



**COLEGIO DE POSTGRADUADOS**

**INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN EN CIENCIAS  
AGRÍCOLAS**

**CAMPUS MONTECILLO**

**POSTGRADO DE SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA  
ESTADÍSTICA**

**Modelos geoestadísticos aplicados a la distribución  
de contaminantes en suelo**

NANCY TORIZ ROBLES

TESIS

PRESENTADA COMO REQUISITO PARCIAL PARA  
OBTENER EL GRADO DE:

MAESTRA EN CIENCIAS

MONTECILLO, TEXCOCO, EDO. DE MÉXICO

2016

La presente tesis titulada: **MODELOS GEOESTADÍSTICOS APLICADOS A LA DISTRIBUCIÓN DE CONTAMINANTES EN SUELO**, realizada por la alumna: Nancy Toriz Robles bajo la dirección del Consejo Particular indicado, ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

MAESTRA EN CIENCIAS

SOCIOECONOMÍA, ESTADÍSTICA E INFORMATICA-ESTADÍSTICA

CONSEJO PARTICULAR

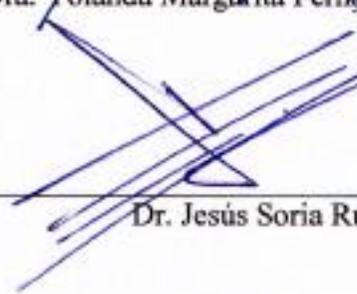
CONSEJERA

  
Dra. Martha Elva Ramírez Guzmán

DIRECTORA

  
Dra. Yolanda Margarita Fernández Ordoñez

ASESOR

  
Dr. Jesús Soria Ruiz

ASESORA

  
Dra. Ma. Carmen Ybarra Moncada

Montecillo, Texcoco de Mora, Estado de México, Julio de 2016.

# MODELOS GEOESTADÍSTICOS APLICADOS A LA DISTRIBUCIÓN DE CONTAMINANTES EN SUELO

Nancy Toriz Robles

Colegio de Postgraduados, 2016

## RESUMEN

El estudio de datos espaciales presenta: dependencia entre las muestras observadas, distribución de datos no normales y heterocedasticidad. Por ello, es importante la implementación de metodologías que permitan predecir niveles de contaminación en suelo con mayor precisión. El presente trabajo analiza y compara el comportamiento de los estimadores de diferentes modelos de regresión lineales (GLM y GLMM) y no lineales (GAM y GAMM), a través de la simulación de una variable de respuesta generada por diferentes tipos de matrices de pesos, niveles de autocorrelación, errores autocorrelacionados y distribuciones con el objetivo de identificar que modelos que son capaces de describir el comportamiento de datos espaciales. Adicionalmente se aplicaron los modelos a un conjunto de 72 datos del Valle del Mezquital, Hgo. para elaborar mapas de predicción de concentración de metales pesados en función de los modelos ajustados. Además se propuso un tamaño de muestra óptimo en función del nivel de autocorrelación. Cuando la variable respuesta es generada a través de la matriz de pesos de tipo U, la estimación de los parámetros es más robusta al nivel de autocorrelación debido a que se acerca más al valor inicial de los parámetros. Los mejores resultados se obtuvieron con los modelos mixtos y no lineales, debido a la integración de efectos aleatorios y a las funciones de suavizamiento respectivamente. Las medidas de autocorrelación espacial permiten calcular tamaños de muestra con el fin de optimizar los recursos sin perder información estadísticamente importante.

**Palabras Clave:** Metales pesados, autocorrelación, distribución no normal, modelos lineales, modelos no lineales, índice de Moran, muestra óptima.

