COMPARACIÓN DE MODELOS LINEALES Y NO LINEALES PARA ESTIMAR EL RIESGO DE CONTAMINACIÓN DE SUELOS

COMPARISON OF LINEAR AND NONLINEAR MODELS TO ESTIMATE THE RISK OF SOIL CONTAMINATION

Nancy **Toriz-Robles**[1*], Martha E. **Ramírez-Guzmán**[1], Yolanda M. **Fernández-Ordoñez**[1], Jesús **Soria-Ruiz**[2], María C. **Ybarra Moncada**[3]

[1]Colegio de Postgraduados, Campus Montecillo. Carretera México-Texcoco, Km. 36.5, Montecillo, Estado de México. [2]Instituto Nacional de Investigaciones Forestales y Agropecuarias, Campo Experimental Valle de Toluca, Zinacantepec, Estado de México. [3]Universidad Autónoma Chapingo, Carretera México-Texcoco, Km. 38.5, Estado de México. (toriz.nancy@colpos.mx)

RESUMEN

El estudio de datos de contaminantes en áreas geográficas se caracteriza por la dependencia espacial, distribución no normal y heteroscedasticidad. Pero, estas características no se han considerado en la modelación de datos edafológicos. Por lo anterior, en este estudio se analizó y comparó el comportamiento de estimadores de modelos de regresión lineal generalizados (GLM), lineales generalizados mixtos (GLMM), aditivos generalizados (GAM) y aditivos generalizados mixtos (GAMM) a través de la simulación de una variable respuesta generada con distribuciones estadísticas diferentes, con cinco tipos de matrices de pesos (W, B, C, U y S) y niveles diferentes de autocorrelación. Los resultados mostraron que la matriz de vecindad U fue robusta a todos los niveles de autocorrelación espacial. Como se esperaba, los modelos GAM y GAMM fueron superiores a GLM y GLMM, debido a su flexibilidad representada por las funciones de suavización (splines) y la incorporación de efectos mixtos. Mapas de predicción de concentración de metales pesados y de probabilidad de riesgo de exceder los límites permisibles se elaboraron para el Valle del Mezquital, Hidalgo.

Palabras clave: metales pesados, autocorrelación, distribución no normal, heterocedasticidad, modelos generalizados mixtos, modelos aditivos mixtos.

ABSTRACT

The study of pollution in geographical areas includes spatial dependence, non-normal distribution, and heteroscedasticity. However, the modelling of edaphological data has not taken these features into consideration. Therefore, this study included the analysis and comparison of the behavior of the estimators of generalized linear regression (GLM), generalized linear mixed (GLMM), generalized additive (GAM), and generalized additive mixed (GAMM) models, through the simulation of a response variable generated with different statistical distributions, with five weighing matrixes (W, B, C, U, and S) and several autocorrelation levels. The results showed a strong U-adjacency matrix for all spatial autocorrelation levels. As was expected, GAMs and GAMMs were higher than GLMs and GLMMs, as a consequence of their flexibility which is represented by smoothing splines and the incorporation of mixed effects. The concentration of heavy metals and the risk probability of surpassing permissible limits in the Mezquital Valley, Hidalgo, were subject to prediction mapping.

Key words: Heavy metals, autocorrelation, non-normal distribution, heteroscedasticity, generalized mixed models, additive mixed models.

INTRODUCCIÓN

El Valle del Mezquital, comprende los Distritos de Riego (DR) 003 Tula, 100 Alfajayucan y 112 Ajacuba en el estado de Hidalgo, y recibe los

INTRODUCTION

The Mezquital Valley, State of Hidalgo, includes the following irrigation districts (DRs): 003 Tula, 100 Alfajayucan, and 112 Ajacuba. The tributary streams of Mexico City and its conurbation are used for agricultural irrigation in this valley. This problem has been studied for over a century using models that are based on the classic linear regression

*Autor responsable ❖ Author for correspondence.
Recibido: febrero, 2017. Aprobado: octubre, 2017.
Publicado como ARTÍCULO en Agrociencia 53: 269-283. 2019.

afluentes de la Ciudad de México y áreas conurbanas, que se usan para riego agrícola. Éste es un problema que se ha estudiado desde hace más de un siglo mediante modelos que asumen los supuestos clásicos de la regresión lineal. Debido a que la distribución de contaminantes como metales pesados es asimétrica, es necesario implementar metodologías nuevas para mejorar la estimación y predicción de los niveles de contaminación. Por lo anterior, el objetivo del presente estudio fue comparar modelos que permitieran incorporar información propia de datos espaciales mediante un estudio de simulación. Los resultados se aplicaron al Valle del Mezquital.

MATERIALES Y MÉTODOS

Los modelos que consideramos fueron lineales y no lineales.

Modelos Lineales Generalizados

Los Modelos Lineales Generalizados (GLM, del inglés Generalized Linear Models) (Nelder y Wedderburn, 1972) permiten modelar variables de respuesta con distribuciones pertenecientes a la familia exponencial, la que incluye distribuciones continuas y discretas. Una distribución pertenece a la familia exponencial si su función de densidad puede escribirse como:

$$f_\theta(y) = \exp\left[\frac{y\theta - b(\theta)}{\alpha(\phi)} + c(y,\phi)\right] \qquad (1)$$

donde a, b y c son funciones arbitrarias, θ es un parámetro natural de la distribución y ϕ es un parámetro de escala. $E[y]=\mu$, donde μ es la media de y, la cual puede depender de covariables a través de la función:

$$g(\mu)=\eta \qquad (2)$$

donde $\eta=X\beta$ es el predictor lineal, g es una función conocida (función liga), X es la matriz diseño y β un vector de parámetros. Los modelos GLM están diseñados para modelar datos independientes.

Modelos Lineales Generalizados Mixtos

Los Modelos Lineales Generalizados Mixtos (GLMM, del inglés Generalized Linear Mixed Models) se obtienen a partir de los GLM, con la incorporación de efectos aleatorios en los predictores lineales (Breslow y Clayton, 1993). La incorporación de efectos aleatorios permite modelar sobredispersión, heterocedasticidad y correlación espacial o temporal (McCulloch, 1997;

assumptions. Since heavy metals have an asymmetrical distribution, new methodologies are required to improve how pollution levels are estimated and predicted. Therefore, this study compared models that incorporated typical spatial data information through a simulation study. The results were applied to the Mezquital Valley.

MATERIALS AND METHODS

Linear and nonlinear models were taken into consideration.

Generalized Linear Models

Generalized Linear Models (GLMs) enable the modelling of response variables with distributions that belong to the exponential family, which includes uniform and discrete distribution (Nelder and Wedderburn, 1972). A distribution belongs to the exponential family, if its destiny function can be described as follows:

$$f_\theta(y) = \exp\left[\frac{y\theta - b(\theta)}{\alpha(\phi)} + c(y,\phi)\right] \qquad (1)$$

where a, b, and c are arbitrary functions; θ is a natural distribution parameter; and ϕ is a scale parameter. $E[y]=\mu$, where μ is the mean of y, which may depend on covariables, according to the following function:

$$g(\mu)=\eta \qquad (2)$$

where $\eta=X\beta$ is the linear predictor; g is a known function (link function); X is the design matrix; and β is a parameter vector. GLMs have been designed to model independent data.

Generalized Linear Mixed Models

Generalized Linear Mixed Models (GLMMs) are obtained from GLMs, incorporating random effects into linear predictors (Breslow and Clayton, 1993). Incorporating random effects enables the modelling of overdispersion, heteroscedasticity, and spatial or temporal correlation (McCulloch, 1997; Torabi, 2015). The structure of a GLMM is given by the following expression:

$$g(\mu)=\eta=X\beta+Zy \qquad (3)$$

where X, β, μ, g, and η are defined in (1) and (2); Z is the design matrix for the random effects; $\gamma \sim N(0, \psi)$ is the vector that contains the random effects; and ψ is the covariance matrix of the effects.

