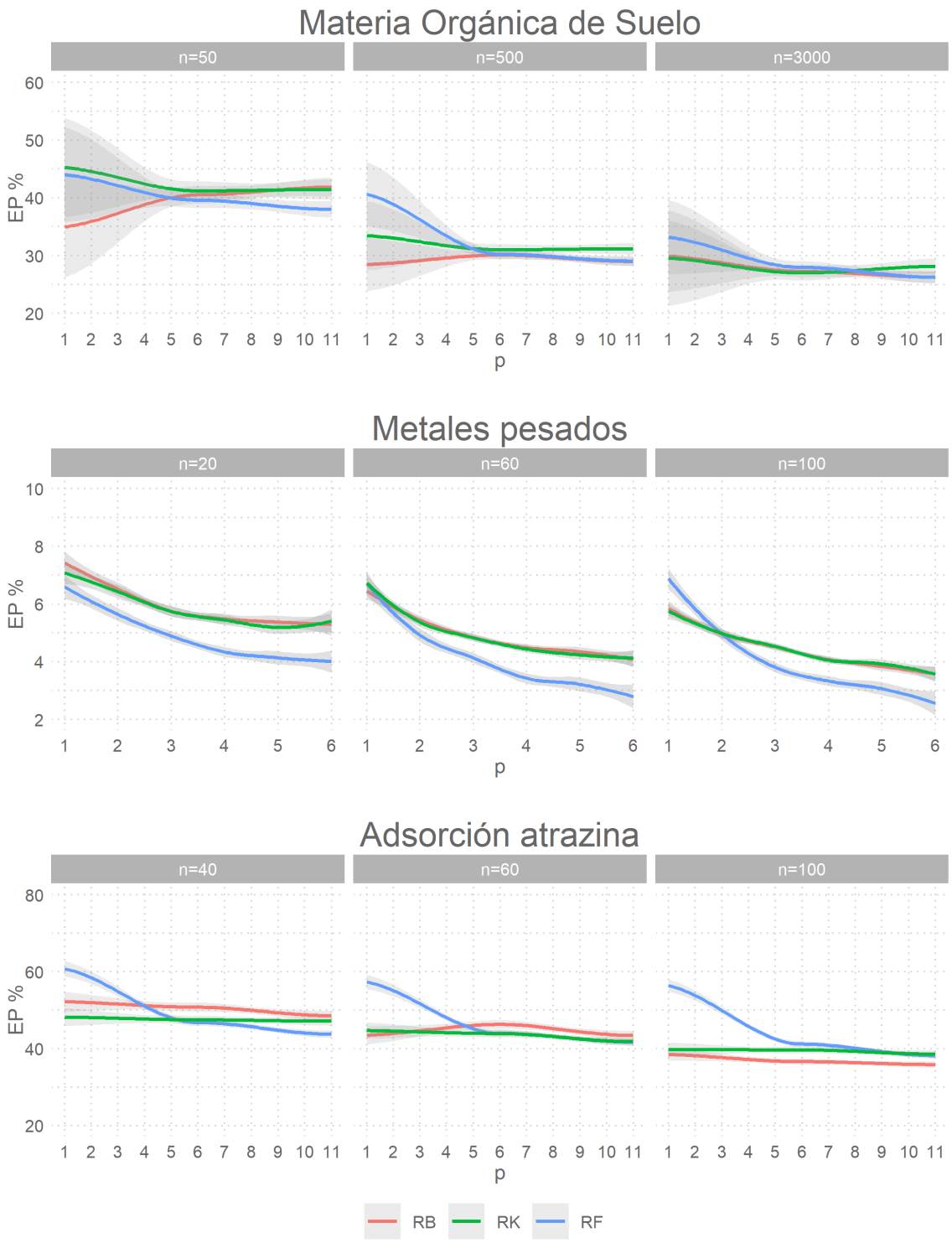


Figura 17: Error de Predicción Global (EP%) expresado como porcentaje de la media predicha para cada sitio por tres algoritmos de predicción espacial: Regresión Bayesiana para datos espaciales (RB), Regresión Kriging (RK) y Random Forest con residuos krigeados (RF) frente a diferentes escenarios configurados según cantidad de variables explicativas (p) y tamaño muestral usado en la estimación (n).

Se realizó un contraste para evaluar la performance de RK y RB con respecto a RF para un escenario de p=2 covariables y luego para p=7 con el menor tamaño muestral en cada base de datos (Figura 18). El valor p del contraste fue ajustado por SIDAK ya que el mismo fue postulado luego de observar las tendencias en la Figura 17. Los resultados del contraste confirman que para un contexto de dos covariables RK y RB producen menor o igual error que RF (MOS:p=0.0730; ln(Zn):p=0.0695; ln(Kd):p=0.0001), pero no en un contexto de mayor p en donde RF registró menor error de predicción (MOS: p<0.0001; ln(Zn):p= p<0.0001; ln(Kd):p<0.0001). En la Figura 18 se muestran las comparaciones de media realizadas luego del ANAVA.

Errores Sitio Específicos

Como medida del error puntual o sitio especifico de las predicciones se presenta la distribución empírica de los errores sitio específicos (ESE%) para el menor tamaño muestral testeado en cada base de datos. Para MOS y ln(Kda) se confirma un mejor desempeño de RB y RK RF comparado con para una dimensionalidad de $p = 2$ (Figura 19). En cambio, los percentiles de la distribución de los ESE de RF son menores para $p = 7$. Es decir, el aumento de la dimensionalidad favorece el desempeño de RF no solo en el desempeño de global de la predicción, sino que predice mayor cantidad de sitios son ESE bajo.