# STATISTICAL MODELS APPLIED TO THE DISTRIBUTION OF CONTAMINANTS IN SOIL

Nancy Toriz Robles

Colegio de Postgraduados, 2016

## ABSTRACT

The study of spatial data presents: dependence between observed samples, non-normal distribution of data, and heteroskedasticity. Therefore, It is important to implement methodologies to predict pollution levels in soil more accurately. This paper analyzes and compares the behavior of the estimators of different linear (GLM and GLMM) and nonlinear (GAM and GAMM) regression models, through simulation of a response variable generated by different types of weight matrices, autocorrelation levels, autocorrelated errors, and distributions in order to identify which models are able to better describe the behavior of geospatial data. Additionally, models were applied to a set of 72 data from the Mezquital Valley, Hidalgo to map prediction of concentration of heavy metals as function of the adjusted models. In addition, an optimal sample size was proposed depending on the level of autocorrelation. When the response variable is generated by the weight matrix type U, the estimation of parameters is more robust at the autocorrelation level because it is closer to the initial value of the parameters. The best results were obtained with the mixed and nonlinear models due to the integration of random effects and smoothing functions respectively. Spatial autocorrelation measures allow calculating sample sizes in order to optimize resources without losing statistically important information.

**Keywords:** Heavy metals, autocorrelation, non-normal distribution, linear models, nonlinear models, Moran Index, optimal sample.

## AGRADECIMIENTOS

*Quisiera que las próximas líneas sirvan como reconocimiento a las personas e instituciones que han hecho posible la realización de esta tesis.*

*Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por el apoyo brindado para la realización de mis estudios de Postgrado.*

*Al Colegio de Postgraduados, en especial al Postgrado de Estadística por ofrecerme la oportunidad de continuar con mi preparación profesional y brindarme los conocimientos necesarios para la culminación de este trabajo de investigación.*

*A la Dra. Martha Elva Ramírez Guzmán por su guía, tiempo, dedicación y sus valiosas aportaciones, sin las que este trabajo no hubiera sido posible.*

*A la Dra. Dra. Yolanda Margarita Fernández Ordoñez, al Dr. Jesús Soria Ruiz y la Dra. Ma. Carmen Ybarra Moncada por la revisión y cada una de las aportaciones que más que enriquecer el trabajo lo han mejorado.*

*Al personal de los Distritos de Riego 003 Tula, 100 Alfajayucan y 112 Ajacuba de la Comisión Nacional del Agua (CONAGUA), por los datos proporcionados para la realización de la presente investigación.*

## DEDICATORIA

*Es mi deseo como sencillo gesto de agradecimiento, dedicar mi trabajo a las siguientes personas.*

*A mis padres, porque no hay palabras que puedan describir mi profundo agradecimiento a quienes admiro, por su comprensión y ayuda en todo momento. Por ser la fuente de mi inspiración, fortaleza y motivación para superarme cada día, Dios los bendiga.*

*A mis hermanas Miriam y Clara por cada momento compartido, su cariño infinito y apoyo incondicional.*

*A Marco y Samara por ser mi motor y fuente de alegría.*

*A las familias Toriz Monroy y Robles Pérez por que en cada uno de ustedes encontré el apoyo para seguir adelante, gracias por hacerme pasar momentos inolvidables. A todos y cada uno de mis primos y sobrinos, en especial a Crysthian, Nadia, Yuly, Emmanuel, Edileni, Hanam, Misael, Anargelia, Mayra, Brenda, David, Omar, Ferdinando, Oscar, Osvaldo y Alexa por los buenos momentos disfrutados a su lado.*

*A María, Alberto, Olga, Diana, Arely, Sócrates, Rodrigo y Joselyn por su amistad, comprenderme, darme ánimo en los malos momentos, acompañándome y apoyándome para ser cada día una mejor persona. A Jesús, Moisés, José, Benito, Alejandro, Yuseel, Magaña, Antonio, Carlos, Ely, Carmen, Isaí, Zacarías y Adrián porque nuestra amistad perdura a pesar de la distancia.*

*A mis compañeros de maestría, por los gratos recuerdos que han dejado en mí durante mi estancia en el colegio y deseándoles el éxito que cada uno se merece.*

*“La dicha de la vida consiste en tener siempre algo que hacer, alguien a quien amar y alguna cosa que esperar”. Thomas Chalmers*