Torabi, 2015). La estructura de un GLMM está dada por la expresión:

$$g(\mu)=\eta=X\beta+Zy \quad (3)$$

donde X, β, μ, g y η son definidos en (1) y (2), Z es la matriz diseño de los efectos aleatorios, $\gamma \sim N(0, \psi)$ es el vector que contiene los efectos aleatorios y ψ es la matriz de covarianzas de los efectos.

Modelos Aditivos Generalizados

Los Modelos Aditivos Generalizados (GAM, del inglés Generalized Additive Models), incluyen funciones de suavizamiento asociadas a las k covariables del predictor lineal. Estos modelos se consideran modelos de regresión no paramétrico (Hastie y Tibshirani, 1986) y su estructura general es:

$$g(\mu)=\eta=X\beta+f_1(x_1)+f_2(x_2)+\ldots+f_k(x_k) \quad (4)$$

donde $\mu, g, \eta, X,$ y β y son definidos en (1, 2 y 3) y f_k son funciones de suavizamiento continuas.

Una función de suavizamiento representa la tendencia de la variable respuesta en función de una covariable. Dicha función depende de las observaciones de un punto dado y de las observaciones vecinas. Entre las técnicas de suavizamiento más empleadas están las funciones polinómicas (en inglés conocido como *splines*) a través de puntos llamados nodos. Estos puntos dividen el rango de x en regiones. Los *splines* dependen de: grado del polinomio, número de nodos y localización de los nodos. La función más utilizada es el *spline* cúbico, la que es una curva construida a través de polinomios de tercer grado alrededor de cada nodo, los cuales se ensamblan para formar una curva continua. Su utilidad radica en que tienen segundas derivadas continuas y puntos de inflexión (Wood, 2006; Liu, 2008; Mamouridis, 2011).

Modelos Aditivos Generalizados Mixtos

Los Modelos Aditivos Generalizados Mixtos (GAMM, del inglés Generalized Additive Mixed Models), son una extensión de los GAM mediante la incorporación de efectos aleatorios en los predictores lineales (Lin y Zhang, 1999). El modelo general es:

$$g(\mu)=\eta=X\beta+f_1(x_1)+f_2(x_2)+\ldots+f_k(x_k)+Z\gamma \quad (5)$$

donde $\mu, g, \eta, X, \beta, Z, \gamma$ y f_k ya se definieron previamente.

Generative Additive Models

Generative Additive Models (GAMs) include smoothing splines that are associated with the k covariables of the linear predictor. These models are considered as nonparametric regression models (Hastie and Tibshirani, 1986) and they have the following general structure:

$$g(\mu)=\eta=X\beta+f_1(x_1)+f_2(x_2)+\ldots+f_k(x_k) \quad (4)$$

where μ, g, η, X, and β are defined in (1, 2, and 3) and f_k are continuous smoothing splines.

A smoothing spline represents the trend of the response variable according to a covariable. The said function depends on the observations of a given point and of the adjacent observations. The most frequently used smoothing techniques include polynomial functions (splines) through points known as nodes. These points divide the range of x in regions. Splines depend on: the degree of the polynomial, the number of nodes, and the location of the nodes. The most frequently used function was the cubic spline: a curve developed through third degree polynomials around each node, which are assembled from a continuous curve. Its usefulness lays in its second continuous derivatives and inflection points (Wood, 2006; Liu, 2008; Mamouridis, 2011).

Generalized Additive Mixed Models

Generalized Additive Mixed Models (GAMMs) extend GAMs incorporating random effects into linear predictors (Lin and Zhang, 1999). The general model is the following:

$$g(\mu)=\eta=X\beta+f_1(x_1)+f_2(x_2)+\ldots+f_k(x_k)+Z\gamma \quad (5)$$

where μ, g, η, X, β, Z, γ, and f_k have been already defined.

Moran index

This index indicates the degree of spatial association of the y variable (Moran, 1948). Its values fluctuate between -1 and 1. A value of zero indicates a random spatial process. It is expressed as follows:

$$I=\frac{n\sum_{t=1}^{n}\sum_{r=1}^{n}W_{tr}(y_t-\bar{y})(y_r-\bar{y})}{\sum_{t=1}^{n}\sum_{r=1}^{n}W_{tr}\sum_{t=1}^{n}(y_t-\bar{y})^2} \quad (6)$$

Índice de Moran

Este índice indica el grado de asociación espacial de la variable *y* (Moran, 1948). Sus valores oscilan entre −1 y 1, y el valor cero es indicativo de un proceso espacial aleatorio. Se expresa como:

$$I = \frac{n\sum_{t=1}^{n}\sum_{r=1}^{n} W_{tr}(y_t - \bar{y})(y_r - \bar{y})}{\sum_{t=1}^{n}\sum_{r=1}^{n} W_{tr}\sum_{t=1}^{n}(y_t - \bar{y})^2} \qquad (6)$$

donde n es el número de observaciones, y_t es la variable de respuesta en el punto t, $\bar{y}$ es la media de y_t, y W_{tr} es el valor de la matriz de pesos asociado al punto t con respecto al punto r. La matriz de pesos indica si una región (objeto o dato) es vecina espacial de otra, la matriz debe ser de tamaño $n \times n$, con elementos (t, r). De forma convencional se ha considerado que los elementos diagonales de la matriz W_{tr}, tienen valor cero cuando el punto t es el punto r. La matriz de pesos puede obtenerse en diferentes maneras. Las cinco más representativas, proporcionadas por la paquetería spdep del software R, son las siguientes.

Matriz tipo W. En la matriz tipo W, también llamada "fila de normalización", las filas suman la unidad y cada peso por fila tiene el mismo valor. Sean $\tilde{w}_t$ el total de vecindades entre t y r en la fila t, el peso para cada vecindad (w_{tr}) se define como: $w_{tr} = \frac{1}{\tilde{w}_t}$.

Matriz tipo B. La matriz tipo B se conoce como "matriz binaria" y consiste en dar el valor de la unidad cuando entre t y r son vecinos y cero de otra manera: $w_{tr}=1$ si es vecino y $w_{tr}=0$ de otro modo. Así, la suma de los elementos de la matriz w_{tr} es el número total de vecindades.

Matriz tipo C. Sea n el número de regiones y $\tilde{w}_{tr}$ el número total de vecindades, entonces los elementos de la matriz de pesos se definen como: $w_{tr} = \frac{n}{\tilde{w}_{tr}}$.

Matriz tipo U. En esta matriz, el valor de cada uno de los elementos de la matriz está determinado por la división entre la unidad y el número total de vecindades, es decir: $w_{tr} = \frac{1}{\tilde{w}_{tr}}$.

