GEOACTA **40**(1): 54-64 (2015) ISSN 1852-7744

PREDICCIÓN DE VARIABLES REGIONALIZADAS CON DATOS NO DETECTADOS: DOS CASOS DE ESTUDIO

PREDICTION OF REGIONALIZED VARIABLES WITH UNDETECTED DATA: TWO CASES OF STUDY

Morvillo, M. C.[1], Diblasi, A. M.[2], Giménez, M. E.[1], Guerra, E. T.[1], Ruiz, S.B.[1]

[1] Facultad de Ciencias Exactas, Físicas y Naturales. Universidad Nacional de San Juan, Argentina.
[2] Facultad de Ciencias Económicas. Universidad Nacional de Cuyo, Argentina.
E-mail: mmorvillo@hotmail.com

RESUMEN

Los datos “no detectados” se encuentran usualmente en el muestreo de variables georreferenciadas en contextos como la exploración de depósitos minerales, el monitoreo ambiental, o el sondeo de depósito de aguas subterráneas, entre otros. Tales datos son “no detectados” porque sus valores están por debajo del límite de detección, o por arriba de la cota máxima, de los instrumentos de medición. Estos datos extremos, cuando no están bien imputados, pueden producir distorsiones significativas en los mapas de contornos que se elaboran con métodos de interpolación. De hecho, si se produce la predicción en sitios con este tipo de datos antes de imputarlos, resultarían valores intermedios, en clara contraposición a la condición de extremo.

En este trabajo se exhiben dos muestreos con datos no detectados, uno sobre exploración de oro con valores por debajo de un límite de detección y el otro sobre profundidad de la capa freática con parte de sus valores por arriba de la cota superior de medición. Tres criterios -Eliminación, LDI y CNPI – se consideran para imputar en sitios con valores no detectados y sus efectos se comparan en la predicción. Los resultados son más favorables al criterio CNPI.

Palabras Claves: datos no detectados, georreferenciados, imputación CNPI, kriging.

ABSTRACT

Non-detected data are usually found in sampling of georeferenced variables in contexts such as mineral deposit exploration, environmental monitoring, and groundwater reservoir among others. Such data are non-detected because their values are below a detection limit or above the maximum detection level of the instrument used to measure. These extreme data, when they are not properly imputed can produce significant distortions in the contour maps that are made with interpolation methods. In fact, if predicted values in locations with this kind of data are calculated before imputing, would intermediate values, in clear contrast to extreme condition.

In this work, two samplings with non-detected data are shown. One, on gold exploration, with values below a detection level. Another, on the depth of groundwater layer with values above of the maximum detection level of measuring. Three criteria -Elimination, LDI and NCIP - are considered to impute it in places with no detected values and their effects on the prediction are compared. The results are more favorable the CNPI criterion.

Keywords: non-detected georeferenced data, imputation CNPI, kriging.

INTRODUCCIÓN

Los datos no detectados son una clase especial de datos faltantes que suelen ocurrir por la censura sistemática de instrumentos de medición cuando el valor verdadero es menor al límite de detección (LD). También se los conoce como datos censurados “por izquierda”.

Estos datos son frecuentes en estudios de ciencias ambientales y de la tierra relacionados con, por ejemplo, la salud ambiental (Succop et al., 2004), los estudios marinos (Huybrechts et al., 2002), la geoquímica (Buccianti et

al., 2014), la calidad del agua (Shumway et al., 2002), por nombrar sólo algunos. En tales contextos, las mediciones refieren a la proporción en que uno o varios elementos químicos están contenidos en el soporte natural (aire, agua, suelo, rocas, sedimentos, etc.) y aun cuando no hayan sido detectadas, se saben cantidades positivas en el entorno del cero, es decir ubicadas en el intervalo real (0, LD).

La censura también puede ocurrir cuando hay observaciones fuera del rango del instrumento de medida. De hecho, si se utiliza una barra métrica con un valor máximo de 4 metros para medir la profundidad del agua subterránea, cuando una observación exceda ese máximo, sólo se sabrá que supera la cota pero no en qué medida se produce. En estos casos, se dice que el dato está censurado "por derecha".

En consecuencia corresponde asumir que los datos censurados, tanto por "izquierda" como "por derecha", son valores extremos del rango de la variable.

En el contexto geoestadístico, cuyo objetivo es completar la representación de la variable regionalizada en un dominio espacial continuo, elaborando mapas de curvas de nivel, los datos censurados o extremos constituyen un problema serio para el Kriging (ó krigeado), método que se emplea para predecir valores en nuevos sitios. Esto es, pues sin un tratamiento o imputación adecuada en forma previa a la predicción, kriging produciría interpolaciones en los sitios de valores no detectados, con magnitudes intermedias o no extremas, creando imágenes "localmente" falsas o incoherentes respecto a la información que se tiene.