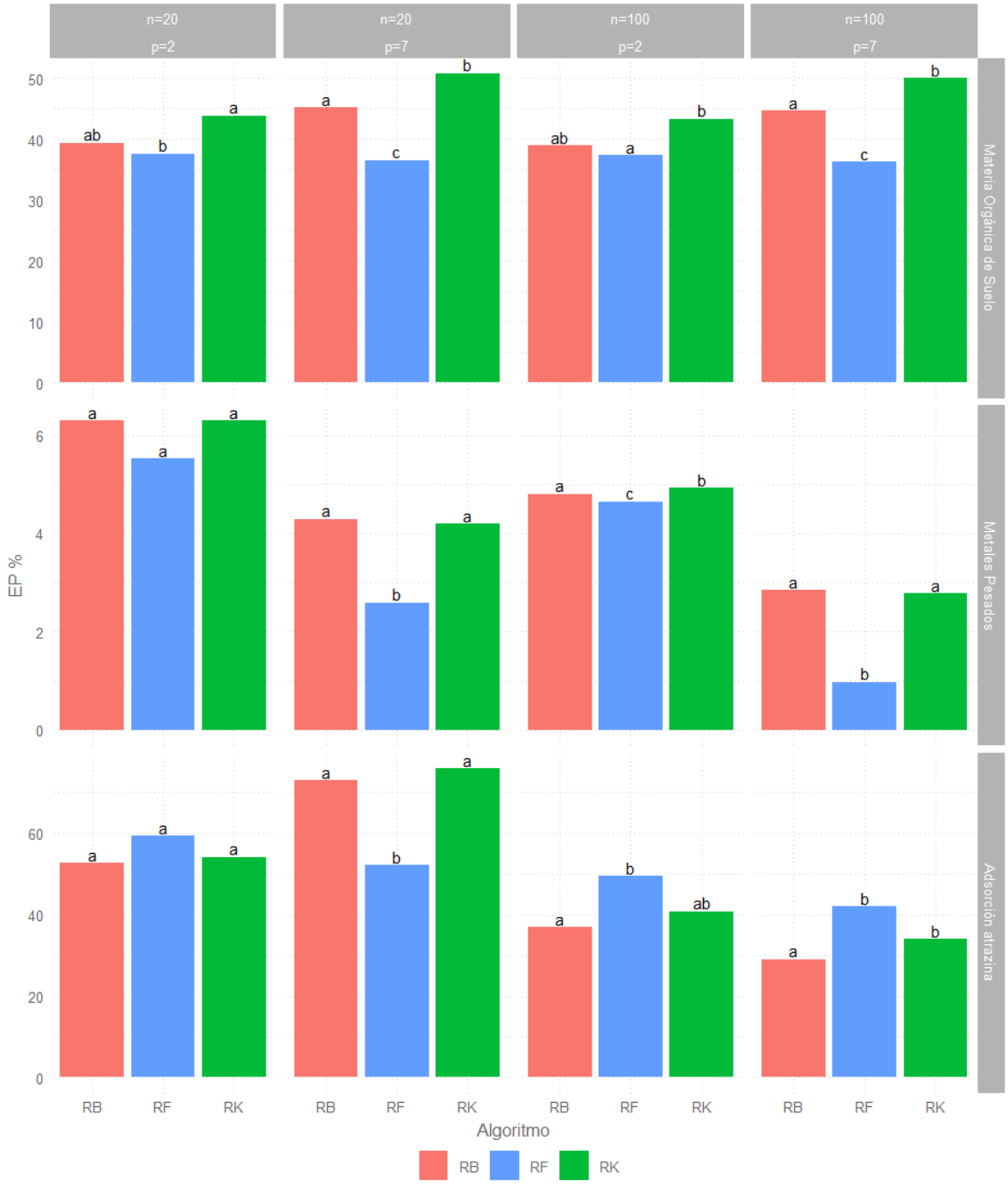


Figura 18: Error de Predicción Global (EP%) expresado como porcentaje de la media predicha para cada sitio por tres algoritmos de predicción espacial: Regresión Bayesiana para datos espaciales (RB), Regresión Kriging (RK) y Random Forest con residuos krigeados (RF) para $\boldsymbol{p = 2}$ y $\boldsymbol{p = 7}$ variables explicativas y $\boldsymbol{n = 20}$ y $\boldsymbol{n = 100}$ tamaño muestral. Letras diferentes implican medias estadísticamente diferentes para los tres algoritmos en cada escenario $\boldsymbol{n}$ y $\boldsymbol{p}$ obtenidos de Análisis de Varianza realizados para cada base de datos.

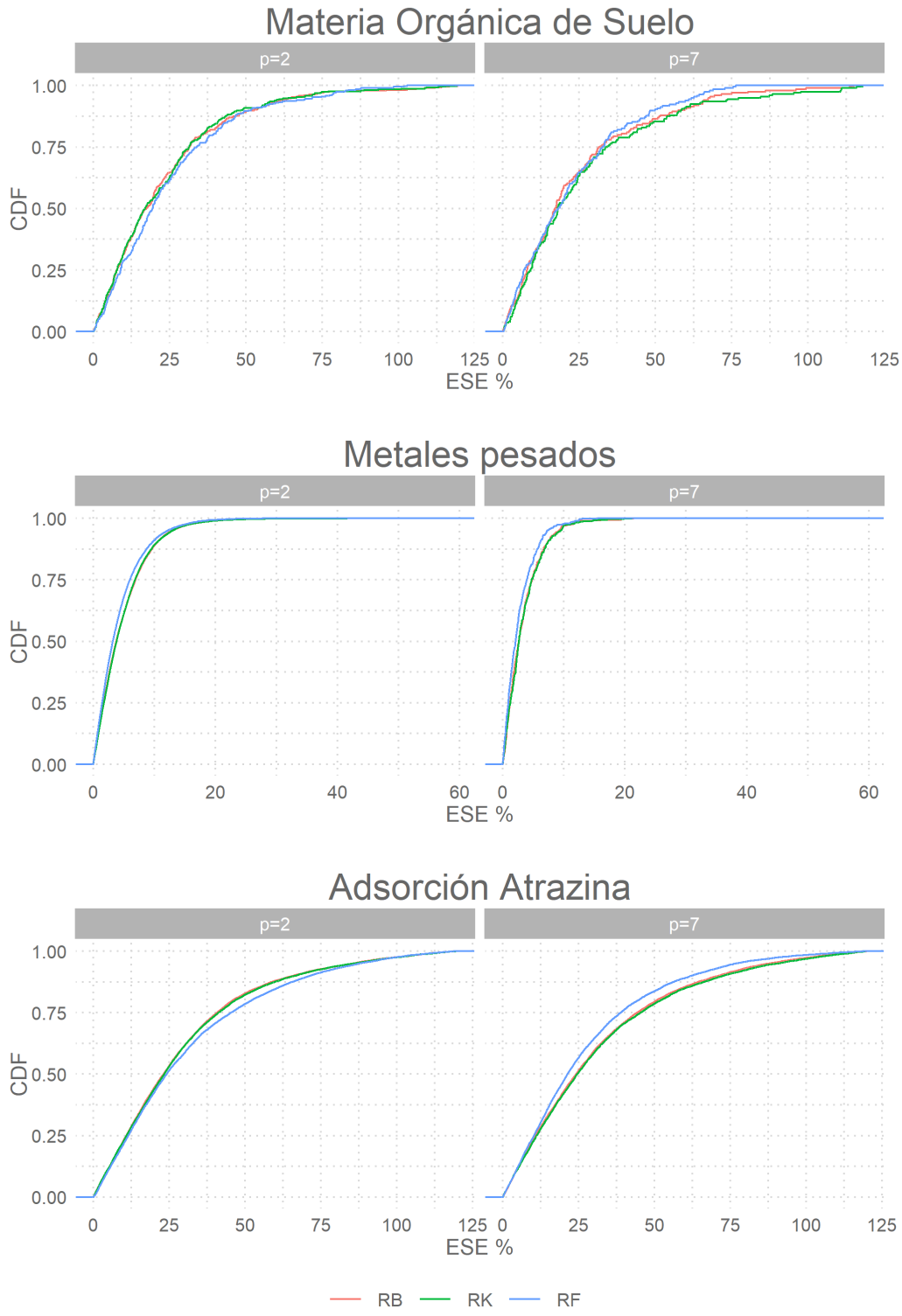


Figura 19: Errores Sitio Específicos (ESE %) expresado como porcentaje de la media observada en cada sitio por tres algoritmos de predicción espacial: Regresión Bayesiana para datos espaciales (RB), Regresión Kriging (RK) y Random Forest con residuos krigeados (RF) frente a diferentes escenarios configurados según cantidad de variables explicativas (p).