Matriz tipo S. La matriz tipo S es un esquema de codificación propuesto por Tiefelsdorf *et al.* (1999). En él se ponderan los valores de la suma de los pesos de cada fila y dicho valor es dividido entre el número de vecindades. Filas con el mismo número de vecindades tienen pesos iguales y la suma de los pesos ponderados de las filas resulta en el número total de regiones. Además, se cumple que filas con número mayor de vecindades tengan peso mayor que aquellas con número menor. Es decir: $w_{1r} = \frac{a_1}{\tilde{w}_{1r}}, w_{2r} = \frac{a_2}{\tilde{w}_{1r}}, \ldots, w_{nr} = \frac{a_n}{\tilde{w}_{nr}}$, donde a_t son las ponderaciones por fila tal que: $a_1 + a_2 + \ldots a_n = n$.

where n is the number of observations; y_t is the response variable in the t point; $\bar{y}$ is the mean of y_t; and W_{tr} is the value of the weighing matrix associated with the t point in relation to the r point. The weighing matrix indicates if a region (object or datum) is spatially adjacent to another: the size of the matrix must be $n \times n$ with (t, r) elements. The diagonal elements of the W_{tr} matrix have been conventionally considered to have a value of zero, when the t point is the r point. There are several ways to obtain a weighing matrix. According to the spdep package from the R software, the five most representative matrices are the following:

Type-W Matrix. In a type-W matrix (also known as a "normalizing row"), the rows add up to the unity and the weight of each row has the same value. When $\tilde{w}_t$ is the total number of adjacencies between t and r in the t row, the weight of each adjacency (w_{tr}) is defined as: $w_{tr} = \frac{1}{\tilde{w}_t}$.

Type-B Matrix. The type-B matrix is known as a "binary matrix" and it provides the value of the unit when t and r are adjacent or otherwise have a value of zero: $w_{tr}=1$ (adjacent) and $w_{tr}=0$ (otherwise). Therefore, the sum of the elements of the w_{tr} matrix is the total number of adjacencies.

Type-C Matrix. When n is the number of regions and $\tilde{w}_{tr}$ is the total number of adjacencies, the elements of the weighing matrix are defined as follows: $w_{tr} = \frac{n}{\tilde{w}_{tr}}$.

Type-U Matrix. The value of each of the elements of this matrix is determined by the division of the unit and the total number of adjacencies, as follows: $w_{tr} = \frac{1}{\tilde{w}_{tr}}$.

Type-S Matrix. The type-S matrix is a coding scheme proposed by Tiefelsdorf *et al.* (1999). It is used to weigh the values of the sum of the weights of each row, before dividing the said value between the number of adjacencies. Rows with the same number of adjacencies have equal weight and the sum of the weighted weights of the rows is the total number of regions. Additionally, the rows with the highest number of adjacencies are heavier than those with less adjacencies. *I.e.*: $w_{1r} = \frac{a_1}{\tilde{w}_{1r}}, w_{2r} = \frac{a_2}{\tilde{w}_{1r}}, \ldots, w_{nr} = \frac{a_n}{\tilde{w}_{nr}}$, where a_t are the weighting factor per row, so that: $a_1 + a_2 + \ldots a_n = n$.

Simulation studies

The simulation attempted to imitate the potential behavior of a certain concentration of heavy metals in the soil, based on the concentration of six pollutants in the Mezquital Valley, Hidalgo. The following conditions were applied to the simulation analysis:

Estudio de simulación

En la simulación se trató de imitar el comportamiento que podría llegar a tener la concentración de metales pesados en el suelo, partiendo de la concentración de seis contaminantes en el Valle del Mezquital, Hidalgo. Las condiciones del análisis de simulación fueron las siguientes.

Análisis exploratorio de datos de concentración de metales

El análisis exploratorio de los seis metales proporcionados y la prueba de Shapiro-Wilk, para probar normalidad, se realizaron. En todos los casos se rechazó la hipótesis nula con un nivel de significancia de 0.05 (Cuadro 1).

Definición de un modelo general

El modelo de Viton (2010) se utilizó como el general, en él la variable de respuesta autocorrelacionada depende de covariables y de errores que también pueden ser autocorrelacionados:

$$y=X\beta+\lambda Wy+\varepsilon+\upsilon \qquad (7)$$

donde X es la matriz diseño, β es el vector de los coeficientes de regresión, λ es el nivel de autocorrelación de la variable respuesta, W es la matriz de pesos espaciales, ε es error autocorrelacionado (definido en el siguiente paso) y υ es ruido blanco.

Selección de una muestra aleatoria de puntos de una malla

Considerando las coordenadas mínimas y máximas de longitud y latitud de la zona de estudio, se crearon 900 puntos equidistantes y se seleccionó una muestra aleatoria de tamaño $n=100$.

Definición de tratamientos

Las variables respuesta se generaron bajo diferentes tratamientos de acuerdo con el modelo general. Primero, las distribuciones

Exploratory analysis of the metal concentration data

An exploratory analysis of the six metals provided and a Shapiro-Wilk test (to test normality) were carried out. The null hypothesis was rejected in every case, with a 0.05 significance level (Table 1).

Defining a general model

Viton's model (2010) was used as a general model. The autocorrelated response variable depends on covariables and errors that can also be autocorrelated as follows:

$$y=X\beta+\lambda Wy+\varepsilon+\upsilon \qquad (7)$$

where X is the design matrix; β is the vector of the regression coefficients; λ is the autocorrelation level of the response variable; W is the spatial weighing matrix; ε is the autocorrelated error (defined in the following step); and υ is white noise.

Selection of a random sample of points in a mesh

Taking into consideration the minimum and maximum longitude and latitude coordinates of the study zone, 900 equidistant points were established and a random sample with a $n=100$ size was chosen.

Defining treatments

Based on the general model, response variables were generated for the various treatments. First, the Gamma, Inverse Gaussian, and Normal (as a reference) distributions were taken into consideration, in order to represent the behavior of Cr in the Mezquital Valley. A value of four was added to the actual data, in order to guarantee positive results for the simulation. In all cases, the parameters were adjusted to make sure that the first and second order moments were the same in every distribution, as follows:

Cuadro 1. Estadísticas descriptivas de la concentración de los metales.
Table 1. Descriptive statistics for the concentration of metals.

Metal	Mínimo	Mediana	Media	Máximo	Prueba Shapiro-Wilks (p-value)
Cadmio	0.000	0.230	0.340	1.440	1.129E-07
Cobre	0.100	2.650	5.159	34.300	2.158E-09
Cromo	0.000	0.000	0.622	10.650	4.570E-16
Níquel	0.000	0.525	1.413	7.080	9.425E-09
Plomo	0.100	1.760	3.080	12.760	1.263E-07
Zinc	0.300	3.500	10.730	80.700	2.411E-11

Gamma, Inversa Gaussiana y Normal (como referencia) se consideraron para representar el comportamiento de Cr en el Valle del Mezquital. A los datos reales se les adicionó un valor de cuatro para asegurar datos positivos en la simulación. En todos los casos se ajustaron los parámetros para que los momentos de primer y segundo orden fuesen los mismos de una distribución a otra, así:

Se consideró como $y \sim N(4.622, 3.582)$, donde $E[y]=\mu=\bar{y}=4.622$ y $Var(y)=\sigma^2=S^2=3.582$

Se consideró como $y \sim G(5.963, 1.290)$, donde $E[y]=\frac{a}{b}=\bar{y}=4.622$ y $Var(y)=\frac{a}{b^2}=S^2=3.582$ para $y \sim G(a, b)$.

Se consideró como $y \sim IG(4.622, 27.561)$, donde $E[y]=\mu=\bar{y}=4.622$ y $Var(y)=\frac{\mu^3}{\gamma}=S^2=3.582$ para $y \sim IG(\mu, \gamma)$.