Este razonamiento nos lleva a descartar algunos métodos heurísticos o reduccionistas que suelen contener los softwares para tratar datos no detectados, como listwise deletion (ó ACC: análisis de casos completos) que elimina las unidades (sitios) con valores faltantes, u otros "inapropiados" como imputar fuera del rango de censura, como por ejemplo, con el mismo umbral o límite de detección LD.

Una forma popular, un tanto más razonable para tratar a los datos censurados "por izquierda", ha sido imputar con la mitad del límite de detección (LD/2). Sin embargo trabajos anteriores de Schafer (2002), Palarea et al. (2008, 2013) y Helsel (2012) han demostrado que ese tipo de reemplazos produce efectos distorsivos en la inferencia de los modelos subyacentes.

Más específicamente, considerando la hipótesis de autocorrelación que caracteriza a los procesos espaciales, se han encontrado pocos trabajos de enfoque geoestadístico clásico o frecuentista, pero en su totalidad han abordado y propuesto soluciones a censura "por izquierda" y con un esquema estacionario unidimensional. El más antiguo, de Militino y Ugarte (1999), presentó una versión espacial del algoritmo EM (Esperanza-Maximización) que requiere conocida la función de autocovarianza, pero ello no ocurre en la práctica y constituye una limitación importante para su aplicación. Luego Rathbun (2006) aplicó el algoritmo de Robbins-Monro para estimar los parámetros de un modelo espacial, empleando un muestreo de importancia para obtener simulaciones condicionales de observaciones censuradas, pero tiene problemas de aplicabilidad pues consume un esfuerzo computacional excesivo, especialmente si hay muchos valores censurados. Por último, Schelin y Sjöstedt (2014) propusieron un método relativamente sencillo, llamado "seminaive", pero se limita a variables de rango positivo y pierde sentido en la escala logarítmica recomendada para la optimización del krigeado con variables asimétricas.

Finalmente, en línea con el mismo tipo de censura y aumentando la dimensión y generalidad del modelo espacial, para un vector de dos procesos estocásticos no estacionarios (presencia de tendenciales globales no constantes o derivas), se presenta y evalúa el criterio CNPI (Imputación del Pivote Cercano Corregionalizado), en Morvillo (2012).

En aquel trabajo, se demostró mediante la simulación de los procesos con tendencias polinómicas de orden cuadrático en las coordenadas y sistemas de semivariogramas exponenciales, que el tratamiento con CNPI mantiene las propiedades deseadas para la inferencia hasta con un nivel de censura del 40 % y que no es conveniente imputar con LD/2 en proporciones mayores al 10%. El buen rendimiento de CNPI se puede comprender desde la particularidad que es independiente de la estimación de la tendencia global de la variable -afectada por la censura- y se apoya en la variabilidad espacial de incrementos de variables (variogramas) y en la información que aportan valores vecinos disponibles.

Una novedad y ventaja más de CNPI, que se presenta en este trabajo, es que su formulación es ajustable para resolver problemas de censura "por derecha".

Para ilustrar la aplicabilidad de CNPI a diferentes dimensiones del modelo y tipos de censura en los datos, se presentan dos casos reales: el primer caso referido al muestreo de una sola variable con datos censurados "por derecha" y el otro, sobre observaciones censuradas "por izquierda" en una variable y una covariable con todos los datos disponibles.

Aunque en tales casos reales no es posible medir los errores de la predicción en sitios con datos censurados, se analiza la "coherencia" de la predicción como consecuencia de aplicar CNPI y los métodos heurísticos "descartables", únicos criterios disponibles adaptables a los casos que se ilustran.

Este artículo está organizado de la siguiente manera. En la Sección 2, se detallan la metodología CNPI y las soluciones heurísticas para tratar los datos censurados. Se incluye la definición medidas para evaluar los efectos de los tratamientos en la predicción. En la Sección 3 se presentan los casos de estudio. En la sección 4 se presentan los resultados obtenidos y por último, en la Sección 5, se efectuaron las conclusiones del trabajo.

METODOLOGIA

Los tres criterios que se describen para tratar los datos no detectados se han propuesto por su aplicabilidad a los dos tipos de censura, “por izquierda” y “por derecha”:

- Eliminación: También conocido como análisis de casos completos (ACC). Consiste en eliminar del muestreo los sitios con datos no detectados.
- LDI: Imputación con el límite de Detección (LD). En todos los sitios en que ocurre un valor no detectado, siendo mayor o menor al límite LD, se imputa el valor umbral LD.
- CNPI: Imputación del Pivote Cercano Corregionalizado.

El criterio consiste en asignar un valor acorde menor (o mayor) al límite de detección, considerando la estructura de autocorrelación espacial que ajusta a los datos, según se argumenta a continuación.
En términos formales, sea $Z(s) = [Z_1(s), Z_2(s)]$ un vector estocástico de procesos espaciales cuyas componentes descriptivas son:

$\vec{\mu}(s) = [\mu_1(s), \mu_2(s)]'$, el vector de tendencias espaciales a gran escala, con definición determinística según funciones polinómicas en las coordenadas (x, y) de *s,* de grado menor o igual a dos.