Representación de factores SCOPRPAN en los modelos de mejor ajuste

A continuación, se presenta el mejor modelo estimado por cada algoritmo. Es decir, la combinación de p y n que obtuvieron los mejores desempeños predictivos. Las mejores configuraciones logran errores de predicción globales menores al 25% en relación a la media, siendo el *ln(Kda)* la variable con mayor error de predicción, seguida por MOS y por último el *ln(Zinc)*. Estas diferencias en los errores globales, no se deben a la cantidad de covariables explicativas o tamaños muestrales, pero si se corresponden con la variabilidad en relación a la media de las variables respuestas (CV ln(Kda)=71%> CV MOS=12%> CV ln(Zn)=12%). Los menores errores de se obtienen con el algoritmo RB para MOS y ln(Kda) y con RF para el ln(Zn).

En la Tabla 5 se presenta la proporción de aparición de variables explicativas en las 100 mejores configuraciones para cada modelo según el factor SCORPAN al que hacen referencia. Las covariables edáficas existentes en cada base de datos se encuentran siempre presentes en las 100 mejores configuraciones, lo que se corresponde con la estructura de correlaciones encontrada con anterioridad. El factor topografía adquiere especial relevancia para explicar MOS y ln(zinc) y aparece en menor medida como factor explicativo de ln(Kda) (Tabla 5).

El proceso de adsorción de una molécula al suelo es un fenómeno de superficie determinado por la interacción de la molécula con características físicas y químicas del suelo (Cheng, 1990; Weber et al., 2004) por lo que la acción de la topografía como factor formador influye de manera indirecta comparado con otras características edáficas como la concentración de MOS o la concentración de Zn. El Kd de atrazina en suelo es un atributo que se puede englobar dentro de las variables que definidas como *funciones de suelo* que ofrecen nuevos desafíos para el MDS principalmente debido a la influencia de las covariables clásicas no es directa (Minasny y McBratney, 2016; Styc y Lagacherie, 2019).

La variabilidad en la concentración de metales pesados contaminantes que no son parte del material parental de los suelos generalmente se explica por los procesos de transporte a los que se ven afectados por lo que en este contexto particular de metales en el Rio Meuse donde la fuente de contaminación es el agua su dinámica está determinada por variables topográficas es de esperar que las mismas tengan predominancia a la hora de explicar el proceso (Burrough & Swindell, 1997; Grifoll & Cohen, 1996). Es importante tener en cuenta que en el grupo de variables explicativas hay concentraciones de otros metales pesados además de Zinc para los cuales los procesos que dan origen a la variabilidad son los mismos. Se conoce que la variabilidad de MOS depende de múltiples

factores formadores, sensible a diversas variables, por eso también se considera en un potente indicador y descriptor de calidad de suelo (Milne, 2009; Yigini et al., 2018).

Tabla 5: 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 *			
				Clima	Aorg	Suelo	Topografía
Materia Orgánica de Suelo n=3000	RB	16.42	11	0.99	0.71	1	1
	RK	17.02	11	0.96	0.78	1	1
	RF	15.13	16	0.99	0.7	1	1
Metales pesados n=100	RB	2.31	4	-	0.62	1	0.89
	RK	2.24	2	-	0.54	1	0.84
	RF	1.40	6	-	0.66	1	0.81
Adsorción Atrazina n=100	RB	20.39	6	0.8	-	0.99	0.43
	RK	20.41	6	0.97	-	1	0.5
	RF	23.04	11	0.85	-	1	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. 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.

Capítulo V

Comentarios Finales

Desafíos estadísticos metodológicos en mapeo digital de suelos

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. 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. La predicción espacial de una variable continua implica una serie de desafíos metodológicos-estadísticos. 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. 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. 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. 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. En este trabajo de tesis se presentan 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 (el código ejemplo de la implementación en R se presenta en el Anexo).

En primer lugar, se introdujo el MDS y se abordaron los fundamentos estadísticos teóricos para la modelación de datos geoestadísticos. Luego se desarrolló el marco conceptual que soporta la

modelación espacial a través de la inferencia bayesiana utilizando INLA y SPDE. 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 (Piumetto, García y Morales, 2018); 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., 2019a). Los resultados de la implementación con RB se compararon con otros dos algoritmos utilizados en la literatura moderna de MDS, Regresión Kriging (RK) (Hengl et al., 2004) y Random Forest con residuos krigeados (RF) (Breiman, 2001; Li et al., 2011). Finalmente, 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.

Relevancia de las contribuciones. Algoritmos de predicción espacial ventajas y desventajas

Los modelos de regresión que consideran la variabilidad espacial subyacente resultan más eficaces que aquellos que tratan los datos como independientes. Los resultados de esta tesis 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 son 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. 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 Matérn 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, por esto hay autores que afirman que es un método más eficiente desde lo computacional. En este estudio la conveniencia de RB respecto a RK y RL en términos computacionales solo fue evidente en el contexto del mapeo digital de MOS para la provincia de Córdoba donde la predicción por RF y RL supero en más de 4 minutos a la predicción por RB.

La predicción sitio específica corresponde a una medida resumen de posición de la distribución conjunta a posteriori predicha en cada sitio. De la misma distribución de densidad se obtienen las medidas de incertidumbre de cada predicción. Estas particularidades posicionan a la regresión bayesiana como una buena alternativa comparada a otros métodos principalmente en la cuantificación de la incertidumbre. Como en la regresión lineal frecuentista, resulta conveniente realizar una selección de predictores para reducir la dimensionalidad y evitar problemas de colinealidad. En este sentido el algoritmo RF es el más práctico ya que trabaja sin inconveniente con gran número de covariables, incluso estando estas altamente correlacionadas. También es importante destacar que utilizando este algoritmo basado en aprendizaje de máquinas no es posible contar con una estimación del efecto que cada predictora tiene sobre la variable respuesta. Con RB se obtiene una estimación de la distribución marginal a posteriori para cada parámetro e hiperparámetro del modelo dado las observaciones y con RK se obtiene una estimación puntual para cada parámetro de los efectos fijos del modelo. No obstante, con RF es posible describir las relaciones parciales da cada covariable con la variable respuesta, lo cual puede ser una alternativa válida frente a la imposibilidad de conocer la contribución en magnitud y sentido de cada predictora.

Este trabajo también confirma que 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, en las bases de datos aquí analizadas este comportamiento es constante entre los algoritmos RB, RK y RF. No ocurre lo mismo con la cantidad de covariables implicadas en el modelo, es decir el número de parámetros a estimar. En este sentido el aumento en el número de variables predictoras tiene un impacto diferencial para en el algoritmo basado en arboles de regresión y clasificación, RF, obteniéndose mejores desempeños comparados con RB y RK en contextos de alta dimensionalidad. Las mejores configuraciones 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. 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.

