FastMapping: software para mapeo de variabilidad en dominios espaciales continuos

Mariano Córdoba[1], Pablo Paccioretti[1], Cecilia Bruno[1], Fernando Aguate[1] y Mónica Balzarini[1]

[1]Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba-CONICET, Ing. Agr. Félix Aldo Marrone 746, Ciudad Universitaria, Córdoba
mbalzari@agro.unc.edu.ar.

Resumen. La difusión de tecnologías agrícolas que producen datos espaciales ha generado la necesidad de herramientas informáticas integradas para el mapeo de variabilidad espacial a escala fina. Para obtener un mapa de rendimiento se usan diferentes software para depurar datos y predecir el rendimiento en lugares sin registro. FastMapping implementa herramientas estadísticas para la depuración de datos espaciales, ajusta y selecciona cuasi automáticamente variogramas para realizar y mapear la variable de interés por interpolación kriging. Es una aplicación web interactiva, con una interfaz amigable desarrollada en lenguaje R. Los usuarios pueden obtener el mapa de rendimiento o suelo para un lote agrícola, así como un mapa varianza de predicción y estadísticas de error de predicción sin necesidad de manipular códigos de programación. FastMapping se encuentran disponibles en forma libre en http://fastmapping.psi.unc.edu.ar/

Palabras Clave: aplicación web, geoestadística, agricultura de precisión, mapa de rendimiento

1 Introducción

El manejo uniforme de lotes agrícolas está siendo reemplazado por el manejo por ambientes, el cual se ve beneficiado por la descripción estadística de la variabilidad espacial intralote de una o más variables de interés. El uso de nuevas maquinarias de precisión acopladas a sistemas de posicionamiento global (GPS) producen miles de datos georreferenciados dentro de un lote; ejemplos comunes son las cosechadoras con monitores de rendimiento y las rastras equipadas con sensores proximales para monitoreos intensivos del suelo [1]. Las nuevas tecnologías asociadas a la agricultura de precisión producen grandes volúmenes de datos espaciales generando demandas de herramientas de software que faciliten el procesamiento rápido y efectivo de estos nuevos datos. Productos que se pueden lograr con estos datos y herramientas son los mapas o capas de información espacial del lote. Los mapas de rendimiento y los mapas de propiedades del suelo, muestran la variación espacial del atributo de interés en un

dominio continuo. Sin embargo, aun cuando se hacen mediciones intensivas de una propiedad sobre el terreno, estas son discretas por lo que deben realizarse interpolaciones para construir un mapa de variación espacial en un continuo. La interpolación espacial de tipo estadístico resulta clave para describir la variación espacial de variables con fuerte componente aleatoria como el rendimiento. La interpolación kriging, basada en el ajuste de funciones de variograma, tiene la pericia de predecir valores de la variable aleatoria en sitios sin mediciones para producir mapas de variabilidad espacial en un continuo. Existen excelentes herramientas de software libre para ajustar variogramas y realizar interpolaciones, como son los paquetes geoR [2] y gstat [3] de R [4], pero en la práctica su implementación, entre técnicos y productores, es escasa por demandar conocimientos de programación. Otros softwares, como PG2000 [5], EasyKrig3.0 [6] o Surfer [7], realizan mapeos por interpolaciones espaciales estadísticas sin demandar programación especializada, pero no ofrecen herramientas para la depuración y preprocesamiento de los datos, siendo que los valores *outliers* han demostrado efecto sobre la caracterización de la variabilidad espacial. Algunos sistemas de información geográfica (GIS) como ArcGIS [8] y QGIS [9] tienen herramientas para análisis geoestadístico, pero también pueden presentar dificultades para su uso y ofrecen limitadas facilidades para la depuración de los datos. FastMapping fue desarrollado para automatizar, en una plataforma amigable, un protocolo de análisis geoestadístico tendiente a mapear la variabilidad espacial de una variable aleatorio en un dominio espacial continuo. Permite el procesamiento de mapas de rendimiento, como los provenientes de agricultura de precisión, pero puede ser usado en otros contextos más generales u otros tipos de mapas con mayor intervención por parte del usuario. La aplicación fue creada usando la librería shiny [10] e involucra la librería automap [11] que permite el ajuste de variogramas y la generación de mapas de variabilidad espacial de manera similar al paquete gstat [3].