# CONTENIDO

RESUMEN .....	iii
ABSTRACT.....	iv
LISTA DE CUADROS.....	ix
LISTA DE FIGURAS .....	x
I. INTRODUCCIÓN .....	1
II. MARCO TEÓRICO.....	2
2.1 Modelos Lineales.....	2
2.1.1 Modelo Lineal .....	2
2.1.2 Kriging .....	4
2.1.3 Modelos Simultáneos Autorregresivos.....	5
2.1.4 Modelo Condicional Autorregresivo.....	6
2.1.5 Modelos Lineales Generalizados.....	6
2.1.6 Modelos Lineales Generalizados Mixtos.....	10
2.2 Modelos No Lineales .....	13
2.2.1 Modelos Aditivos Generalizados.....	13
2.2.2 Modelos Aditivos Generalizados Mixtos.....	15
2.3 Matriz de vecindad y matriz de pesos.....	16
2.4 Índice de Moran.....	18
III. JUSTIFICACIÓN.....	19
IV. OBJETIVOS.....	20
V. MATERIALES Y MÉTODOS .....	21
5.1 Estudio de simulación .....	21
5.2 Aplicación .....	26
5.2.1 Ubicación del área de estudio .....	26
5.2.2 Muestras de suelo y obtención de datos .....	27
VI. RESULTADOS .....	28
6.1 Estudio de simulación .....	28
6.1.1 Efecto del tipo de matriz .....	28
6.1.2 Efecto del nivel de autocorrelación.....	30
6.1.3 Efecto del tipo de error .....	31

6.1.4	Efecto de los modelos .....	33
6.2	Aplicación .....	35
6.2.1	Análisis exploratorio .....	35
6.2.2	Ajuste de modelos .....	38
6.2.3	Predicción y análisis de riesgo .....	41
6.2.4	Índice de Moran y tamaño de muestra óptimo .....	44
VII.	CONCLUSIONES Y RECOMENDACIONES.....	47
VIII.	BIBLIOGRAFÍA CONSULTADA.....	48
IX.	ANEXOS .....	54
Anexo 1.	Tratamientos para generar las variables de respuesta.....	54
Anexo 2.	Mapas de predicción de metales pesados .....	56
Anexo 3.	Efectos de altas concentraciones de metales pesados .....	60
Anexo 4.	Métodos de tratamiento de aguas residuales y recuperación de suelos .....	62

## LISTA DE CUADROS

Cuadro 1. Características de algunas distribuciones de la familia exponencial.....	8
Cuadro 2. Tratamientos para la generación de variable respuesta. ....	23
Cuadro 3. Efecto del tipo de matriz en la estimación de $\beta_1$ y $\beta_2$ con modelos GLM y GAM.....	29
Cuadro 4. Efecto del tipo de matriz en la estimación de $\beta_1$ y $\beta_2$ con modelos GLMM y GAMM. .....	30
Cuadro 5. Efecto del nivel de autocorrelación en la estimación de $\beta_1$ y $\beta_2$ con modelos GLM y GAM.....	31
Cuadro 6. Efecto del nivel de autocorrelación en la estimación de $\beta_1$ y $\beta_2$ con modelos GLMM y GAMM. ....	31
Cuadro 7. Efecto del tipo de error en la estimación de $\beta_1$ y $\beta_2$ con modelos GLM y GAM. ....	32
Cuadro 8. Efecto del tipo de error en la estimación de $\beta_1$ y $\beta_2$ con modelos GLMM y GAMM. .....	32
Cuadro 9. Efecto del modelo GAM y GAMM en la estimación de $\beta_1$ y $\beta_2$ con matriz de tipo U. ....	34
Cuadro 10. Efecto de los modelos en la estimación de $\beta_1$ y $\beta_2$ con matriz de tipo C, error tipo SMA y nivel de autocorrelación de 0.5.....	34
Cuadro 11. Estadísticas descriptivas de la concentración de los metales. ....	35
Cuadro 12. Comparación de modelos ajustados para concentración de metales. ....	39
Cuadro 13. Límites máximos permisibles de metales pesados. ....	41
Cuadro 14. Porcentaje de puntos que exceden los límites permisibles. ....	43
Cuadro 15. Índices de Moran empleando diferente matriz de pesos. ....	44
Cuadro 16. Tamaño de muestra empleando diferente matriz de pesos. ....	45
Cuadro 17. Combinaciones para generar las variables de respuesta. ....	54
Cuadro 18. Efectos de algunos metales tóxicos en plantas y animales.....	61
Cuadro 19. Efectos de algunos metales tóxicos en el ser humano.....	61
Cuadro 20. Principales técnicas de recuperación de suelos. ....	65