Futuras líneas de investigación

Para el contexto de este trabajo se ha profundizado en la medición de los errores de predicción, pero aún es necesario profundizar el conocimiento respecto al dimensionamiento y la propagación de la incertidumbre en las predicciones espaciales generadas por regresiones bayesianas y los otros métodos. A su vez se puede llegar a conocer mediante estudios de simulación y sensibilidad como impactan las covariables en el dimensionamiento de la incertidumbre. Por otro lado, las implementaciones bayesianas se aplican aquí como estrategias estadístico-computacional para la predicción espacial pero aún queda por abordar como utilizar el enfoque bayesiano incluyendo el conocimiento a priori que se tiene sobre los procesos ambientales que determinan la variabilidad de las propiedades edáficas estudiadas, más allá de las que proponen las covariables incluidas en el modelo. Este aspecto adquiere especial relevancia a la hora de modelar variables que describen funciones del suelo en las que se encuentran implicadas interacciones de mayor orden comparadas a los clásicos atributos de suelo, como fue el caso en ente estudio de la modelación del proceso de adsorción de atrazina al suelo. Por último, así como en la regresión RF y RL se modela la estructura de covarianza sobre los residuos a través del ajuste de un semivariaograma y el posterior krigeado sobre los residuos, algunos autores proponen incorporar la metodología SPDE (configuración de la malla por triangulación y estimación de la estructura de correlación espacial por SPDE) a otros algoritmos distintos de INLA.

Respecto a las potencialidades en general de la aplicación de los modelos jerárquicos bayesianos utilizando SPDE al mapeo digital de suelo, dada principalmente por la manera en la que construyen la matriz de varianzas y covarianzas, los modelos bayesianos ofrecen una serie de ventajas a la hora de describir la variabilidad espacial de datos no normales es decir en el contexto de modelos generalizados mixtos donde la estimación del efecto aleatorio de sitio no es factible de realizarse en dos pasos como en los ejemplos aquí abordados. Si bien este trabajo se concentra en herramientas para modelar datos geoestadísticos dentro del MDS también surgen interrogantes que demandan la modelación de variables categóricas como las unidades taxonómicas o las

capacidades de uso de los suelos con mucho menor desarrollo. Las aplicaciones de regresiones espaciotemporales con INLA utilizando SPDE han sido exitosamente aplicados a la modelación de procesos espaciotemporales en otros estudios ambientales. También, el abordaje en el reconocimiento no solo los patrones espaciales sino temporales, es un tópico de creciente relevancia en el MDS. Además, las propiedades edáficas frecuentemente se miden a distintas profundidades dentro de un perfil, generalmente los patrones de variación en profundidad se describen a partir de funciones de modelos no-lineales cuyos parámetros pueden ser mapeados utilizando técnicas del MDS.

Bibliografía

Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

Arrouays, D., Grundy, M. G., Hartemink, A. E., Hempel, J. W., Heuvelink, G. B. M., Hong, S. Y., Lagacherie, P., Lelyk, G., McBratney, A. B., & McKenzie, N. J. (2014). GlobalSoilMap: Toward a fine-resolution global grid of soil properties. In *Advances in agronomy* (Vol. 125, pp. 93–134). Elsevier.

Bakka, H., Rue, H., Fuglstad, G. A., Riebler, A., Bolin, D., Illian, J., Krainski, E., Simpson, D., & Lindgren, F. (2018). Spatial modeling with R-INLA: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, *February*, 1–24. https://doi.org/10.1002/wics.1443

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 192–236.

Besag, J. (1977). Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika*, 616–618.

Besag, J. (1981). On a system of two-dimensional recurrence equations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *43*(3), 302–309.

Besag, J., Green, P., Higdon, D., & Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science*, 3–41.

Best, N., Richardson, S., & Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, *14*(1), 35–59.

Bivand, R S, Gomes-Rubio, V., & Rue, H. (2015). Spatial Data Analysis with R - INLA with Some Extensions. *Journal of Statistical Software*, *63*(20), 1–31. https://doi.org/http://dx.doi.org/10.18637/jss.v063.i20

Bivand, Roger S. (2014). GeoComputation and Open-Source Software. *GeoComputation*, *November*, 329. https://doi.org/10.2139/ssrn.1972280

Bivand, Roger S, Pebesma, E. J., Gomez-Rubio, V., & Pebesma, E. J. (2008). *Applied spatial data analysis with R* (Vol. 747248717). Springer.

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

Bosq, D. (2012). *Nonparametric statistics for stochastic processes: estimation and prediction* (Vol. 110). Springer Science & Business Media.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brenning, A. (2012). Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, 5372–5375.

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

Cameletti, M., Lindgren, F., Simpson, D., & Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*, *97*(2), 109–131. https://doi.org/10.1007/s10182-012-0196-3

Correa Morales, J. C., Causil, B., & Javier, C. (2018). *Introducción a la estadística bayesiana: notas de clase*. Instituto Tecnológico Metropolitano.

Cressie, N. (1990). The origins of kriging. *Mathematical Geology*, *22*(3), 239–252.

Cressie, N. (1993). Spatial statistics. *New York*.

Diggle, P. J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns*. Chapman and Hall/CRC.

Diggle, P. J., Ribeiro, P. J., & Christensen, O. F. (2003). An introduction to model-based geostatistics. In *Spatial statistics and computational methods* (pp. 43–86). Springer.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., & Leitão, P. J. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46.

Durbán, M., Lee, D.-J., & Ugarte, M. D. (2008). *Splines con penalizaciones (P-splines): Teoría y aplicaciones*.

Efron, B., & Hastie, T. (2016). Computer age statistical inference: Algorithms, evidence, and data science. In *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. https://doi.org/10.1017/CBO9781316576533

Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, *92*(438), 548–560.

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., & Lehmann, A. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151.

Florinsky, I. V. (2012). The Dokuchaev hypothesis as a basis for predictive digital soil mapping (on the 125th anniversary of its publication). *Eurasian Soil Science*, *45*(4), 445–451.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, *185*, 1–17.

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, *21*(2), 215–223.

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

Goovaerts, P. (1999). Geostatistics in soil science: state-of-the-art and perspectives. *Geoderma*, *89*(1–2), 1–45.

Hamby, D. M. (1994). A review of techniques for parameter sensitivity analysis of environmental models. *Environmental Monitoring and Assessment*, *32*(2), 135–154.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized Additive Models. Chapman&Hall/CRC*. New York.

Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., & Bauer-Marschallinger, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One*, *12*(2), e0169748.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *6*, e5518.

Higgins, J. J. (2003). *Introduction to modern nonparametric statistics*.

Huang, J., Malone, B. P., Minasny, B., McBratney, A. B., & Triantafilis, J. (2017). Evaluating a Bayesian

modelling approach (INLA-SPDE) for environmental mapping. *Science of the Total Environment*, *609*, 621–632.