FastMapping integra en un mismo ambiente las distintas etapas de procesamiento que en otros entornos hay que realizar separadamente [12]. La innovación metodológica embebida en FastMapping es la integración de una etapa de pre-procesamiento de datos orientada a eliminar anomalías comunes en mapas de rendimiento o variables agrícolas colectadas automáticamente. Genera, mapas de variabilidad espacial con mínima intervención del usuario. La calidad del mapa es evaluada por validación cruzada [13]. La secuencia lógica de análisis (protocolo) (Figura 1), que va desde el preprocesamiento de los datos a la validación del mapa construido, puede implementarse en FastMapping para producir resultados gráficos (mapas) de alta calidad. Técnicos y productores interesados en obtener mapas de rendimiento desde las bases de datos que producen los monitores de las cosechadoras encontrarán en este software una herramienta útil para mapear rendimiento y conocer los errores de predicción. El objetivo de este trabajo es especificar el paso a paso del protocolo implementado en el software FastMapping para obtener un mapa de variabilidad espacial e ilustrar la obtención de un mapa de rendimiento de en un lote agrícola de la región pampeana Argentina.

2 Protocolo para análisis de variabilidad espacial

2.1 Depuración de datos

La presencia de valores raros entre los datos de una variable impacta los estudios de variabilidad espacial. Consecuentemente, la depuración previa o pre-procesamiento de los datos debe ser el punto de partida en un protocolo de análisis de datos espaciales (Figura 1) [12], [14], [15].

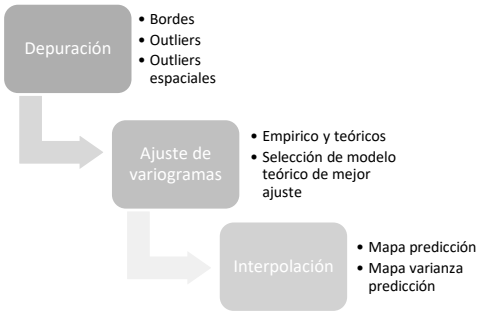

Fig. 1. Pasos del protocolo analítico para el análisis de variabilidad espacial implementado en FastMapping.

Eliminación de bordes

La primera opción de depuración consiste en la remoción de datos erróneos relacionados a los efectos de bordes. Este tipo de efectos, frecuentemente denominado "efecto cabecera", es común de observar en mapas de rendimiento. FastMapping, por defecto eliminará las observaciones que se ubican a una distancia de hasta 20 m desde los límites del lote, correspondiente a más de una pasada de la cosechadora. Otros valores de zona *buffer*, pueden ser ingresados por el usuario.

Eliminación de *outliers* globales

Los *outliers*, o valores atípicos, son observaciones con valores que se encuentran fuera del patrón general o distribución del conjunto de datos. Estas observaciones se pueden eliminar a través de un proceso donde se complementan distintas técnicas y teorías: 1) el conjunto de datos se limita dentro de un rango de variación razonable donde los valores máximos y mínimos se obtienen desde el conocimiento previo de su distribución, 2) para el conjunto de datos de una variable, se calcula la media y la desviación estándar (DE) y se identifican los valores que se encuentran fuera de la media ± 3 DE. Teóricamente, casi el 90% de los datos de cualquier variable aleatoria se encuentran entre la media ± 3 SD. En mapas de rendimiento donde los datos son sesgados por procesos no aleatorios tales como malas lecturas de monitores,

cosechadoras funcionando a medio llenar o con el cabezal hacia abajo sobre áreas cosechadas, puede realizarse una modificación de estos límites [12].