Dos tipos de error autocorrelacionado se consideraron (Anselin, 2005), el espacial autorregresivo aleatorio (SAR, del inglés Spatial Autoregressive Random), que se expresa como: $\varepsilon=(I-\rho W)^{-1}u$ y espacial de promedios móviles (SMA, del inglés Spatial Moving Average), que se expresa como: $\varepsilon=\rho Wu+u$, donde W es la matriz de pesos espaciales, ρ es el nivel de autocorrelación correspondiente a los residuales y u es el error autocorrelacionado. Para la matriz de vecindades, implementamos los cinco tipos de matrices de pesos: *W*, *B*, *C*, *U* y *S*. Los niveles de autocorrelación para λ y ρ, fueron los mismos: {0, 0.2, 0.5, 0.7, 0.9}. El total de combinaciones de "tratamientos" fue de 150 (5ρ*$5W$*3 distribuciones*2 ε)para probar la efectividad de estimación y precisión de los modelos GLM, GLMM, GAM y GAMM.

Determinación de los valores de los coeficientes de regresión

Los valores de los coeficientes de la regresión $y=\beta_{00}+\beta_{01}x_1+\beta_{02}x_2+\beta_{03}x_3+\beta_{04}x_4+\lambda Wy+\varepsilon+\upsilon$ fueron $\beta_{00}=1$, $\beta_{01}=5$, $\beta_{02}=10$, $\beta_{03}=1$ y $\beta_{04}=1$. Las covariables que se tomaron en cuenta fueron cuatro, con las primeras dos se intentó reproducir la presencia de cualquier tipo de covariable por lo que se definió: $x_1 \sim N(0, 4)$ y $x_2 \sim N(0, 9)$. Las dos covariables restantes (x_3, x_4) correspondieron a las coordenadas de longitud y latitud del punto seleccionado en la malla, ε se simuló como se definió en la selección de la muestra aleatoria de puntos de una malla y υ como $N(0, 1)$, para cada punto de la malla. De acuerdo con el modelo general (7), un dato en cada punto de la malla se simuló en cada una de las 1000 repeticiones.

Ajuste de modelos lineales y no lineales

Para cada combinación de la selección de la muestra aleatoria de puntos de una malla, los siguientes modelos lineales y no lineales se ajustaron en cada repetición.

Considered as $y \sim N(4.622, 3.582)$, where $E[y]=\mu=\bar{y}=4.622$ and $Var(y)=\sigma^2=S^2=3.582$

Considered as $y \sim G(5.963, 1.290)$, where $E[y]=\frac{a}{b}=\bar{y}=4.622$ and $Var(y)=\frac{a}{b^2}=S^2=3.582$ for $y \sim G(a, b)$.

Considered as $y \sim IG(4.622, 27.561)$, where $E[y]=\mu=\bar{y}=4.622$ and $Var(y)=\frac{\mu^3}{\gamma}=S^2=3.582$ for $y \sim IG(\mu, \gamma)$.

Two types of autocorrelated errors were taken into consideration (Anselin, 2005): SAR and SMA. SAR (Spatial Autoregressive Random) is expressed as follows: $\varepsilon=(I-\rho W)^{-1}u$. Meanwhile, SMA (Spatial Moving Average) is expressed as follows: $\varepsilon=\rho Wu+u$. In both cases, W is the spatial weight matrix, ρ is the autocorrelation level that matches the residuals, and u is the autocorrelated error. In the case of the adjacency matrix, the *W*, *B*, *C*, *U*, and *S* weighing matrices were applied. The autocorrelation levels for λ and ρ were the same: {0, 0.2, 0.5, 0.7, 0.9}. In order to test the effectiveness of the estimate and accuracy of the GLMs, GLMMs, GAMs, and GAMMs, 150 "treatment" combinations (5ρ*$5W$*3 distributions*2 ε) were used.

Determining the values of the regression coefficients

The regression coefficients had $y=\beta_{00}+\beta_{01}x_1+\beta_{02}x_2+\beta_{03}x_3+\beta_{04}x_4+\lambda Wy+\varepsilon+\upsilon$ and $\beta_{00}=1$, $\beta_{01}=5$, $\beta_{02}=10$, $\beta_{03}=1$ y $\beta_{04}=1$ values. Four covariables were considered. The first two were used to attempt to replicate the presence of any type of covariable, defining the following expression: $x_1 \sim N(0, 4)$ and $x_2 \sim N(0, 9)$. The other two covariables (x_3, x_4) matched the longitude and latitude coordinates of the point selected in the mesh; ε was simulated as defined in the selection of the random sample of points in a mesh and υ, as $N(0,1)$ for each point in the mesh. Based on the general model (7), a datum per point in the mesh was simulated in each of the 1000 repetitions.

Adjusting linear and nonlinear models

The following linear and nonlinear models were adjusted in each repetition for each combination of the selection from the random sample of points in the mesh.

Generalized linear model: $\eta=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4$,

Generalized linear mixed model:
$\eta=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+Z\gamma$,

Generalized additive model: $\eta=\beta_0+\beta_1x_1+\beta_2x_2+f_3x_3+f_4x_4$, and

Modelo lineal generalizado: $\eta=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4$,

Modelo lineal generalizado mixto:
$\eta=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+Z\gamma$,

Modelo aditivo generalizado: $\eta=\beta_0+\beta_1x_1+\beta_2x_2+f_3x_3+f_4x_4$, y

Modelo aditivo generalizado mixto:
$\eta=\beta_0+\beta_1x_1+\beta_2x_2+f_3x_3+f_4x_4+Z\gamma$

donde $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ son los coeficientes de regresión, x_1, x_2, x_3, x_4 son las covariables, f_3 y f_4 son las funciones de suavizamiento, Z es la matriz diseño de los efectos aleatorios y γ es el vector que contiene los efectos aleatorios.

Estimación de parámetros

Para cada modelo ajustado y en cada repetición los parámetros β_0, β_1, β_2 se estimaron mediante simulación *bootstrap*. Los parámetros β_3, β_4 correspondieron a las coordenadas de longitud y latitud del punto seleccionado en la malla. Estos parámetros fueron lineales para GLM y GLMM y no lineales para GAM y GAMM.

Comparación de los modelos

La comparación de los modelos se hizo mediante el cálculo del sesgo, varianza y error cuadrado medio (ECM).

RESULTADOS Y DISCUSIÓN

Estudio de simulación

Efecto del tipo de matriz y nivel de autocorrelación

La matriz de vecindad U presentó gran robustez en presencia de los diferentes grados de asociación espacial. Esto fue similar para los cuatro modelos ajustados y para los coeficientes de regresión β_1 y β_2. Con esa matriz se obtuvieron las varianzas y sesgos menores. La distribución de β_1 generada por el modelo GLM mostró la robustez de la matriz U al nivel de autocorrelación (Figura 1) con error de tipo SMA). La estimación de los coeficientes con la matriz tipo B fue la más afectada por la variabilidad del nivel de autocorrelación espacial.

En general los valores del ECM tendieron a aumentar poco con el nivel de autocorrelación (Cuadro 2).

Generalized additive mixed model:
$\eta=\beta_0+\beta_1x_1+\beta_2x_2+f_3x_3+f_4x_4+Z\gamma$

where $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ are the regression coefficients; x_1, x_2, x_3, x_4 are the covariables; f_3 and f_4 are the smoothing splines; Z is the design matrix for the random effects; and γ is the vector that contains the random effects.

Estimating parameters

For each adjusted model and in each repetition, the parameters were estimated using the bootstrap simulation. The parameters matched the longitude and latitude coordinates of the point selected in the mesh. These parameters were linear for GLM and GLMM, and nonlinear for GAM and GAMM.

Comparing the models

The models were compared calculating the bias, variance, and mean squared error (MSE).

RESULTS AND DISCUSSION

Simulation studies

Effect of the matrix type and autocorrelation level

The U-adjacency matrix showed great strength in the presence of several degrees of spatial association. The four adjusted models and the β_1 and β_2 regression coefficients had a similar result. This matrix achieved the lowest variances and biases. The β_1 distribution generated by GLM showed that the U matrix was strong at the autocorrelation level (Figure 1) with a SMA-type error. The variability of the spatial autocorrelation level had a greater effect on the estimation of coefficients with a type B matrix than in others.