$\vec{\delta}(s) = [\delta_1(s), \delta_2(s)]'$, el vector gaussiano de procesos estocásticos isotrópicos alrededor de las tendencias, con medias cero y funciones de semivariogramas simples γ_1 y γ_2 , y semivariograma cruzado γ_{12}, éstas dependientes de la separación *h* entre pares de sitios y parametrizadas mediante un modelo lineal de corregionalización (MLC), en el cual intervienen las varianzas nugget simples y cruzada τ_1^2 , τ_2^2 y τ_{12}^2 de las variables en la microescala de h -, las varianzas parciales simples y cruzada σ_1^2 , σ_2^2 y σ_{12}^2 y el alcance Φ .

$$\text{MLC}\begin{cases}\gamma_1(h) = \frac{1}{2}\text{var}[\delta_1(s+h) - \delta_1(s)] = \tau_1^2 + \sigma_1^2.g(h/\Phi)\,; \\ \gamma_2(h) = \frac{1}{2}\text{var}[\delta_2(s+h) - \delta_2(s)] = \tau_2^2 + \sigma_2^2.g(h/\Phi)\;\; ; \\ \gamma_{12}(h) = \frac{1}{2}\text{cov}[(\delta_1(s+h) - \delta_1(s)),(\delta_2(s+h) - \delta_2(s))] = \tau_{12}^2 + \sigma_{12}^2.g(h/\Phi)\end{cases} \quad (1)$$

La imputación de los datos no detectados en el proceso $Z_1(s)$, se condiciona al tipo de censura “por izquierda” o “por derecha” y a la dimensión del modelo.

Imputación CNPI “por izquierda”: A partir de la estacionariedad local de dos procesos simultáneos, se ha formulado y ensayado la “Imputación de Pivote Cercano Corregionalizado” para reemplazar valores no detectados “por izquierda”, es decir menores a un límite de detección ($z_1(s) < LD$).

El criterio se apoya en valores vecinos $z_1(s+h)$ disponibles -mayores al límite LD- y en el supuesto de normalidad del vector de incrementos en cada variable - $IZ_1(h) = Z_1(s+h) - Z_1(s)$ y $IZ_2(h) = Z_2(s+h) - Z_2(s)$ - según se detalla en Morvillo (2012):

$$z_1^*(s) = \begin{cases} z_{1,k,min} - \left[\frac{\gamma_{12.}(h)}{\gamma_2(h)}.Iz_2(h) + \sigma_{12}(h).\frac{\phi\left(-\frac{\mu_{1.2.}(h)}{\sigma_{1.2}(h)}\right)}{\psi\left(\frac{\mu_{1.2.}(h)}{\sigma_{1.2}(h)}\right)}\right] & \text{si} \quad Iz_2(h) > 0 \\ z_{1,k,min} - 2.\sqrt{\frac{\gamma_1(h)}{\pi}} & \text{en otro caso} \end{cases} \quad (2)$$

Siendo:

$z_{1,k,min}$, el pivote, valor mínimo disponible en la vecindad de s con k- vecinos próximos (k número natural mayor al total de datos no detectados en la muestra).
h , la distancia entre el pivote $z_{1,k,min}$ y el valor no detectado $z_1(s)$

$$\mu_{1.2}(h) = \frac{\gamma_{12}(h)}{\gamma_2(h)}.Iz_2(h) \qquad ; \qquad \sigma^2_{1.2}(h) = 2.\left(\gamma_1(h) - \frac{\gamma^2_{12}(h)}{\gamma_2(h)}\right)$$

ϕ y ψ , funciones de densidad y distribución de una normal estándar.

Cuando el modelo estructural corresponda a un proceso estocástico univariado, la expresión se reduce a:

$$z^*_1(s) = z_{1,k,min} - 2.\sqrt{\frac{\gamma_1(h)}{\pi}} \tag{3}$$

Imputación CNPI "por derecha": Bajo el mismo marco teórico anterior, para reemplazar valores no detectados mayores al rango de medición ($z_1(s) > LD$), se acude a los vecinos disponibles $z_1(s+h)$ - menores al límite LD – conjuntamente con propiedades de incrementos gaussianos en la variable y covariable,

$$z^*_1(s) = \begin{cases} z_{1,k,max} + \left[-\frac{\gamma_{12.}(h)}{\gamma_2(h)}.Iz_2(h) + \sigma_{12}(h).\frac{\phi\left(-\frac{\mu_{1.2.}(h)}{\sigma_{1.2}(h)}\right)}{\psi\left(\frac{\mu_{1.2.}(h)}{\sigma_{1.2}(h)}\right)} \right] & \text{si} \quad Iz_2(h) < 0 \\ z_{1,k,max} + 2.\sqrt{\frac{\gamma_1(h)}{\pi}} & \text{en otro caso} \end{cases} \tag{4}$$

Siendo $z_{1,k,max}$ el pivote, valor máximo disponible en la vecindad de s con k- vecinos más próximos, ubicado a una distancia h.