Isaaks, E. H., & Srivastava, R. M. (1989). *An introduction to applied geostatistics*. Oxford university press.

Kanevski, M., Timonin, V., Pozdnukhov, A., & Ritter, G. (2009). Machine learning for spatial environmental data: theory, applications, and software. In *Ssrn*. EPFL press. https://doi.org/10.2139/ssrn.3015609

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, *14*(2), 1137–1145.

Koo, H., Iwanaga, T., Croke, B. F. W., Jakeman, A. J., Yang, J., Wang, H.-H., Sun, X., Lü, G., Li, X., Yue, T., Yuan, W., Liu, X., & Chen, M. (2020). Position paper: Sensitivity analysis of spatially distributed environmental models- a pragmatic framework for the exploration of uncertainty sources. *Environmental Modelling & Software*, *134*, 104857. https://doi.org/https://doi.org/10.1016/j.envsoft.2020.104857

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

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). Springer.

Li, J., Heap, A. D., Potter, A., & Daniell, J. J. (2011). Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling & Software*, *26*(12), 1647–1659.

Lindgren, F, Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields 670 and Gaussian Markov random fields: the SPDE approach (with discussion). *JR 671 Stat Soc, Series B*, *73*, 423–498.

Lindgren, Finn, & Rue, H. (2015a). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, *63*(19), 1–25.

Lindgren, Finn, & Rue, H. (2015b). Bayesian Spatial Modelling with *R* - **INLA**. *Journal of Statistical Software*, *63*(19). https://doi.org/10.18637/jss.v063.i19

Lindgren, Finn, Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(4), 423–498.

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2005). *Geographic information systems and science*. John Wiley & Sons.

Lovelace, R., Nowosad, J., & Muenchow, J. (n.d.). *Geocomputation with R*.

Margules, C. R., & Pressey, R. L. (2000). Systematic conservation planning. *Nature*, *405*(6783), 243–253. https://doi.org/10.1038/35012251

Matérn, B. (1986). Spatial variation, vol. 36. *Lecture Notes in Statistics*, *2*.

McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. (2003). On digital soil mapping. In *Geoderma* (Vol. 117, Issues 1–2). https://doi.org/10.1016/S0016-7061(03)00223-4

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (Vol. 37). CRC press.

Miller, D. L., Glennie, R., & Seaton, A. E. (2020). Understanding the Stochastic Partial Differential Equation Approach to Smoothing. *Journal of Agricultural, Biological and Environmental Statistics*, *25*(1), 1–16. https://doi.org/10.1007/s13253-019-00377-z

Milne, E. (2009). *Soil organic carbon Mapping Cookbook* (Issue Ipcc). http://www.eoearth.org/view/article/156087/

Moraga, P. (2019). *Geospatial health data: Modeling and visualization with R-INLA and shiny*. CRC Press.

Moraga, P., Cramb, S. M., Mengersen, K. L., & Pagano, M. (2017). A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE. *Spatial Statistics*, *21*, 27–41.

Mulder, V. L., De Bruin, S., Schaepman, M. E., & Mayr, T. R. (2011). The use of remote sensing in soil and terrain mapping—A review. *Geoderma*, *162*(1–2), 1–19.

Page, G. L., Liu, Y., He, Z., & Sun, D. (2017). Estimation and prediction in the presence of spatial confounding for spatial linear models. *Scandinavian Journal of Statistics*, *44*(3), 780–797.

Plant, R. E. (2012). *Spatial data analysis in ecology and agriculture using R*. cRc Press.

Poggio, L., Gimona, A., Spezia, L., & Brewer, M. J. (2016). Bayesian spatial modelling of soil properties and their uncertainty: The example of soil organic matter in Scotland using R-INLA. *Geoderma*, *277*, 69–82.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Reich, B. J., & Fuentes, M. (2007). A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *The Annals of Applied Statistics*, *1*(1), 249–264.

Rue, Hååvard, & Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, *29*(1), 31–49.

Rue, Havard, & Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC press.

Rue, Håvard, Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series b (Statistical Methodology)*, *71*(2), 319–392.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2009). Semiparametric regression during 2003–2007. *Electronic Journal of Statistics*, *3*, 1193.

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

Stroup, W. W. (2016). *Generalized linear mixed models: modern concepts, methods and applications*. CRC press.

Wang, X., Ryan, Y. Y., & Faraway, J. J. (2018). *Bayesian Regression Modeling with INLA*. Chapman and Hall/CRC.

Webster, R., & Oliver, M. A. (2007). Geostatistics for environmental scientists. In *Vadose Zone Journal* (Vol. 1, Issue 2). John Wiley & Sons. https://doi.org/10.2136/vzj2002.0321

Weniger, E. J., & Cížek, J. (1990). Rational approximations for the modified Bessel function of the second kind. *Computer Physics Communications*, *59*(3), 471–493.

West, B. T., Welch, K. B., Galecki, A. T., & Edition, S. (2014). *Linear mixed models: a practical guide using statistical software*. Chapman and Hall/CRC.

Willmott, C. J. (1981). On the validation of models. *Physical Geography*, *2*(2), 184–194.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

Zienkiewicz, O. C., Taylor, R. L., Nithiarasu, P., & Zhu, J. Z. (1977). *The finite element method* (Vol. 3). McGraw-hill London.

Anexos

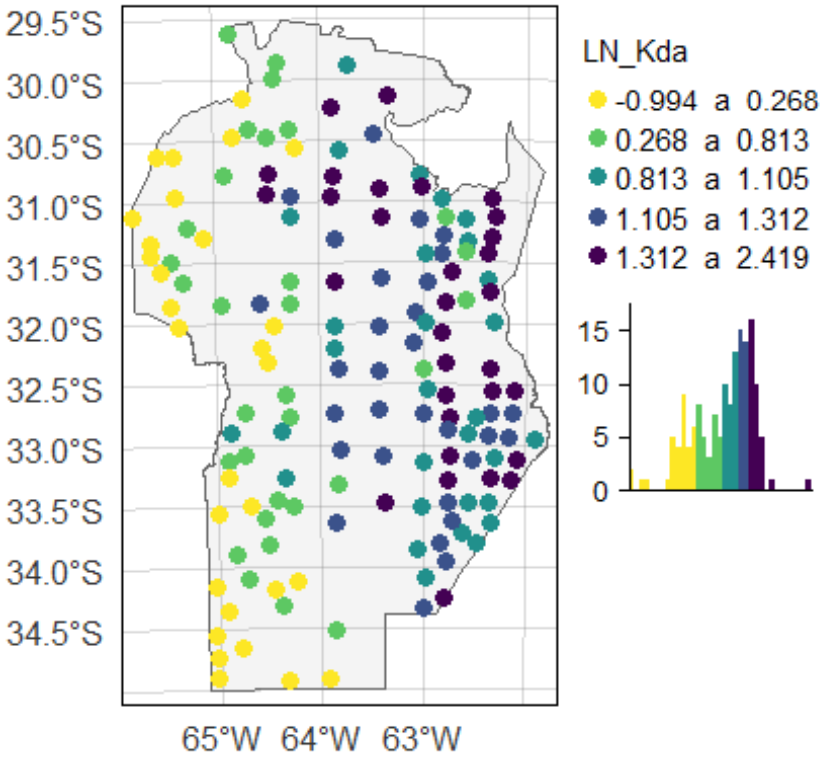
Anexo I

Secuencia de implementación MDS utilizando R-INLA SPDE

Regresión espacial para datos geoestadísticos ejemplo predicción coeficiente de adsorción de atrazina (Kda). Los datos y el código se encuentran en el siguiente repositorio https://github.com/francagiannini/anexoI_mea.