Overall, the SME values showed little tendency to increase with the autocorrelation level (Table 2).

Effects of the error type

The results of the bias and the SME generated by the type-S matrix —with a 0.2 and 0.5 correlation, but with different error types— showed certain patterns. The estimator bias for the three distributions was the same for the same autocorrelation levels, in

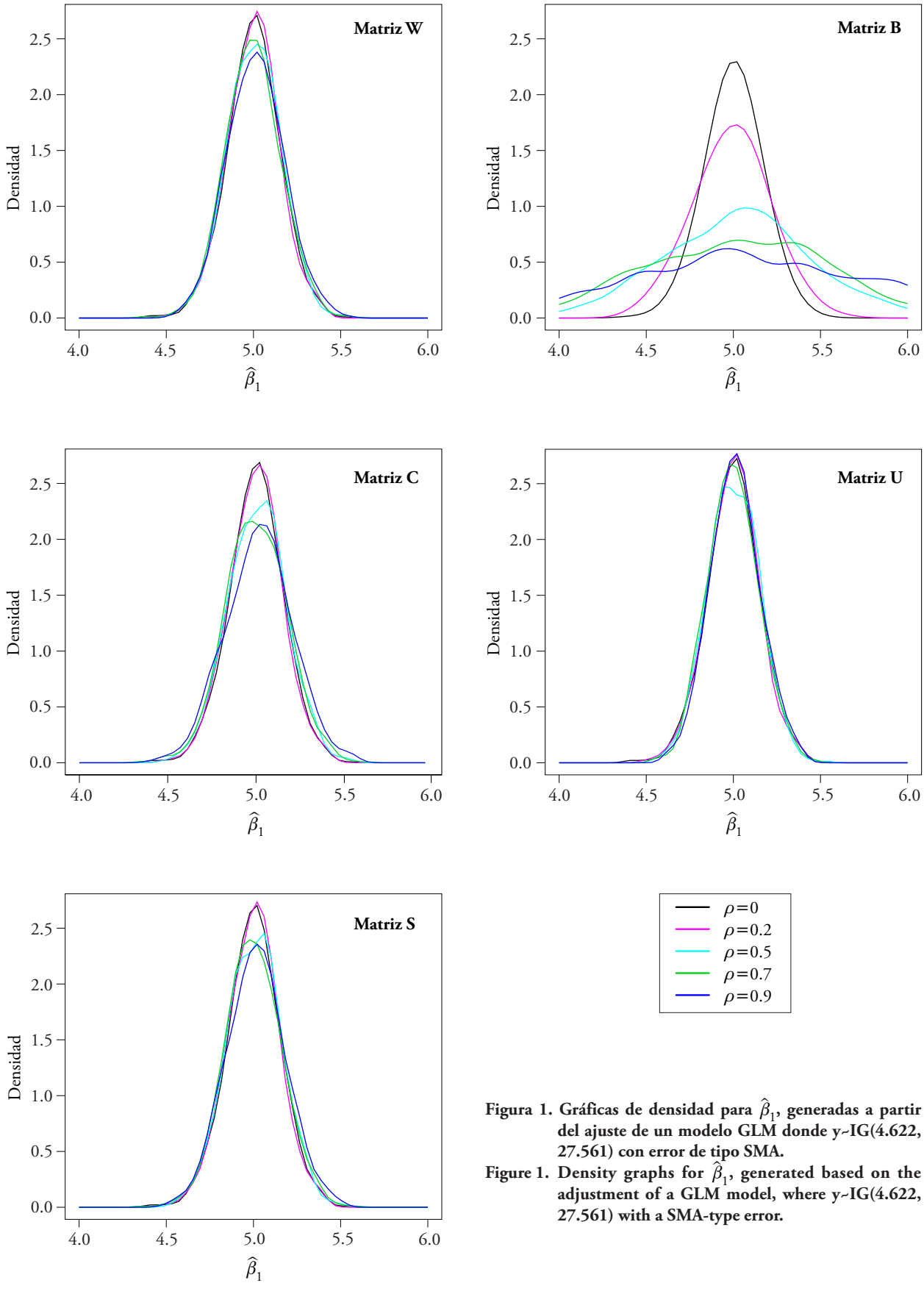

Figura 1. Gráficas de densidad para $\hat{\beta}_1$, generadas a partir del ajuste de un modelo GLM donde y~IG(4.622, 27.561) con error de tipo SMA.

Figure 1. Density graphs for $\hat{\beta}_1$, generated based on the adjustment of a GLM model, where y~IG(4.622, 27.561) with a SMA-type error.

Cuadro 2. Efecto de ρ en la estimación de β_1 y β_2 con distribución Inversa Gaussiana.
Table 2. Effect of ρ in the estimate of β_1 and β_2 with an inverse Gaussian distribution.

Dist.†	Error	W	λ, ρ	GLMM				GAMM			
				Sesgo		ECM		Sesgo		ECM	
				$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
Inv. Gau.¶	SAR	W	0.00	0.001	0.002	0.023	0.022	−0.001	0.001	0.023	0.023
Inv. Gau.¶	SAR	W	0.20	−0.001	0.001	0.022	0.022	0.001	0.003	0.022	0.022
Inv. Gau.¶	SAR	W	0.50	0.004	−0.001	0.024	0.026	0.000	0.002	0.026	0.025
Inv. Gau.¶	SAR	W	0.70	−0.002	0.005	0.028	0.031	0.006	0.007	0.027	0.030
Inv. Gau.¶	SAR	W	0.90	0.001	−0.011	0.038	0.042	−0.001	0.002	0.039	0.039

† Distribución de *y*. ¶ Inversa Gaussiana. ❖ † Distribution of *y*. ¶ Inverse Gaussian.

Efecto del tipo de error

Los resultados de sesgo y ECM generados de la matriz tipo S, con correlación de 0.2 y 0.5, pero con diferente tipo de error, mostraron ciertos patrones. El sesgo de los estimadores de las tres distribuciones fue el mismo para los mismos niveles de autocorrelación, en los modelos GLM, GAM y GLMM. Los valores menores en ECM se presentaron con el error de tipo SMA, cuando el nivel de autocorrelación fue 0.50; con 0.20, no se identificaron cambios en los estimadores (Cuadro 3).

Efecto de los modelos

La estimación de los parámetros con modelos no mixtos y mixtos no mostró diferencias significativas

Effect of the error type

Sesgo and ECM values generated from the S-type matrix, with correlation of 0.2 and 0.5, but with different error type, showed certain patterns. The sesgo of the estimators of the three distributions was the same for the same autocorrelation levels in the GLM, GAM and GLMM models. The lower ECM values were presented with the SMA-type error, when the autocorrelation level was 0.50; with 0.20, no changes were identified in the estimators (Table 3).

Effect of the models

Estimating parameters with non-mixed and mixed models did not show significative differences in variance and SME in relation to linear models, when the response variable was the result of the implementation of the type-U matrix and 0.5 and 0.9 autocorrelation levels. Some changes between GAMs and GAMMs (Table 4) indicated that GAMMs had lower variance and SME values in the three distributions, with a SAR-type error and both autocorrelation levels for both estimators.

Cuadro 3. Efecto del tipo de error en la estimación de β_1 y β_2 con distribución Gamma.
Table 3. Effect of the error type in the estimate of β_1 and β_2 with a Gamma distribution.