Eliminación de *outliers* espaciales

Los *outliers* espaciales son datos que difieren significativamente de su vecindario, pero se sitúan dentro del rango general de variación del conjunto de datos. Existen herramientas estadísticas diseñadas específicamente para identificarlos. Tal es el caso del índice autocorrelación espacial local de Moran (IMl) [16]. Dado un grupo de datos que pertenecen a diferentes vecindarios, el IMl es aplicado a cada dato individualmente y da idea del grado de similitud o diferencia entre el valor de una observación en un sitio respecto al valor de la observación en sus sitios vecinos. Se consideran sitios vecinos a puntos contiguos ubicados dentro de un rango de distancia pre-establecido. En otros estudios en los que se procesaron mapas de rendimiento provenientes de la región pampeana Argentina, la distancia utilizada, con buenos resultados, como límite para definir los puntos vecinos de cada observación fue de 20 m [17]. Este es el valor usado por defecto en FastMapping.

2.2 Interpolación espacial

Ajuste de variogramas

La teoría de variables regionalizadas define funciones para modelar variabilidad espacial denominadas variogramas [18]. Bajo este marco teórico, el primer paso para analizar variabilidad espacial es construir un variograma empírico (a partir de los datos). Los parámetros de la función variograma son: la varianza nugget o efecto pepita, la varianza estructural o sill parcial y el rango [19]. Luego se ajusta un modelo teórico de variograma sobre el variograma empírico. El variagrama ajustado es usado para obtener predicciones de la variable de interés para cualquier interdistancia perteneciente al dominio espacial estudiado. Existen distintos modelos teóricos (Figura 2) entre ellos los más usados son: modelo exponencial, modelo esférico y el modelo gaussiano. En un procesamiento de 600 mapas de rendimiento en el que se ajustaron estos tres modelos teóricos, en un 90% de los mapas el modelo exponencial fue el de mejor ajuste [17]. La selección del modelo teórico de mejor ajuste se realizará mediante un proceso de validación cruzada que evalúa la capacidad predictiva del modelo de correlación espacial seleccionado [13].