Cuando el modelo estructural refiera a solo un proceso estocástico, la expresión se reduce a:

$$z^*_1(s) = z_{1,k,max} + 2.\sqrt{\frac{\gamma_1(h)}{\pi}} \tag{5}$$

Evaluación de Tratamientos

La aplicación de tratamientos a los datos censurados en la muestra constituye una distorsión que afectará tanto a la estimación de parámetros del modelo subyacente como a la predicción.

En la línea de investigación de datos faltantes se ha establecido que la evaluación de los tratamientos no puede estar al margen de la modelización, estimación y procedimientos de prueba en los que encajan.

En tal sentido se realizan estudios de simulación, en los que una vez aplicado el tratamiento a una muestra, que ha sido generada bajo determinado modelo y proporción de censura, se realizan los procesos de inferencia y extraen medidas como errores y verificación de coberturas, Schafer (2002).

En el caso del modelo geoestadístico que contiene gran cantidad de parámetros, pierde sentido analizar la estimación de los parámetros uno por uno, resultando más práctico evaluar la estimación del modelo en su conjunto, a través de la predicción por kriging con validación cruzada- $\hat{Z}_{cv}(s_i)$, de su error estándar $\hat{\sigma}_{cv}(s_i)$ y de la verificación de la cobertura (COV) de los verdaderos valores $z_1(s_i)$ de la muestra por intervalos de la predicción, según se especifica:

$$\mathrm{COV}_{(i)} = \begin{cases} 1 & \text{si} \quad z_1(s_i) \in \left(\hat{z}_{cv}(s_i) - 1.96.\hat{\sigma}_{cv}(s_i); \hat{z}_{cv}(s_i) + 1.96.\hat{\sigma}_{cv}(s_i)\right) \\ 0 & \text{en otro caso} \end{cases} \tag{6}$$

El porcentaje de cobertura real (COVP) se obtiene promediando la verificación en todos los sitios de la muestra:

$$\mathrm{COVP} = \frac{\sum_{i=1}^{n} \mathrm{COV}_{(i)}}{n} \tag{7}$$

Bajo el supuesto de normalidad, la cobertura esperada de los intervalos basada en los predichos $\hat{z}_{cv}(s_i)$ y su desviación estándar $\hat{\sigma}_{cv}(s_i)$ es del 95%.

Por esa vía, en el trabajo Morvillo (2012) se ha demostrado con simulaciones basadas en un modelo espacial bivariado, que el criterio CNPI valida la cobertura nominal de las predicciones, siempre que el porcentaje de sitios con valores no detectados no exceda el 40%.

Pero en este trabajo, cuyo objetivo es aplicar en casos reales, no es posible verificar la cobertura de la predicción en los sitios con datos censurados pues no se tienen disponibles sus valores. Sin embargo, puede evaluarse la coherencia de tales intervalos con el intervalo de pertenencia o censura $(0, \mathrm{LD})$ ó (LD, ∞) según se define a continuación.

En términos simbólicos, cuando el límite de detección es por izquierda, la coherencia (COH) de la predicción en sitios con datos censurados o no disponibles (*na*) se verificará mediante:

$$\mathrm{COH}_{(i)} = \begin{cases} 1 & \text{si} \quad \hat{z}_1(s_i) - 1.96.\hat{\sigma}_1(s_i) < \mathrm{LD} \quad \text{para} \quad z_1(s_i) < \mathrm{LD} \\ 0 & \text{en otro caso} \end{cases} \tag{8}$$

En tanto que si el límite se encuentra por derecha corresponde la verificación con:

$$\mathrm{COH}_{(i)} = \begin{cases} 1 & \text{si} \quad \hat{z}_1(s_i) + 1.96.\hat{\sigma}_1(s_i) > \mathrm{LD} \quad \text{para} \quad z_1(s_i) > \mathrm{LD} \\ 0 & \text{en otro caso} \end{cases} \tag{9}$$

Por último el promedio de las coherencias en los na- sitios con datos no disponibles –*COHP*- indicará la proporción de sitios censurados en que parte del intervalo de predicción tiene coincidencia con el intervalo de censura, cuyo valor ideal es como mínimo 0.975 :

$$\mathrm{COHP} = \frac{\sum_{t=1}^{na} \mathrm{COH}_{(t)}}{na} \tag{10}$$

CASOS DE ESTUDIO

A continuación se exponen dos situaciones reales de muestreo georreferenciado, correspondientes a diferentes temáticas, en los que se manifiesta el problema de datos no detectados.

Archivo de Agua Subterránea

Estos datos provienen del Instituto Nacional de Tecnología Agropecuaria (INTA), Estación Experimental Agropecuaria San Juan. La medición corresponden al mes noviembre de 2012, de la profundidad de la capa

freática en 198 sitios de una zona urbana del departamento Sarmiento, provincia de San Juan, R.A. El archivo contiene tres columnas: la profundidad V medida en metros y las coordenadas (x,y) de cada sitio s medidas en el sistema de proyección Gauss-Krüger.