Dist.†	Error	W	λ, ρ	GLMM				GAMM			
				Sesgo		ECM		Sesgo		ECM	
				$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
Gamma	SAR	S	0.20	−0.006	−0.005	0.022	0.023	−0.004	0.002	0.022	0.022
Gamma	SAR	S	0.50	−0.003	0.002	0.027	0.025	0.005	−0.005	0.026	0.023
Gamma	SMA	S	0.20	−0.006	−0.005	0.022	0.023	−0.006	0.001	0.021	0.023
Gamma	SMA	S	0.50	−0.003	0.002	0.025	0.024	0.009	0.000	0.025	0.023

† Distribución de *y*. ❖ † Distribution of *y*.

en varianza y ECM en los modelos lineales, cuando la variable respuesta provenía de la implementación de la matriz tipo U y niveles de autocorrelación de 0.5 y 0.9. Algunos cambios entre los modelos GAM y GAMM (Cuadro 4) indicaron que los modelos GAMM presentaron valores menores de varianza y ECM para las tres distribuciones, con un tipo de error SAR y con los dos niveles de autocorrelación en ambos estimadores.

Aplicación

La Comisión Nacional del Agua proporcionó para el estudio 72 observaciones de Cd, Cu, Cr, Ni, Pb y Zn (mg kg^{-1}), de un muestreo en los DR 003, 100 y 112, en septiembre y octubre de 2013. La prueba de Shapiro-Wilk (Cuadro1) indicó que la distribución de los metales fue asimétrica. El efecto fue de latitud y altitud y no de longitud. Los valores mayores se presentaron a mayor altitud, debido a que el agua es distribuida por gravedad, la cual llega primero a las zonas altas.

Ajuste de modelos

Seis de nueve índices de Moran de los residuales con valores más pequeños a modelos mixtos con índices de Moran no significativos (Cuadro 5); esto confirmó que pueden describir la correlación espacial. Tres de esos seis valores corresponden a modelos GAMM y los otros tres a modelos GAM. Esto demostró que en la aplicación los modelos no lineales fueron mejores

Implementation

In order to carry out this study, the Comisión Nacional del Agua provided 72 observations of Cd, Cu, Cr, Ni, Pb, and Zn (mg kg^{-1}), out a sample taken from DRs 003, 100, and 112, during September-October 2013. According to the Shapiro-Wilk test (Table 1), the metals were asymmetrical distributed. Latitude and altitude had an effect, but longitude did not. The highest values appeared at a higher altitude, because water —which arrives first to the high lands— is distributed by gravity.

Adjusting the models

Six out of nine Moran Index of the lower values residuals correspond to mixed models with non-significant Moran Index (Table 5). This confirms that they can describe spatial correlation. Three out of this six values match GAMMs and the remaining three match GAMs. Therefore, as far as their implementation is concerned, nonlinear models had a better performance than linear models. Although GAMs were not designed for correlated data, smoothing splines can describe the said behavior (Figure 2).

Prediction

The Sistema Generador de Modelos Altimétricos (SIGMA) software (Pedraza, 2000) was used to develop a mesh of points based on the quadrant

Cuadro 4. Efecto del modelo GAM y GAMM en la estimación de β_1 y β_2 con matriz U y error SAR.
Table 4. Effect of GAMs and GAMMs in the estimating of β_1 and β_2 with a U-matrix and a SAR error.

Dist.†	λ, ρ	GAM				GAMM			
		Varianza		ECM		Varianza		ECM	
		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
Normal	0.50	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022
Normal	0.90	0.023	0.022	0.023	0.022	0.022	0.021	0.022	0.021
Gamma	0.50	0.022	0.023	0.022	0.023	0.021	0.022	0.021	0.022
Gamma	0.90	0.022	0.022	0.022	0.022	0.021	0.021	0.021	0.021
Inv. Gau.¶	0.50	0.022	0.023	0.022	0.023	0.021	0.022	0.021	0.022
Inv. Gau.¶	0.90	0.022	0.025	0.022	0.025	0.021	0.024	0.021	0.024

† Distribución de *y*. ¶ Inversa Gaussiana. ❖ † Distribution of *y*. ¶ Inverse Gaussian.

Cuadro 5. Comparación de modelos ajustados para concentración de metales.
Table 5. Comparison of adjusted models for metal concentration.

Metal	Distribución	Modelo	AIC	ECM	Ind. Mor.†	p-value
Cadmio	Normal	LM	40.141	0.091	0.342	1.026E-07
	Gamma	GLM	30.859	0.007	0.340	1.279E-07
		GLMM	NA	0.068	0.251	7.162E-05
		GAM	−23.582	0.003	**0.024**	5.635E-01
		GAMM	6.030	0.037	0.121	4.098E-02
	Inv. Gau.¶	GLM	26.623	0.002	0.339	1.443E-07
		GLMM	NA	0.068	0.253	6.269E-05
		GAM	**−28.951**	0.001	**0.026**	5.465E-01
		GAMM	2.310	0.040	0.156	1.035E-02
Cobre	Normal	LM	447.927	26.370	0.100	7.419E-02
	Gamma	GLM	379.569	0.229	0.142	1.994E-02
		GLMM	NA	27.196	0.116	4.043E-02
		GAM	365.376	0.165	0.057	2.857E-01
		GAMM	403.547	26.136	0.095	8.486E-02
	Inv. Gau.¶	GLM	360.911	0.028	0.175	4.931E-03
		GLMM	NA	27.909	0.127	2.562E-02
		GAM	**339.061**	0.018	0.049	3.425E-01
		GAMM	378.953	26.372	0.099	7.546E-02
Cromo	Normal	LM	297.171	3.249	0.150	4.692E-03
	Gamma	GLM	229.177	0.108	0.146	9.099E-03
		GLMM	NA	3.252	0.155	3.551E-03
		GAM	229.259	0.100	0.120	2.874E-02
		GAMM	285.942	3.252	0.155	3.551E-03
	Inv. Gau.¶	GLM	203.588	0.023	0.150	8.564E-03
		GLMM	NA	3.255	0.157	3.051E-03
		GAM	**201.572**	0.020	0.106	5.566E-02
		GAMM	278.751	3.255	0.157	3.051E-03
Níquel	Normal	LM	275.425	2.402	0.231	2.455E-04
	Gamma	GLM	249.460	0.095	0.254	6.959E-05
		GLMM	NA	1.659	**0.025**	5.529E-01
		GAM	233.424	0.057	0.098	9.502E-02
		GAMM	236.018	1.659	**0.025**	5.529E-01
	Inv. Gau.¶	GLM	239.419	0.020	0.263	4.081E-05
		GLMM	NA	1.666	0.032	4.914E-01
		GAM	**223.939**	0.012	0.103	8.172E-02
		GAMM	244.692	1.666	0.026	5.477E-01
Plomo	Normal	LM	359.832	7.758	0.320	5.736E-07
	Gamma	GLM	328.987	0.163	0.367	1.692E-08
		GLMM	NA	4.243	**0.008**	7.400E-01
		GAM	277.782	0.058	0.035	4.607E-01
		GAMM	293.255	4.243	**0.008**	7.400E-01
	Inv. Gau.¶	GLM	317.864	0.026	0.382	5.166E-09
		GLMM	NA	4.281	**0.010**	7.133E-01
		GAM	**268.903**	0.009	0.038	4.387E-01
		GAMM	314.233	3.311	**−0.103**	1.739E-01
Zinc	Normal	LM	594.723	202.562	0.233	9.682E-05
	Gamma	GLM	495.104	0.613	0.319	5.909E-07
		GLMM	NA	153.538	0.139	1.397E-02
		GAM	**444.220**	0.217	**−0.035**	7.540E-01
		GAMM	502.633	149.639	0.113	3.884E-02

† Índice de Moran de los residuales. ¶ Inversa Gaussiana. ❖ † Moran Index of the residuals. ¶ Inverse Gaussian.

que los lineales. Aunque los modelos GAM no están diseñados para datos correlacionados las funciones de suavizamiento pueden describir dicho comportamiento (Figura 2).