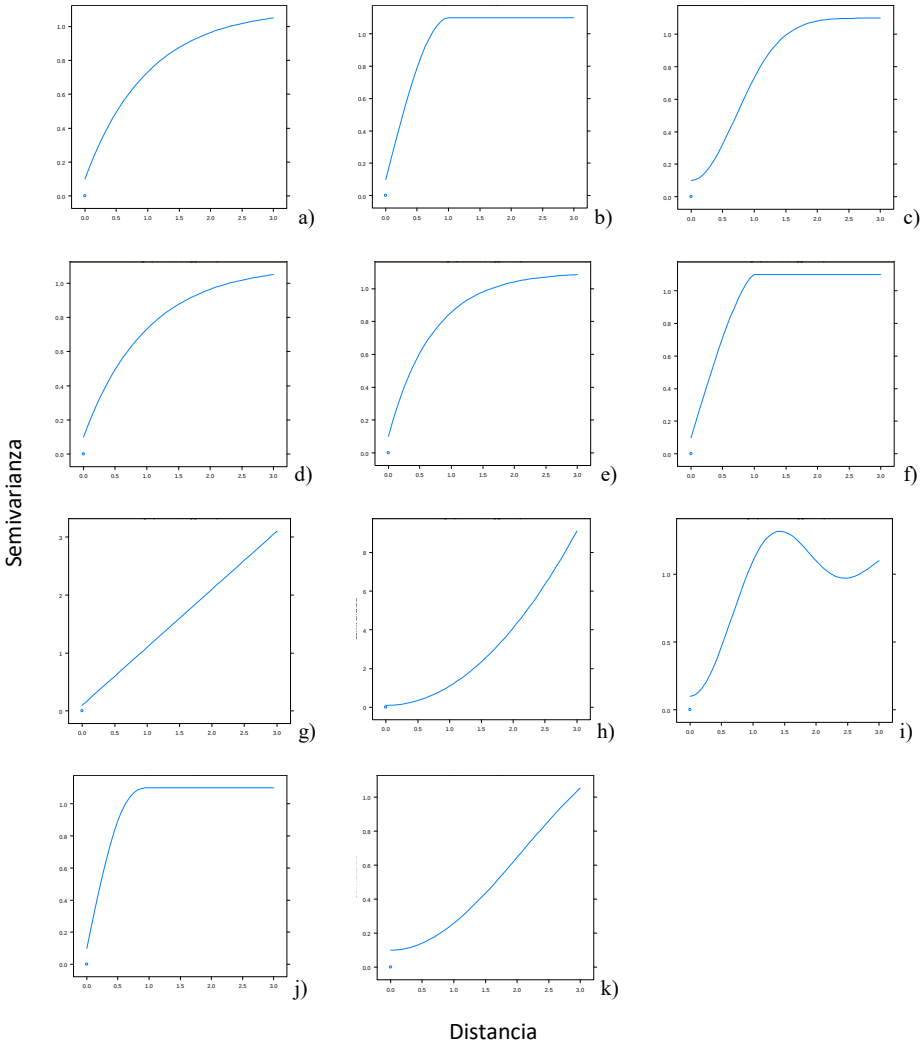

Fig. 2. Modelos de semivariogramas (semivarianza en función de distancia) disponibles en Fastmapping: a) exponencial, b) esférico, c) gaussiano, d) Matern, e) parametrización de Matern Stein, f) circular, g) lineal, h) *power*, i) *wave*), j) pentaesférico, k) *hole.*

Predicción espacial

Kriging es una técnica utilizada en geostadística para realizar interpolaciones espaciales y poder predecir los valores de la variable en sitios no muestreados. El método kriging proporciona el mejor predictor lineal para el valor de la variable en un sitio, suministrando además un error de predicción conocido como varianza kriging, que depende del modelo de variograma ajustado y de las localizaciones de los datos

originales. La varianza kriging brinda la posibilidad de analizar la calidad de las predicciones obtenidas por interpolación. Para evitar el uso de información proveniente de muestras redundantes, el método kriging pondera de forma distintas muestras que están muy cerca entre sí y de la misma región que muestras que estén en lados opuestos al sitio al que se quiere asignar un valor por interpolación. Los parámetros del variograma ajustado tienen importancia a la hora de asignar ponderadores a las muestras que rodean el punto a interpolar.

Entre las opciones de interpolación espacial se destacan los métodos de kriging ordinario, simple y universal. En el kriging ordinario la media de la variable es estimada localmente. En caso de conocer la media de la variable, hecho que raramente ocurre, se utiliza el kriging simple. En el kriging universal la media es estimada y se incluye también la influencia de una tendencia espacial de los datos. La predicción asignada a los puntos incógnita puede realizarse de manera puntual (kriging puntual) o definiendo bloques (kriging en bloques) [19]. La interpolación puntual es la estimación del valor de la variable en el sitio incógnita, mientras que la interpolación por bloques estima la media de puntos de un área predeterminada que rodea al sitio incógnita. La interpolación por bloques (que produce un "suavizado" de las estimaciones) suele correlacionar mejor con los valores verdaderos. Así, la interpolación espacial puede realizarse utilizando todos los datos simultáneamente (kriging global) o alternativamente puede usarse la información de datos vecinos para la realizar la predicción sobre un punto dado (kriging local). Cuando se trabajan bases de datos con miles de observaciones pueden producirse problemas computacionales referidos a falta de memoria. Para bases de datos de tamaños grandes (n> 1000) utilizar kriging global puede ser lento [20]. En el procesamiento de mapas de rendimiento las bases de datos tienen usualmente más de 1000 datos. Por ello, se recomienda la opción kriging local.