La cota superior del instrumento empleado en las mediciones fue de 4 metros y en 43 sitios, esto es el 22% en la muestra, la profundidad fue superior a ésta, registrándose con el umbral "4". Tales cifras no corresponden a mediciones reales y siguiendo el criterio de notación del caso anterior correspondería haberlos registrado con otro símbolo, por ejemplo, "+4", pues son "mayores a 4". En este caso estamos frente a datos no detectados "por derecha" que pueden ser ajustados a un modelo espacial unidimensional.

En el mapa de posiciones del muestreo (Figura 1A) se han destacado con rojo los sitios que contienen datos no detectados, considerando las coordenadas en unidades de km.

El Análisis Exploratorio de Datos Espaciales (AEDE) del caso confirmó las hipótesis de normalidad y de autocorrelación espacial para la escala logarítmica de la profundidad $\ln V = Z$.

Para tal escala del proceso espacial también se verificó el ajuste a un modelo estructural con tendencia global cuadrática en las coordenadas (x, y) y semivariograma de tipo esférico según las especificaciones siguientes:

$$Z(s) = \mu(s) + \delta(s) \tag{11}$$

Donde:

- $\mu(s) = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 y^2 + \beta_5 xy$ la media global del proceso determinada por un polinomio de segundo grado en las coordenadas xy, que representa la tendencia de la profundidad de la capa freática –en escala logarítmica- en el dominio geográfico.
- $\delta(s)$ la componente estocástica estacionaria gaussiana en torno a la tendencia, con autocorrelación espacial determinada por un semivariograma esférico:
-

$$\gamma(h) = \tau^2 + \frac{\sigma^2}{2}\left[3\frac{h}{\Phi} - \left(\frac{h}{\Phi}\right)^3\right] \tag{12}$$

Por cuestiones técnicas del proceso de estimación las coordenadas espaciales se redujeron a menor escala. A continuación, en la Tabla 1 figuran las estimaciones obtenidas en base a las observaciones disponibles, aplicando el método de máxima verosimilitud restringida (REML), con el paquete geoR del entorno R de distribución libre (http://www.r-project.org).

Parámetro	β_0	β_1	β_2	β_3	β_4	β_5	τ	σ	Φ
Estimador	46.118	-0.425	-0.135	0.0105	0.0011	0.0070	0.15	0.047	9

Tabla 1. Parámetros estimados del modelo estructural ajustado a la escala logarítmica de la profundidad de la capa freática.
Table 1. Estimates parameters for the structural model adjusted for the logarithmic scale of the depth of groundwater table.

Archivo de Mineral

Estos archivos, corresponden a una exploración de oro de una empresa privada, cuyas coordenadas se expresan transformadas y sin referencia métrica, a fin de respetar el derecho de propiedad de la información. Contienen la composición química de muestras de roca en 99 sitios, con 31 observaciones de oro no detectadas registradas con el símbolo "-1". Estas requieren la imputación con valores verosímiles antes de trazar las curvas de nivel del yacimiento. En la Figura 1B se ilustran las posiciones del muestreo contrastando en color rojo a los sitios con datos no detectados.

Además de las coordenadas (x, y) de los sitios y del elemento oro (Au) objetivo del yacimiento –conteniendo valores no detectados- , el archivo también contiene otros elementos químicos con observaciones completas que sirven de covariables, tales como: arsénico (As), bario (Ba), cobre (Cu), manganeso (Mn), plomo (Pb) y zinc (Zn).

El análisis exploratorio de los datos espaciales (AEDE) confirmó hipótesis de distribución log-normal y de autocorrelación espacial para todos los elementos.

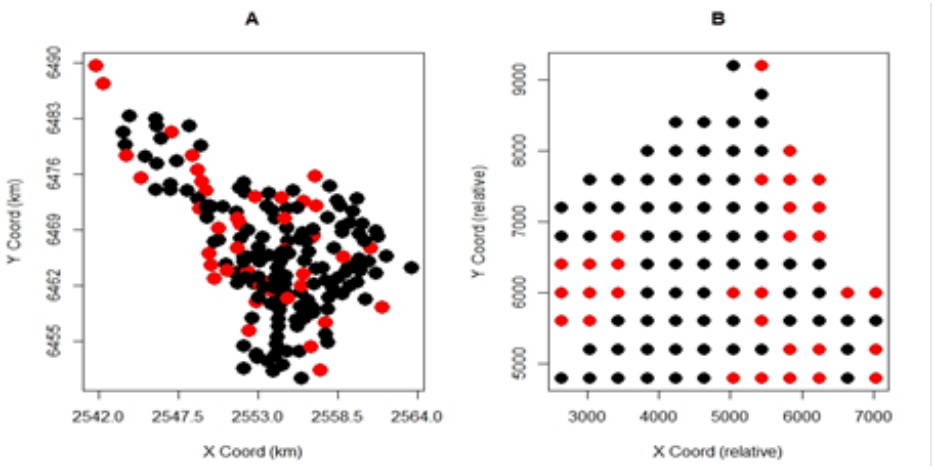

Figura 1. Distribución espacial de observaciones de profundidad de la capa freática (A) y de contenido de oro (B), con valores no detectados en rojo.
Figure 1. Spatial distribution of the observed data depth of groundwater layer (A) and gold content (B), with non-detected values in red color.