## LISTA DE FIGURAS

Figura 1. Diagrama de bloques del proceso de simulación.....	21
Figura 2. Localización de los DR 003 Tula, DR 100 Alfajayucan y DR 112 Ajacuba. ....	26
Figura 3. Gráficas de densidad para $\beta_1$ , generadas a partir del ajuste de un modelo GLM donde $y \sim IG(4.622, 27.561)$ con error de tipo SMA. ....	28
Figura 4. Gráficas de densidad para $\beta_2$ , generadas a partir del ajuste de un modelo GLM donde $y \sim IG(4.622, 27.561)$ con error de tipo SMA. ....	29
Figura 5. Histogramas de la concentración de los metales. ....	36
Figura 6. Longitud vs. concentración de metales.....	36
Figura 7. Latitud vs. concentración de metales. ....	37
Figura 8. Altitud vs. concentración de metales. ....	37
Figura 9. Gráfica boxplot para residuales del ajuste de modelos para [Cd] con distribución Normal en LM e Inversa Gaussiana para el resto de los modelos.....	40
Figura 10. Mapas de predicción para [Cd] con distribución Normal para LM e Inversa Gaussiana para el resto de los modelos. ....	42
Figura 11. Mapas de predicción para [Ni] con distribución Normal para LM y Gamma para el resto de los modelos.....	42
Figura 12. Probabilidad de que el valor predicho sobrepase el límite permisible .....	43
Figura 13. Muestreo a) aleatorio y b) sistemático para el DR 003, DR 100 y DR. ....	46
Figura 14. Mapas de predicción para [Cd] con distribución Normal para LM y Gamma para el resto de los modelos.....	56
Figura 15. Mapas de predicción para [Cr] con distribución Normal para LM y Gamma para el resto de los modelos.....	56
Figura 16. Mapas de predicción para [Cr] con distribución Normal para LM e Inversa Gaussiana para el resto de los modelos. ....	57
Figura 17. Mapas de predicción para [Cu] con distribución Normal para LM y Gamma para el resto de los modelos.....	57
Figura 18. Mapas de predicción para [Cu] con distribución Normal para LM e Inversa Gaussiana para el resto de los modelos. ....	58
Figura 19. Mapas de predicción para [Ni] con distribución Normal para LM e Inversa Gaussiana para el resto de los modelos. ....	58

Figura 20. Mapas de predicción para [Pb] con distribución Normal para LM y Gamma para el resto de los modelos.....	59
Figura 21. Mapas de predicción para [Pb] con distribución Normal para LM e Inversa Gaussiana para el resto de los modelos. ....	59
Figura 22. Mapas de predicción para [Zn] con distribución Normal para LM y Gamma para el resto de los modelos.....	60

## I. INTRODUCCIÓN

Debido al crecimiento demográfico, a la urbanización, deforestación y el cambio climático global, se ha presentado una reducción considerable de aguas dulces. Esto ha dado lugar al uso creciente de aguas residuales para la agricultura, la acuicultura, la recarga de aguas subterráneas, etc. En algunas regiones cercanas a las grandes ciudades, las aguas residuales son el único recurso hídrico de comunidades rurales que se dedican a la agricultura (Ayres y Mara, 1997).

Las aguas residuales sin tratamiento se aplican en terrenos agrícolas, por lo que resulta necesario conocer la dispersión y nivel de contaminación de sustancias nocivas en suelo. Entre los contaminantes incorporados se encuentran los metales pesados, los cuales tienden a acumularse en el suelo y su remoción es prácticamente imposible. Por ello, es importante la implementación de metodologías que permitan predecir los niveles de contaminación del suelo con la mayor precisión posible. Sin embargo, el comportamiento de datos espaciales irrumpe los supuestos gaussianos clásicos de los modelos de regresión, es decir; presentan dependencia entre las muestras, distribución de datos no normales, y heterocedasticidad en la varianza de los errores. Por lo tanto, el presente trabajo analiza y compara diferentes modelos de regresión, que pueden ser aplicados a datos de concentración de contaminantes en suelo.