3 Ilustración

3.1 Depuración de datos

En la pestaña *Depuration* de FastMapping una de las opciones de limpieza de datos consiste en la remoción de datos erróneos debido a los efectos de bordes y cabecera de lote (Figura 3). La opción por defecto elimina las observaciones que se ubican a una distancia de hasta 20 m desde los límites del lote. El usuario puede modificar este límite como así también no utilizar esta opción de depuración. Para detectar *outliers* globales FastMapping primero elimina por defecto a datos con valor menor o igual a cero. El usuario también puede especificar un valor máximo admisible para la variable analizada. Posteriormente, identifica y elimina datos que se encuentran por fuera del intervalo media ± 3 DE. Finalmente, para identificar *outliers* espaciales el software considera por defecto a sitios vecinos a aquellos contiguos ubicados dentro de un rango de distancia máximo de 20 m. Los límites mínimos y máximos, la DE y la distancia para definir el vecindario también pueden ser modificadas por el usuario. Como resultado de la depuración se generará una pestaña (*Data Extracted*) en la ventana *Results* que contiene la clasificación de cada observación donde se considera si el dato

pertenece al borde del espacio mapeado, si es *outlier* global u *outlier* espacial y se mostrará la ubicación espacial de cada tipo de observación en la pestaña *Plot Condition Depurated* (Figura 5).

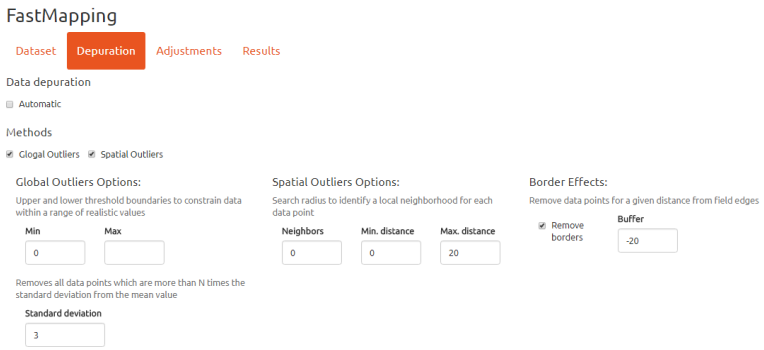

Fig.3. Opciones de depuración de datos espaciales de FastMapping.

3.2 Interpolación espacial

El siguiente paso será el ajuste de un variograma como función para modelar la variabilidad espacial de la variable regionalizada. Se realizará el ajuste de varios modelos de correlación espacial y se selecciona aquel que presenta mejor performance en la predicción espacial del fenómeno bajo estudio, *i.e.* menor error de predicción. Los modelos que pueden seleccionarse para el ajuste son: Exp (exponencial), Sph (esférico), Gau (gaussiano), Mat (Matern), Ste (parametrización de Matern Stein), Cir (circular), Lin (lineal), Pow (potencia o *power*), Wav (ondulado o *wave*), Pen (pentaesférico), Hol (holístico o *hole*) (Figura 2). Con la opción *Automatic* se ajustan todos esos modelos y se identifica el mejor en términos de error de predicción (Figura 4).

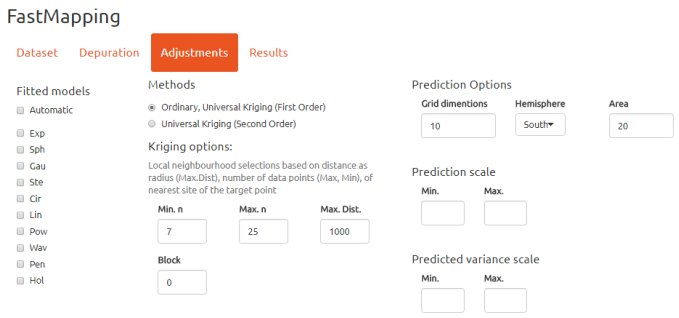

Fig. 4. Opciones de ajuste de modelos de variograma e interpolación espacial de FastMapping.