Además, para emprender la predicción espacial del oro, se determinó un modelo estructural para un proceso espacial bidimensional $Z(s)=[Z_1(s),Z_2(s)]$ con componentes correlacionadas, $Z_1(s)=\ln Au(s)$ y $Z_2(s)=\ln As(s)$, determinados por tendencias polinómicas cuadráticas y semivariogramas simples y cruzado del tipo esférico, con las siguientes especificaciones:

$$[Z_1(s),Z_2(s)]=[\mu_1(s),\mu_2(s)]+[\delta_1(s),\delta_2(s)] \tag{13}$$

Siendo:

$\mu_1(s)=\beta_{01}+\beta_{11}x+\beta_{21}y+\beta_{31}x^2+\beta_{41}y^2+\beta_{51}xy$, media global del oro –en escala logarítmica- en el dominio geográfico, ajustada por un polinomio de segundo grado en las coordenadas (x,y).

$\mu_2(s)=\beta_{02}+\beta_{12}x+\beta_{22}y+\beta_{32}x^2+\beta_{42}y^2+\beta_{52}xy$, media global del arsénico–en escala logarítmica- en el dominio geográfico, ajustada por un polinomio de segundo grado en las coordenadas (x,y).

$\delta_1(s),\delta_2(s)$, componentes estocásticas gaussianas estacionarias del oro y arsénico respectivamente, definidas en torno a sus tendencias, con semivariogramas simples y cruzados, dados por:

$$\begin{cases}\gamma_i(h)=\tau_i^{\,2}+\dfrac{\sigma_i^{\,2}}{2}\left[3\dfrac{h}{\Phi}-\left(\dfrac{h}{\Phi}\right)^3\right] & ,i=1,2\\ \gamma_{12}(h)=\tau_{12}^{\,2}+\dfrac{\sigma_{12}^{\,2}}{2}\left[3\dfrac{h}{\Phi}-\left(\dfrac{h}{\Phi}\right)^3\right]\end{cases} \tag{14}$$

La Tabla 2 contiene las estimaciones REML obtenidas para los parámetros del modelo bidimensional, los que también para este caso, corresponden a las coordenadas reducidas:

Parámetros	β_0	β_1	β_2	β_3	β_4	β_5	τ	σ	Φ
Est. para Au	25.766	0.0672	-0.0007	-0.0034	-0.0027	0.0037	0.448	0.716	20.16
Est. para As	44.144	0.1031	-0.0858	-0.0041	-0.0020	0.0051	0.383	0.491	20.16
Est.| Au x As							0.264	0.570	20.16

Tabla 2. Parámetros estimados del modelo estructural ajustado a las escalas logarítmicas del oro y el arsénico.
Table 2. Estimates parameters of the structural model fitted to the logarithmic scales of gold and arsenic.

RESULTADOS

Una vez examinadas las condiciones de cada archivo, se trataron los datos no detectados con cada uno de los criterios Eliminación, LDI y CNPI.

Las Figuras 2 y 3 contienen, para cada caso de estudio, el mapa inicial (a izquierda) de la muestra con los datos imputados según cada criterio, Eliminación (arriba), LDI (centro) y CNPI (abajo), representando las magnitudes con escalas proporcionales de color y de tamaño. En el sector derecho de ambas figuras, figuran las curvas de nivel resultantes de la predicción aplicada a la muestra tratada con cada criterio.