Predicción

Una malla de puntos del cuadrante de la zona de estudio se elaboró con el software Sistema Generador de Modelos Altimétricos (SIGMA) (Pedraza, 2000), después con ArcMap se hizo un recorte de esos puntos, con la capa vectorial de parcelas de los DR, para obtener los puntos que se encontraban dentro de los DR. Así se obtuvieron 112 728 puntos. Esta malla se consideró para elaborar mapas de predicción. Las predicciones se vieron alteradas por el tipo de distribución estadística que se consideró. Para Cd la combinación de la distribución Inversa Gaussiana y el modelo GAM describió mejor el comportamiento, de acuerdo con el valor del AIC (Figura 3).

Análisis de riesgo

Los límites máximos permisibles para metales pesados en suelo de uso agrícola los establece la NOM-001-SEMARNAT-1996. Se obtuvieron los porcentajes de puntos predichos que sobrepasan el límite permisible, de acuerdo a la predicción por modelos GAM, al considerar distribución inversa Gaussiana para Cd, Cu, Cr, Ni y Pb y distribución Gamma para Zn (Cuadro 6); así como, la probabilidad de que el valor predicho sobrepase el límite permisible para cada uno de los metales dadas las distribuciones mencionadas (Figura 4).

CONCLUSIONES

La estimación de los parámetros β_1 y β_2 es más robusta al nivel de autocorrelación cuando se utiliza la matriz de pesos de tipo U. La integración de efectos aleatorios y funciones de suavizamiento describe mejor el comportamiento de datos autocorrelacionados. Condiciones nuevas, como la distribución no normal, funciones de suavizamiento y efectos aleatorios describen con confiabilidad estadística mayor el comportamiento de la distribución espacial de los metales pesados. La zona más contaminada corresponde al DR 003. La contaminación por Cd es la mayor y por Pb la menor.

of the study zone. Subsequently, ArcMap was used to cut these points with the vector layer of the DR plots, in order to find out which were the points included in the DRs. As a result, 112 728 points were found. This mesh was used to develop prediction maps. The type of statistical distribution that was taken into consideration altered the predictions. The best description of the behavior of Cd —according to the AIC value— was provided by the combination of the inverse Gaussian distribution and the GAM (Figure 3).

Risk analysis

The NOM-001-SEMARNAT-1996 standard establishes the maximum permissible limits for heavy metals in agricultural soils. The predicted percentages of points exceeding the permissible limit were obtained, according to the GAM-based prediction which considered an inverse Gaussian distribution of Cd, Cu, Cr, Ni, and Pb, and a gamma distribution of Zn (Table 6). Given the said distributions, the likelihood that the value predicted exceeds the permissible limit for each metal was also obtained (Figure 4).

CONCLUSIONS

In terms of autocorrelation, the estimate of the β_1 and β_2 is stronger when the type-U weighing matrix

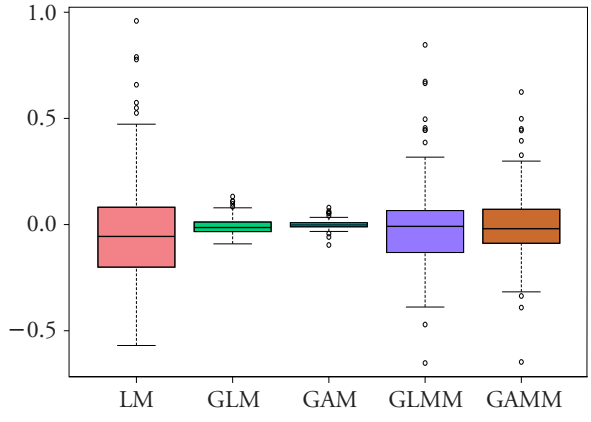

Figura 2. Gráfica boxplot de residuales del ajuste de modelos para [Cd] con distribución Normal en LM e Inversa Gaussiana para el resto de los modelos.

Figure 2. Boxplot of the residuals of the model adjustment for [Cd] with normal distribution in LM and inverse Gaussian in the remaining models.

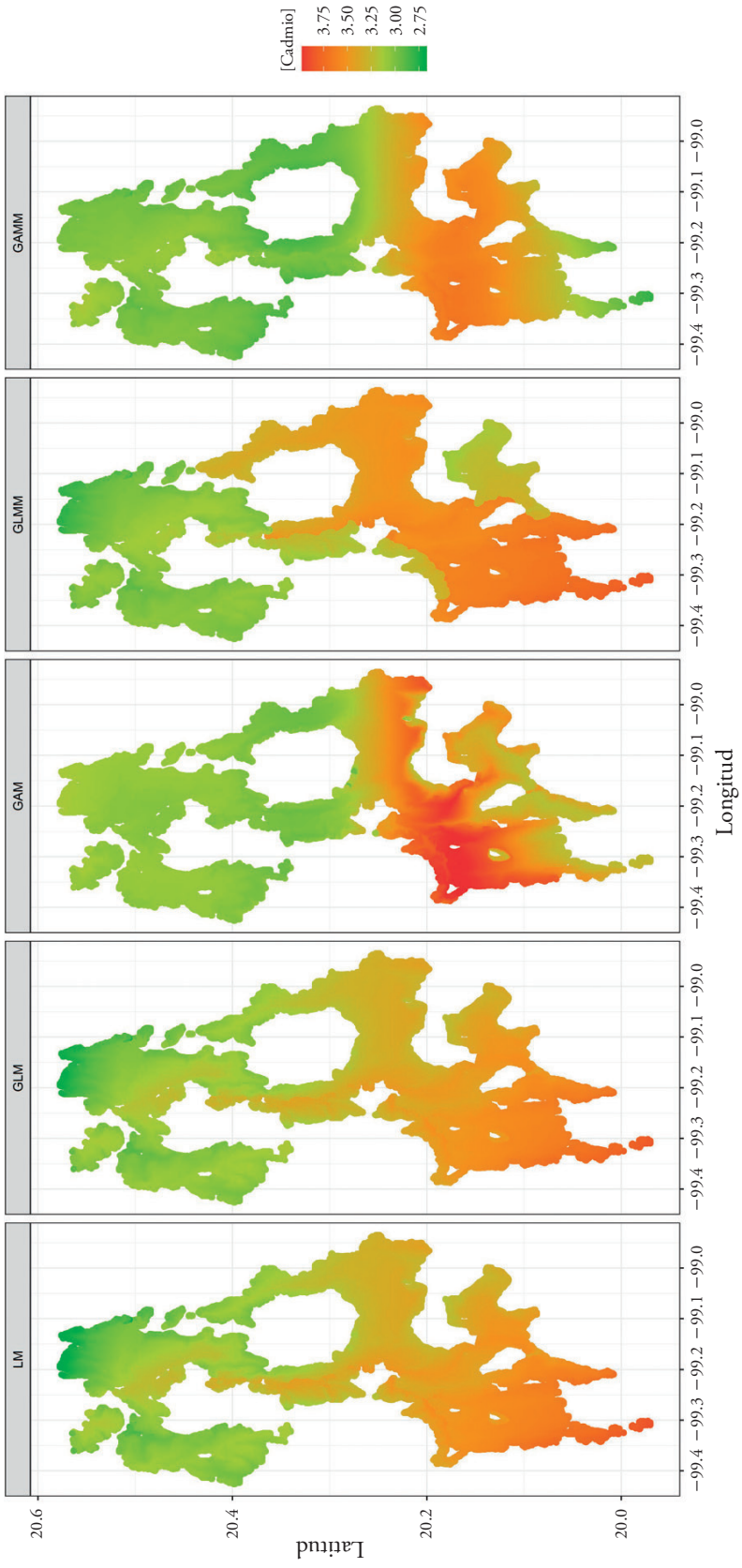

Figura 3. Mapas de predicción para [Cd] con distribución Normal para LM e Inversa Gaussiana para el resto de los modelos.
Figure 3. Prediction mapping for [Cd] with normal distribution for LM and inverse Gaussian for the remaining models.