En todos los ajustes que se realizan se asume que la variación del valor de la variable con el espacio es igual en todas las direcciones de éste (variograma omnidireccional). Los ajustes de los modelos se realizan por el método de mínimos cuadrados ponderados (WLS). FastMapping estima los parámetros iniciales de acuerdo a lo especificado en la descripción del paquete automap. Una opción que puede ser elegida es realizar la interpolación y el ajuste previo del variograma contemplando tendencias con las coordenadas (kriging universal) de primer orden o de segundo orden.

En agricultura de precisión cuando se utilizan datos registrados con sensores proximales se recomienda realizar la predicción usando kriging en bloque con tamaño similar a la separación de las líneas de muestreo, mientras que para datos recolectados con monitores de rendimiento la dimensión recomendada es de 20 m [12] (opción por defecto). En la opción *Methods*, el usuario puede elegir el número mínimo (*Min. n.)* y máximo (*Max. n*.) de puntos utilizados para realizar la estimación en cada uno de los sitios de la grilla de predicción. Esta opción permite realizar interpolación en un contexto local (Kriging *neighbourhood*). Este tipo de interpolación también puede contemplarse definiendo una distancia máxima (*Max. Dist.*) la cual indica que únicamente las observaciones que se encuentran dentro de dicho rango (hasta la máxima distancia espacial) son usadas para la predicción. En caso de que la predicción se realice utilizando toda la información disponible (kriging global) se debe escribir en la opción *Max. n.* y en *Dist.Max.* la sigla Inf (infinito).

En la opción *Prediction Options*, se especifica la dimensión de la grilla de predicción (*Grid dimentions*). Por ejemplo, si se trabaja con coordenadas cartesianas y se ingresa el valor 10, la grilla de predicción será de 10 m × 10 m. En caso de que se disponga de las coordenadas de los puntos que conforman el polígono del área sobre la cual se desea realizar la interpolación, estas pueden cargarse desde un archivo *.txt* utilizando para ello la opción *Upload edges file* en la pestaña *Dataset*. En caso de no contar con esta información, el software creará un polígono utilizando los datos que tienen ubicaciones espacialmente marginales. Con las opciones *Hemisphere* y *Zone*, permite asignar el hemisferio y la zona o faja UTM a la cual pertenecen los datos. Estos son importantes para generar un mapa *geotiff* que contiene la información de la predicción espacial realizada. Con las opciones *Prediction scale* y *Predicted variance scale*, se pueden especificar las escalas mostradas en los mapas de interpolación resultante, tanto de los valores predichos como de su varianza de predicción.

Cuando se selecciona la pestaña *Results* (Figura 5) el software evalúa la capacidad predictiva de cada modelo de correlación espacial seleccionado. Para ello, realiza una validación cruzada del tipo N-fold, en la cual los datos son particionados en N grupos excluyentes. Luego se realiza la predicción para todas las observaciones de un grupo utilizando los N-1 grupos de datos restantes [3]. Con las diferencias entre los valores observados y predichos se calcula el error cuadrático medio de predicción, que se expresa como raíz cuadrada (RMSE). El modelo de mejor capacidad predictiva es aquel con menor RMSE. Una vez que el modelo es seleccionado, el variograma experimental y teórico ajustados pueden ser visualizados en la opción *Plot* como así también el mapa de variabilidad espacial y de varianza de predicción, obtenido mediante interpolación kriging. El panel *Results* también muestra los parámetros estimados para el modelo ajustado y seleccionado junto a la RMSE y el error de predicción relativo a la media en porcentaje.

FastMapping integra los pasos del análisis espacial en un único ambiente. Para ilustrar el uso de FastMapping en datos de rendimiento de soja, las opciones utilizadas para la depuración fueron: rendimiento mínimo y máximo de 0 y 6 t ha^{-1}, respectivamente; remoción de bordes de 20 m; 3 DE para eliminación de *outliers* globales y 20 m para la definición de los vecindarios en la eliminación de *outliers* espaciales. Las opciones de interpolación fueron las disponibles por defecto incluyendo el ajuste de todos los modelos disponibles. FastMapping generó un mapa que ajustó bien a la estructura espacial subyacente.