Datos

```
kda <- read.table("kda.txt", header = TRUE, sep = "\t")

kda_sf = st_as_sf(kda, coords=c("x","y") ,crs = 22174 )

limits <- st_read("Cordoba_f4.shp")

## Reading layer `Cordoba_f4' from data source
##   `C:\Users\franc\Dropbox\Franca\Doctorado\Maestria\mea-tesis\mea-tes
is\Cordoba_f4.shp'
##   using driver `ESRI Shapefile'
## Simple feature collection with 1 feature and 4 fields
## Geometry type: POLYGON
## Dimension:     XY
## Bounding box:  xmin: 4234245 ymin: 6125323 xmax: 4614851 ymax: 673552
9
## Projected CRS: POSGAR 98 / Argentina 4

limits_sf <- st_transform(limits, st_crs(kda_sf))

tm_shape(limits_sf)+
  tm_polygons(col="#F4F4F4")+
  tm_graticules(ticks = FALSE, alpha=0.3,labels.size = 1)+
  tm_shape(kda_sf)+
  tm_dots("LN_Kda", title='LN_Kda',
          pal="-viridis",#n=4,
          style = "quantile",
          size = 0.4,
          title.size=2,
          legend.hist = TRUE)+
  tm_layout(
    legend.format = list(text.separator = " a "),
    legend.outside = TRUE,
    legend.hist.width = 1,
    legend.hist.size = 1) +
  tm_legend(text.size=1)
```



Para modelar la correlación espacial con del modelo de Matern a traves de una solución utilizando SPDE se debe construir una malla de traiangulos, en los nodos de esta malla se estima el campo aleatorio utilizando FEM (Metodo de los elementos finitos).

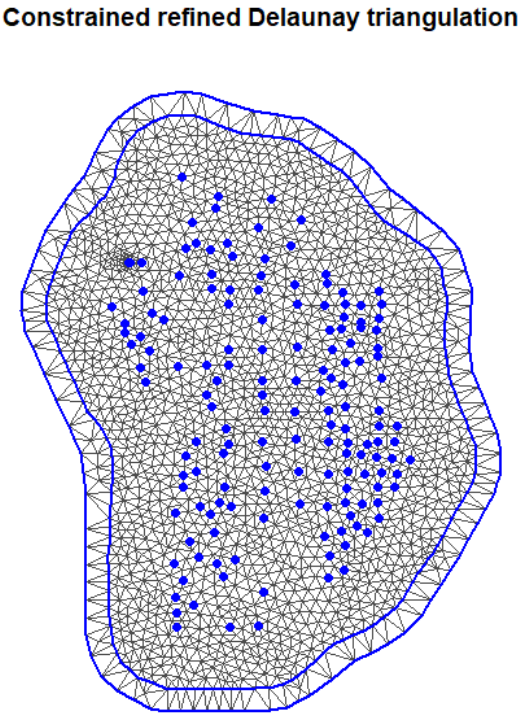
Luego, para la malla se construye una matriz de pesos espaciales que por notación llamamos matriz A.

La malla se puede construir de diversas formas utilizando limites o no y la y la recomendación es que los triángulos sean homogéneos en tamaño y forma

Construcción de la malla

```
#SPDE
#sitios observados
loc.obs <- st_coordinates(kda_sf)

#definición del dominio espacial
boundary.loc <- SpatialPoints(as.matrix(loc.obs))
boundary <- list(
  inla.nonconvex.hull(coordinates(boundary.loc), 81000),
  inla.nonconvex.hull(coordinates(boundary.loc), 111000))

#definición de los parámetros de la malla
mesh <- inla.mesh.2d(boundary.loc,boundary=boundary,
                     max.edge=c(20000, 80000),
                     min.angle=c(30, 20),
                     max.n=c(48000, 16000),
                     max.n.strict=c(128000, 128000),
                     cutoff=200,
```

```
                      offset=c(81000, 111000))

#gráfico
plot(mesh)
points(loc.obs, pch=16 ,col = "blue")
```



Matriz A

```
#definición la estructura de covarianza sobre la malla
spde <- inla.spde2.matern(mesh = mesh, alpha = 2)

#Se proyecta esta estructura sobre los sitios observados es decir se def
ine A
A <- inla.spde.make.A(mesh = mesh, loc = loc.obs)

s <- inla.spde.make.index(name = "s", n.spde = spde$n.spde)
```

Ajuste del modelo

Para facilitar el manejo se implementa un sistema de organización denominado stack que organiza los elementos de la estimación, el vector de la variable respuesta, la matriz A, la matriz de covariables

```
stk.est <- inla.stack(
  data = list(LN_Kda = kda$LN_Kda),
```

```
  A = list(A, 1),
  effects=list(s=1:spde$n.spde,
               data.frame(Intercept=1,
                          Elevation=kda$Elevation,
                          PPanual=kda$PPanual,
                          SOC=kda$SOC,
                          TvsPP=kda$TvsPP,
                          Clay=kda$Clay)),
  tag = 'est'
)

#aJuste del modelo

formula = LN_Kda~ -1+Intercept+TvsPP+PPanual+SOC+Clay+ f(s, model = spde
)


res_est <- inla(
  formula,
  family = "gaussian",
  data = inla.stack.data(stk.est),
  control.predictor = list(A = inla.stack.A(stk.est),
  link=1, compute = TRUE)
)
```

Resultados del modelo

```
summary(res_est)

##
## Call:
##    c("inla(formula = formula, family = \"gaussian\", data =
##    inla.stack.data(stk.est), ", " control.predictor = list(A =
##    inla.stack.A(stk.est), link = 1, ", " compute = TRUE))")
## Time used:
##     Pre = 2.19, Running = 12.8, Post = 1.04, Total = 16.1
## Fixed effects:
##             mean    sd 0.025quant 0.5quant 0.975quant   mode kld
## Intercept -4.643 1.087     -6.717   -4.667     -2.431 -4.715   0
## TvsPP      4.105 1.590      0.901    4.131      7.165  4.183   0
## PPanual    0.005 0.001      0.003    0.005      0.007  0.005   0
## SOC        0.038 0.004      0.030    0.038      0.045  0.037   0
## Clay       0.007 0.003      0.001    0.007      0.013  0.007   0
##
## Random effects:
##   Name     Model
##     s SPDE2 model
##
## Model hyperparameters:
##                                           mean    sd 0.025quant 0.5qu
ant
## Precision for the Gaussian observations  32.24 8.447      18.50    31
```

```
.29
## Theta1 for s                              10.85 0.344      10.19    10.85
## Theta2 for s                             -10.71 0.371     -11.46   -10.71
##                                            0.975quant   mode
## Precision for the Gaussian observations      51.49  29.50
## Theta1 for s                                 11.54  10.82
## Theta2 for s                                -10.00 -10.68
##
## Expected number of effective parameters(stdev): 62.42(19.59)
## Number of equivalent replicates : 2.47
##
## Marginal log-Likelihood:  -53.18
## Posterior marginals for the linear predictor and
##  the fitted values are computed
```