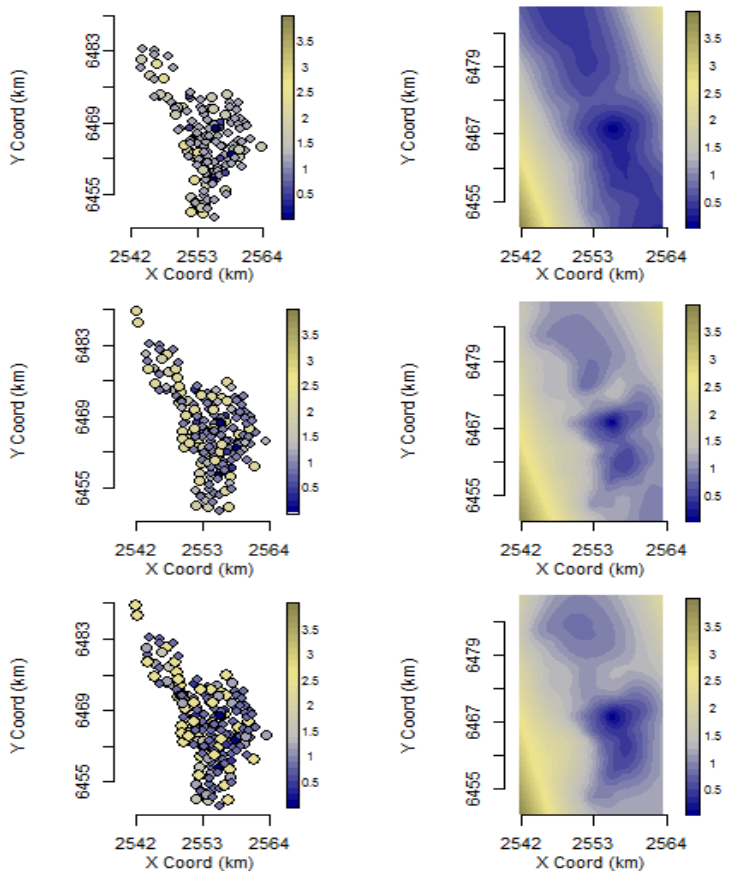

Figura 2. Mapas de predicciones de la profundidad de la capa freática (derecha) basadas en las muestras tratadas (izquierda) con los criterios de "Eliminación" (arriba), "LDI" (centro) y "CNPI" (abajo)
Figure 2. Prediction maps of groundwater layer depth (right) based on data with imputed values under the “Elimination” (top), “LDI” (center), and “CNPI” (bottom) criteria.

Las curvas de nivel obtenidas de la muestra tratada con cada criterio evidencian diferencias en las vecindades de los sitios con datos no detectados. Desde un punto de vista gráfico y global, sobresale la disimilitud de la imagen basada en la muestra incompleta (Eliminación), mientras que las diferencias entre las imágenes obtenidas con las imputaciones de LDI y CNPI son menos evidentes.

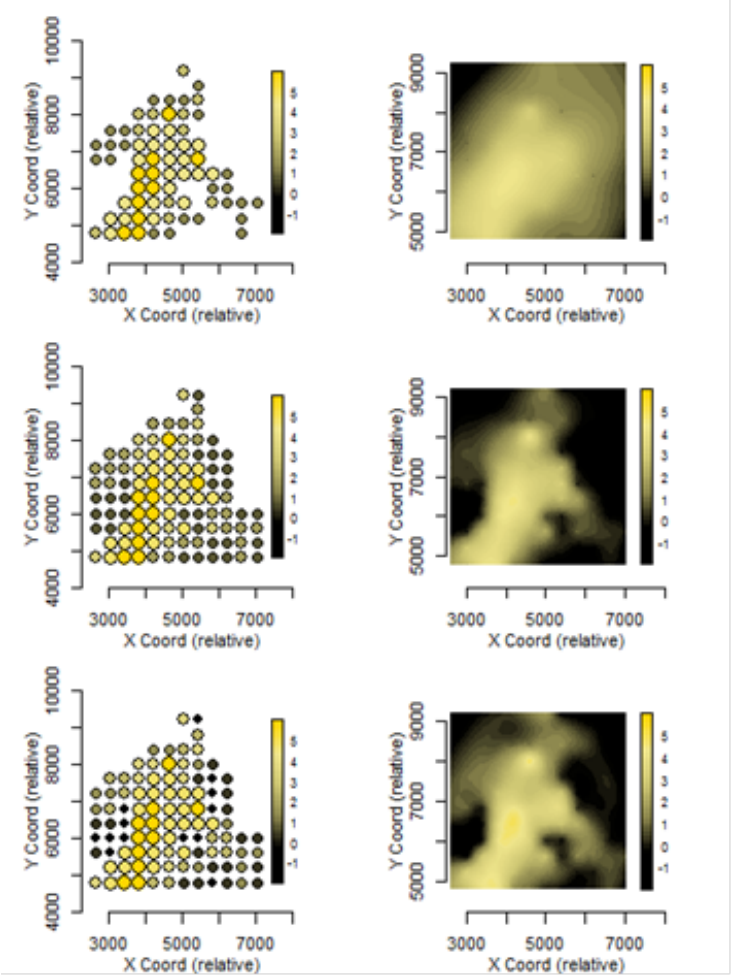

Figura 3. Mapas de predicciones del oro (derecha) basadas en las muestras tratadas (izquierda) con los criterios de Eliminación (arriba), LDI (centro) y CNPI (abajo).
Figure 3. Prediction maps of gold (right) based on data with imputed values under the "Elimination" (top), "LDI" (center), and "CNPI" (bottom) criteria.

Para una conclusión más formal y cuantitativa, evaluando el impacto de los tratamientos en la coherencia de la predicción respecto del intervalo de pertenencia o censura, en sitios con ese tipo de datos no detectados o no disponibles (na), se informan en las Tablas 3 y 4, para cada caso de estudio, los niveles coherencia (COHP) obtenidos, incluyendo LD límite de detección, P proporción.

LD	P	Medida	Eliminación	LDI	CNPI
0	31%	COHP	0,548	0,903	1

Tabla 3. Porcentaje de coherencia en las predicciones del agua subterránea basadas en las muestras con datos no detectados tratados con los distintos criterios de Eliminación, LDI y CNPI.
Table 3. Coherence percentage in prediction of groundwater layer depth based on non-detected data imputed with the Elimination, LDI, and CNPI criteria.