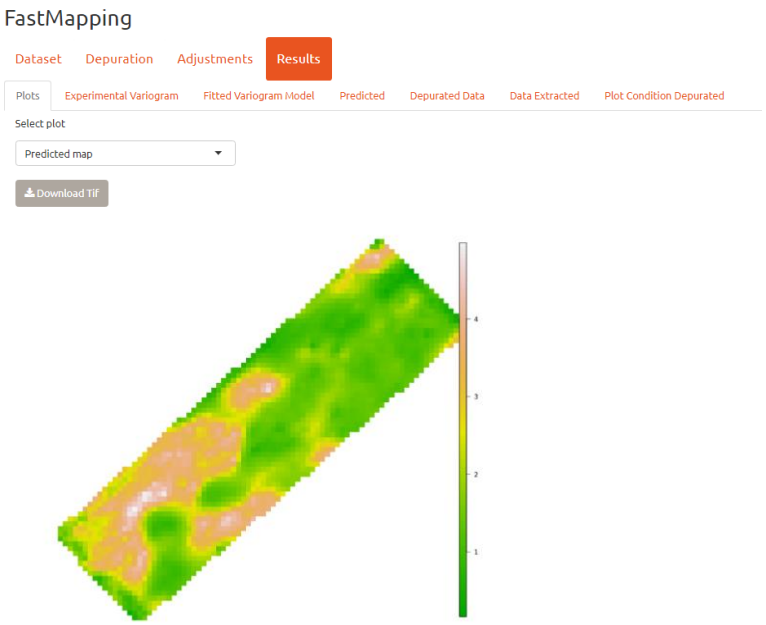

Fig. 5. Mapa de variabilidad espacial del rendimiento en un lote de soja

4 Otras características del Software

FastMapping permite importar archivos en formato *.txt* con diferentes opciones de separación de caracteres. Los resultados pueden exportarse en una tabla con formato CSV en la cual se incluyen las coordenadas de los sitios interpolados, los valores predichos y la varianza de predicción. Además, los mapas de variabilidad pueden ser copiados o exportados como archivo GeoTiff. Este último puede ser cargado en software para datos espaciales o software GIS. FastMapping permite a los usuarios interactuar con sus datos sin tener que manipular códigos de programación. Realiza una programación reactiva que vincula los valores de entrada con los de salida, es decir que cuando una entrada cambia, el servidor reconstruye cada salida que depende de ella.

5 Implementación

Este software es desarrollado en el lenguaje de programación R [4] siendo compatible con diferentes plataformas (e.g. Windows and Linux). Utiliza las librerías Shiny [10] y Shinythemes [22], R como interfase grafica con el usuario. La depuración espacial se realiza con funciones de la librería spdep. Los ajustes de variogramas e interpolación espacial se obtienen a través de los procedimientos desarrollados y disponibles en las librerías automap [11], fields [23], geoR [2] y raster [24]. La aplicación web está alojada en un servidor de la Universidad Nacional de Córdoba (UNC). Se accede ingresando a http://fastmapping.psi.unc.edu.ar/, desde cualquier navegador de internet. Solo se necesita una conexión a internet y un navegador web, no se necesita descargar ningún programa ni instalarlo. Tampoco requiere la instalación de R ni de sus librerías. El tamaño máximo de archivo a procesar es 20MB. La velocidad de cómputo dependerá de la memoria disponible en el servidor.

Agradecimientos

Este desarrollo ha sido producido por investigadores y becarios del Consejo Nacional de Ciencia y Tecnología (CONICET) con lugar de trabajo en la Catedra de Estadística y Biometría de la Facultad de Ciencias Agropecuarias de la Universidad Nacional de Córdoba, Argentina. Las investigaciones metodológicas que sustentan el protocolo de análisis implementado en FastMapping son subsidiadas por el Proyecto PICT 2014 1071 del MinCyT de Argentina, SECyT-UNC y por el proyecto PIODO 2015 del MinCyT de la Provincia de Córdoba, Argentina.