Resultados de la estructura espacial

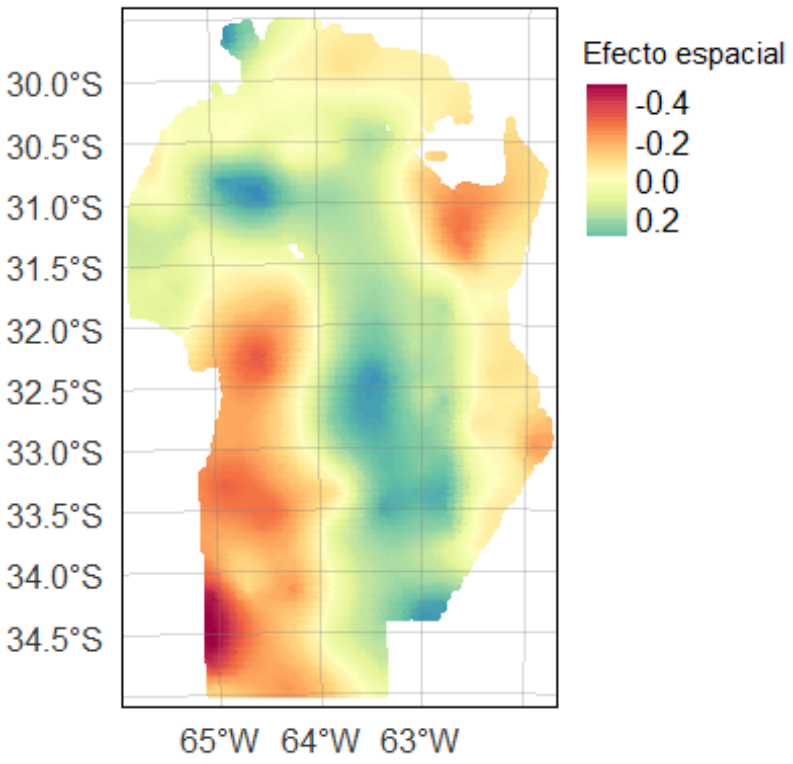
```
spde.est = inla.spde2.result(inla = res_est,
                                  name = "s",
                                  spde = spde,
                                  do.transform = TRUE)

#Estadísticos de posición Varianza nominal
inla.zmarginal(spde.est$marginals.variance.nominal[[1]])

## Mean            0.0652057
## Stdev           0.0244048
## Quantile  0.025 0.0303343
## Quantile  0.25  0.0477266
## Quantile  0.5   0.0608066
## Quantile  0.75  0.0778257
## Quantile  0.975 0.125008

#Rango
inla.zmarginal(spde.est$marginals.range.nominal[[1]])

## Mean            136163
## Stdev           52536
## Quantile  0.025 62512.6
## Quantile  0.25  98649.5
## Quantile  0.5   126212
## Quantile  0.75  162711
## Quantile  0.975 266100
```

Predicción sobre una grilla de predicción

```
# Datos
grid <- st_read("modelsel_serv/grid_kda_sf.gpkg") %>% st_transform(crs =
st_crs(kda_sf))
```

El efecto aleatorio espacial estimado debe proyectarse soobre la grilla de sitios no observados sobre los que se quiere predecir. Se puede graficar los ponderadores espacales en la matriz A

```
#prohjección y definición de A de predicción
A.pred <- inla.spde.make.A(mesh = mesh, loc = st_coordinates(grid))
project <- inla.mesh.projector(mesh, loc = st_coordinates(grid))
grid$sp.mean <- inla.mesh.project(project, res_est$summary.ran$s$mean)

tm_shape(grid)+
  tm_dots(
    "sp.mean",
    style="cont",
    pal="Spectral",
    title='Efecto espacial',
    size =0.1,
    title.size=2) +
  tm_graticules(ticks = FALSE, alpha=0.3,labels.size = 1)+
  tm_layout(
    legend.format = list(text.separator = " a "),
    legend.outside = TRUE,
    legend.hist.width = 1,
    legend.hist.size = 1) +
  tm_legend(text.size=1)
```



Se debe generar un stack de predicción y luego juntarla con el stack de estimación

```
#organización los inputs en un stack de predicción
stk.pred = inla.stack(data = list(LN_Kda = NA),
                      A = list(A.pred, 1),
                      effects = list(s=1:spde$n.spde,
                        data.frame(Intercept=1,
```

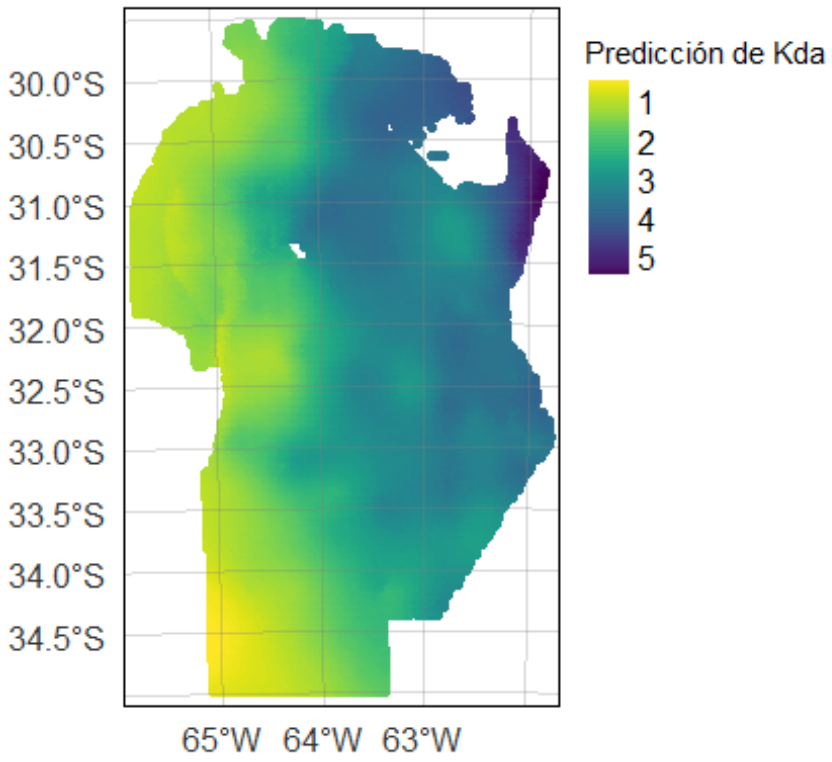
```
                      Elevation=grid$Elevation,
                      PPanual=grid$PPanual,
                      SOC=grid$SOC,
                      TvsPP=grid$TvsPP,
                      Clay=grid$Clay)),
                    tag = "pred")

#union de inputs de estimación y predicción en una stack conjunta
stk.all <- inla.stack(stk.est, stk.pred)

#Ajuste del modelo con datos faltandes de la variable
#respuestas en los sitios no observados

res_pred = inla(formula = formula,
              data = inla.stack.data(stk.all, spde = spde),
              family = "gaussian",
              control.predictor = list(A = inla.stack.A(stk.all)
                                       ,compute = TRUE))
```