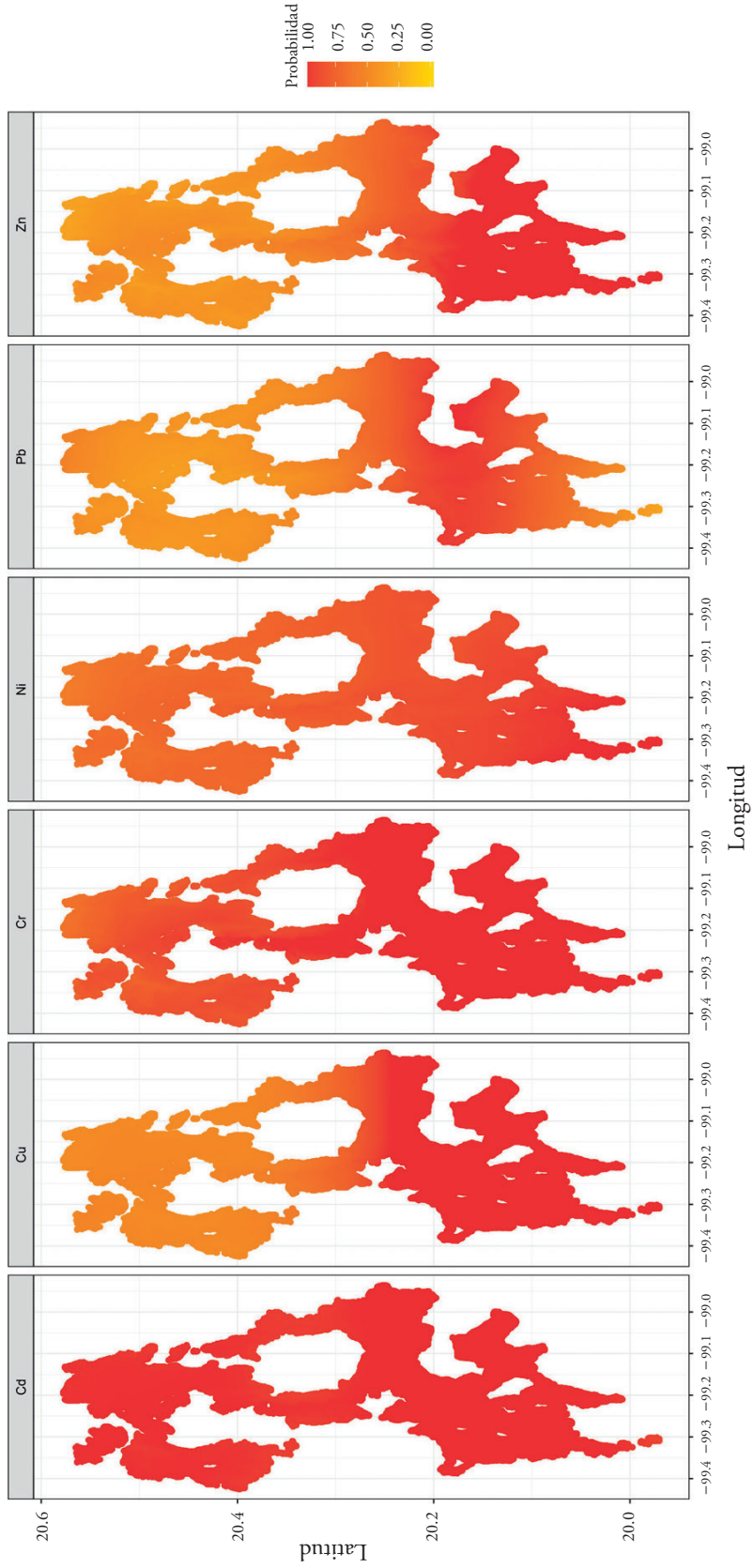

Figura 4. Probabilidad de que el valor predicho sobrepase el límite permisible.
Figure 4. Likelihood that the predicted value exceeds the permissible level.

Cuadro 6. Porcentaje de puntos que exceden los límites permisibles.
Table 6. Percentage of points that exceed permissible levels.

	Metal					
	Cd	Cu	Cr	Ni	Pb	Zn
Límite (mg/kg) †	3.05	7.00	3.50	5.00	8.00	13.00
Porcentaje	84.22	44.58	33.68	30.04	28.59	42.42

† Límite más tres. ❖ † Limit plus three.

AGRADECIMIENTOS

A la Comisión Nacional del Agua (CONAGUA), por los datos proporcionados para la aplicación de la presente investigación.

LITERATURA CITADA

Anselin, L. 2005. Spatial Regression Analysis in R: A Workbook. University of Illinois. Center for Spatially Integrated Social Science. Department of Agricultural and Consumer Economics. 141 p.

Breslow, N. E., and D. G. Clayton. 1993. Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. 88: 9-25.

Hastie, T., and R. Tibshirani. 1986. Generalized additive models (with discussion). Stat. Sci.e 1: 297-318.

Lin, X., and D. Zhang. 1999. Inference in generalized additive mixed models using smoothing splines. J. R. Stat. Soc. 61: 381-400.

Liu, H. 2008. Generalized Additive Model. University of Minnesota Duluth. Department of Mathematics and Statistics.

Mamouridis, V. 2011. Additive Mixed Models applied to the study of red shrimp landings: comparison between frequentist and Bayesian perspectives. Universidad de Coruña. Departamento de Matemáticas. España. 94 p.

McCulloch, C. E. 1997. Maximum likelihood algorithms for generalized linear mixed models. J. Am. Stat. Assoc. 92: 162-170.

Moran, P. A. P. 1948. The Interpretation of Statistical Maps. J. R. Stat. Soc. 10: 243-251.

is used. The behavior of autocorrelated data is best described by the integration of random effects and smoothing splines. New conditions —such as non-normal distribution, smoothing splines, and random effects— describe with greater statistical reliability the behavior of the spatial distribution of the heavy metals. The most polluted zone is DR 003. Cd and Pb are the greatest and lowest pollutants, respectively.

—End of the English version—

Nelder, J. A., and R. W. M. Wedderburn. 1972. Generalized linear models. J. R. Stat. Soc. 135: 370-384.

Pedraza O., F. 2000. SIGMA: Sistema Generador de Modelos Altimétricos. Colegio de Postgraduados. Campus Montecillo. México.

Tiefelsdorf, M., D. A. Griffith, and B. Boots. 1999. A variance-stabilizing coding scheme for spatial link matrices. Environ. Plan. A 31: 165-180.

Torabi, M. 2015. Likelihood Inference for Spatial Generalized Linear Mixed Models. Communications in Statistics. Simul. Comput. 44: 1692-1701.

Viton, P. A. 2010. Notes on Spatian Econometric Models. The Ohio State University.

Wood, S.N. 2006. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC Press.