Referencias

[1] K. Piikki, M. Söderström, and B. Stenberg, “Sensor data fusion for topsoil clay mapping,” *Geoderma*, vol. 199, pp. 106–116, 2013.
[2] P. J. Ribeiro Jr and P. J. Diggle, “geoR: Analysis of Geostatistical Data.” 2016.
[3] E. J. Pebesma, “Multivariable geostatistics in S: the gstat package,” *Comput. Geosci.*, vol. 30, pp. 683–691, 2004.
[4] R Core Team, “R: A Language and Environment for Statistical Computing.” Vienna, Austria, 2016.
[5] I. Clark and W. V Harper, “Practical geostatistics: Ecosse North American LLC,” *Columbus, OH*, 2000.
[6] D. Chu, “The GLOBEC kriging software package--EasyKrig3. 0,” *online], http//globec. whoi. edu/software/kriging/easy_krig/easy_krig. html*, 2004.
[7] G. S. Surfer, “User’s guide. Golden Software Inc.” Colorado, USA, 2011.
[8] A. D. Esri, “Release 10,” *Doc. Manual. Redlands, CA, Environ. Syst. Res. Inst.*, 2011.
[9] QGIS Development Team, “QGIS Geographic Information System.” 2016.
[10] W. Chang, J. Cheng, J. J. Allaire, Y. Xie, and J. McPherson, “shiny: Web Application Framework for R.” 2016.
[11] P. H. Hiemstra, E. J. Pebesma, C. J. W. Twenh"ofel, and G. B. M. Heuvelink, “Real-time automatic interpolation of ambient gamma dose rates from the Dutch Radioactivity

Monitoring Network," *Comput. Geosci.*, 2008.
[12] J. A. Taylor, A. B. McBratney, and B. M. Whelan, "Establishing Management Classes for Broadacre Agricultural Production," *Agron. J.*, vol. 99, no. 5, p. 1366:1376, 2007.
[13] R. Kohavi and others, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, 1995, vol. 14, no. 2, pp. 1137–1145.
[14] M. A. Córdoba, C. I. Bruno, J. L. Costa, N. R. Peralta, and M. G. Balzarini, "Protocol for multivariate homogeneous zone delineation in precision agriculture," *Biosyst. Eng.*, vol. 143, pp. 95–107, 2016.
[15] W. Sun, B. Whelan, A. B. McBratney, and B. Minasny, "An integrated framework for software to provide yield data cleaning and estimation of an opportunity index for site-specific crop management," *Precis. Agric.*, vol. 14, no. 4, pp. 376–391, 2013.
[16] L. Anselin, "Local Indicators of Spatial Association-LISA," *Geogr. Anal.*, vol. 27, no. 2, pp. 93–115, Sep. 1995.
[17] A. Vega, M. Córdoba, and M. Balzarini, "Using the local Moran index to remove errors from crop yield maps," in *XXVIIIth International Biometric Conference*, 2016.
[18] G. Matheron, *The theory of regionalized variables and its applications*, vol. 5. {É}cole national sup{é}rieure des mines, 1971.
[19] R. Webster and M. A. Oliver, *Geostatistics for Environmental Scientists*. Chichester, UK: John Wiley & Sons, Ltd, 2007.
[20] R. S. Bivand, E. Pebesma, and V. Gómez-Rubio, *Applied Spatial Data Analysis with R*. New York: Springer, 2013.
[21] B. G. Amidan, T. A. Ferryman, and S. K. Cooley, "Data outlier detection using the Chebyshev theorem," in *2005 IEEE Aerospace Conference*, 2005, pp. 3814–3819.
[22] W. Chang, "shinythemes: Themes for Shiny." 2016.
[23] Douglas Nychka, Reinhard Furrer, John Paige, and Stephan Sain, "fields: Tools for spatial data." Boulder, CO, USA, 2015.
[24] R. J. Hijmans, "raster: Geographic Data Analysis and Modeling." 2016.