Resultados de la prediccion

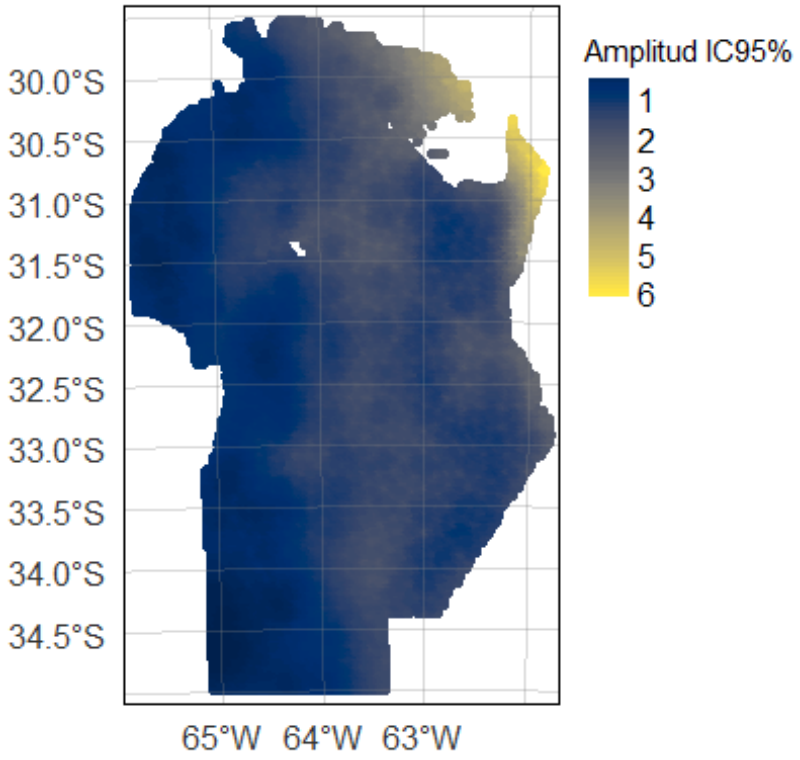
```
#Cambian ligeramente las estimaciones del modelo
spde.pred = inla.spde2.result(inla = res_pred,
                                    name = "s",
                                    spde = spde,
                                    do.transform = TRUE)
#Varianza nominal
inla.zmarginal(spde.est$marginals.variance.nominal[[1]])

## Mean            0.0652057
## Stdev           0.0244048
## Quantile  0.025 0.0303343
## Quantile  0.25  0.0477266
## Quantile  0.5   0.0608066
## Quantile  0.75  0.0778257
## Quantile  0.975 0.125008

inla.zmarginal(spde.pred$marginals.variance.nominal[[1]])

## Mean            0.0663477
## Stdev           0.0262493
## Quantile  0.025 0.02984
## Quantile  0.25  0.0476278
## Quantile  0.5   0.0612897
## Quantile  0.75  0.0794939
## Quantile  0.975 0.131489
```

Extracción de resultados y mapeo

```r
# Mapeo

igr <- inla.stack.index(stk.all,'pred')$data

grid_map <- grid %>%
  bind_cols(res_pred$summary.fitted.values[igr, ]) %>%
  mutate(Kda_pred=exp(mean),
         Kda0.025quant=exp(`0.025quant`),
         Kda0.975quant=exp(`0.975quant`),
         KdaIC95=exp(`0.975quant`)-exp(`0.025quant`)
         )

# Media de la distribución a posteriori
tm_shape(grid_map) +
  tm_dots(
    "Kda_pred",
    style = "cont",
    palette = "-viridis",
    title = "Predicción de Kda",
    size = 0.1) +
  tm_graticules(ticks = FALSE, alpha=0.3,labels.size = 1)+
  tm_layout(
    legend.format = list(text.separator = " a "),
    legend.outside = TRUE,
    legend.hist.width = 1,
    legend.hist.size = 1) +
  tm_legend(text.size=1)
```



```r
# Medida de incertidumbre IC95
tm_shape(grid_map)+
  tm_dots("KdaIC95",
          style="cont",
          pal="cividis",
          title='Amplitud IC95%',
          size =0.1)+
  tm_graticules(ticks = FALSE, alpha=0.3,labels.size = 1)+
  tm_layout(
    legend.format = list(text.separator = " a "),
    legend.outside = TRUE,
    legend.hist.width = 1,
    legend.hist.size = 1) +
  tm_legend(text.size=1)
```



Anexo II

ANAVA Error de Predicción

Tabla 6: Error de Predicción Global (EP%) expresado como porcentaje de la media predicha para cada sitio en función de: algoritmos de predicción espacial (Regresión Bayesiana para datos espaciales (RB), Regresión Kriging (RK) y Random Forest con residuos krigeados (RF)), tamaño muestral ($\boldsymbol{n}$), y número de variables explicativas $\boldsymbol{p}$ y sus respectivas interacciones.

Factor	Materia Orgánica de Suelo n=64825			Metales Pesados n= 45720			Adsorción atrazina n=58840		
	β	EE		β	EE		β	EE	
Intercepto (ref RB)	37.1530	1.8688	*	7.4705	0.0702	*	45.589	4.548	*
RF	0.9100	2.6654		-0.6370	0.0997	*	19.070	6.447	*
RK	3.9816	2.6708		0.0013	0.0707		0.215	6.436	
p	1.1704	0.1600	*	-0.4032	0.0187	*	5.482	0.571	*
n	-0.0043	0.0016	*	-0.0189	0.0011	*	-0.055	0.069	
RF*p	-1.3805	0.2281	*	-0.1452	0.0265	*	-6.909	0.809	*
RK*p	0.2281	0.2288		-0.0147	0.0265		0.326	0.808	
RF*n	0.0007	0.0022		0.0116	0.0015	*	-0.069	0.097	
RK*n	-0.0009	0.0022		0.0020	0.0015		0.033	0.097	
n*p	-0.0005	0.0001	*	0.0001	0.0003		-0.071	0.009	*
RF*p*n	0.0005	0.0002	*	-0.0020	0.0004	*	0.070	0.012	*
RK*p*n	-0.0001	0.0002		-0.0002	0.0004		-0.001	0.012	