LD	P	Medida	Eliminación	LDI	CNPI
1,38	22%	COHP	0,814	0,953	0,977

Tabla 4. Porcentaje de coherencia de las predicciones del oro basadas en las muestras con datos no detectados tratados con los distintos criterios de Eliminación, LDI y CNPI.
Table 4. Coherence percentage in prediction of gold based on non-detected data imputed with the Elimination, LDI, and CNPI criteria.

Para los casos de agua y mineral, la medida de coherencia resultó en los niveles más bajos, de 81,4% y 54,8% respectivamente, cuando se aplicó el tratamiento de Eliminación. Recíprocamente, podría decirse que la eliminación de datos no detectados ha producido predicciones incoherentes en una proporción importante de sitios en los que hay censura. Un tanto mejor fueron los resultados del tratamiento con LDI alcanzando para cada caso los niveles de coherencia del 95,3 % y 90,3% respectivamente, sin embargo siguen siendo insuficientes pues el nivel mínimo esperado para la coherencia es del 97,5 %. Por último la aplicación del criterio CNPI ha mostrado niveles "aceptables" de coherencia (97,7 % y 100 %), resultando el único criterio que cumple con las expectativas de la predicción bajo los niveles de censura tratados, en línea con otros estudios anteriores.

CONCLUSIONES

En este trabajo se ha presentado el problema de valores no detectados, menores o mayores a un límite de detección LD, en dos archivos de datos georreferenciados que responden a supuestos de correlación espacial.

Desde la perspectiva de la predicción espacial, se ha planteado la necesidad de imputar los datos no detectados con valores acordes al intervalo extremo censurado, pues de lo contrario el método de interpolación genera en esas vecindades, valores típicos contrarios (incoherentes) a los reales extremos.

Se propusieron y aplicaron tres criterios para tratar los datos no detectados y se analizaron sus efectos en la predicción. Uno de ellos, CNPI (imputación de pivote cercano corregionalizado) incluye conceptos de autocorrelación espacial para procesos uni o bivariados y es adaptable a límites de detección por izquierda y por derecha.

El porcentaje de coherencia definido para evaluar el rendimiento de los tratamientos arroja mejores resultados del criterio CNPI en el archivo mineral, aun cuando presenta mayor porcentaje de valores no detectados. Este aspecto puede explicarse por la disposición de los datos no detectados en el borde del dominio geográfico y por efectos inherentes al método de predicción.

La evaluación de las predicciones en estas aplicaciones da crédito al criterio CNPI, para imputar datos no detectados en un contexto espacial.

REFERENCIAS

Buccianti, A., Nisi, B., Martín-Fernández, J. A. y Palarea-Albaladejo, J., (2014). Methods to investigate the geochemistry of groundwaters with values for nitrogen compounds below the detection limit. Journal of Geochemical Exploration, 141: 78-88.

Helsel, D. R., (2012). Statistics for Censored Environmental Data Using Minitab and R, John Wiley & Sons.

Huybrechts, T., Dewulf, O., y Van Langenhove, H., (2002). How to estimate moments and quantiles of environmental data sets with non-detect observations? A case study of volatile organic compounds in marine water samples. Journal of Chromatography, A , 975:123-133.

Militino, A. y Ugarte, M., (1999). Analyzing censored spatial data. Mathematical Geology, 31: 551-561.

Morvillo, M., (2012). Imputación para datos no detectados en un contexto espacial, Tesis para Magister en Estadística Aplicada, Universidad Nacional de Córdoba. Disponible en línea: http://www2.famaf.unc.edu.ar/institucional/biblioteca/trabajos/921/16912.pdf.

Palarea-Albaladejo, J. and J. A. Martín-Fernández, (2008). A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. Computers & Geosciences, 34: 902-917.

Palarea-Albaladejo, J. y Martín-Fernández, J. A., (2013). Values below detection limit in compositional chemical data. Analytica Chimica Acta, 764: 32-43.

Rathbun, S.L., (2006). Spatial prediction with left-censored observations. Journal of Agricultural, Biological, and Environmental Statistics, 11: 317-336.

Shumway, R.H., Azari, R. y Kayhanian, M., (2002). Statistical approaches to estimating mean water quality concentrations with detection limits. Environmental Science & Technology, 36: 3345-3353.

Schafer, J.L. y Graham, J., (2002). Missing data: our view of the state of the art. Psychological Methods, 7: 147-177.
Schelin, L. y Sjostedt-de Luna, S., (2014). Spatial prediction in the presence of left-censoring. Computational Statistics and Data Analysis, 74: 125-141.
Succop, P.A., Clark, S., Chen, M. y Galke, E., (2004). Imputation of data values that are less than a detection limit. Journal of Occupational and Environmental Hygiene, 1:7, 436-441.

Recibido: 6-04-2015
Aceptado: 15-07-2015