## II. MARCO TEÓRICO

A lo largo del tiempo se han desarrollado diferentes modelos de regresión que han sido aplicados a datos espaciales. El surgimiento de cada uno de ellos ha considerado una nueva característica del comportamiento de dichos datos, tales como dependencia de las observaciones, valores de la variable respuesta mayor o igual a cero, no normalidad y heterocedasticidad. Por otra parte, es de interés conocer el efecto lineal o no lineal de un conjunto de covariables. A continuación se presentan los Modelos Lineales, Kriging, Modelos Condicionales Autorregresivos, Modelos Autorregresivos Simultáneos, Modelos Lineales Generalizados, Modelos Lineales Generalizados Mixtos y modelos no lineales: como los Modelos Aditivos Generalizados y los Aditivos Generalizados Mixtos. La diferencia entre los modelos lineales y no lineales, recae en que en los primeros, los parámetros multiplican directamente a las covariables en un modelo aditivo; en cambio en los segundos, los parámetros se pueden expresar por ejemplo de manera exponencial o logarítmica, para representar la relación no lineal entre la variable de respuesta y la(s) covariable(s).

### 2.1 Modelos Lineales

#### 2.1.1 Modelo Lineal

El Modelo Lineal (LM, *Linear Model*) analiza la relación lineal entre una variable respuesta  $y \sim N(X\beta, \sigma^2 I)$  y una o más covariables. Este modelo se expresa de la siguiente forma:

$$E[y] = X\beta \quad (1)$$

Donde  $X$  es la matriz diseño correspondiente a las covariables que definen los componentes fijos del modelo y  $\beta$  el vector de los coeficientes de regresión.

Los LM tienen numerosas aplicaciones, solo por citar algunos ejemplos: Cortés *et al.* (2013), quienes determinaron la relación espacial entre la conductividad eléctrica y el contenido de sodio en un suelo Andisol. Por otro lado, Chicas *et al.* (2014) estimaron la capacidad de campo y el punto de marchitez permanente a partir del porcentaje de arcilla, limo y arena, densidad aparente, conductividad eléctrica y contenido de materia orgánica para conocer la disponibilidad de agua en el suelo. Aunque estos modelos son muy fáciles de aplicar, una desventaja que presentan, es que únicamente son aplicables a datos normales de  $y$ , situación que por lo general nunca se comprueba.

### ***Estimación por máxima verosimilitud***

Bajo la expresión 1, se asume que  $y \sim N(X\beta, \sigma^2 I)$  y dado un vector  $y_0 = (y_1, y_2, \dots, y_N)$  de observaciones de la variable respuesta, la estimación de máxima verosimilitud de  $\beta$  es posible si las  $y_i$  son i.i.d. (Radhakrishna y Toutenburg, 1999; McCullagh y Searle, 2001), de manera que:

$$L(\beta) = (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right\} \quad (2)$$

Al aplicar el logaritmo

$$l_\beta = \log L(\beta) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta) \quad (3)$$

y derivar la expresión anterior con respecto a  $\beta$  e igual a cero:

$$\frac{\partial l_\beta}{\partial \beta} = \frac{X'(y - X\beta)}{\sigma^2} = 0 \quad (4)$$

se obtiene:

$$\hat{\beta} = (X'X)^{-1}X'y \quad (5)$$

Denotando  $\beta$  como  $\hat{\beta}$  y resolviendo la expresión anterior se obtiene un estimador calculado por máxima verosimilitud el cual es insesgado y de varianza mínima. Se dice que  $\hat{\beta}$  es de varianza mínima de  $\beta$  si para todas las estimaciones  $\beta^*$  tales que  $E[\beta^*] = \beta$ , se tiene  $Var(\hat{\beta}) \leq Var(\beta^*)$ .

Independencia lineal entre las covariables, implica que cada covariable contiene información sobre la variable respuesta que no está contenida en otras, es decir; cada columna de la matriz  $X$  es linealmente independiente del resto (Meyer, 1992).

### 2.1.2 Kriging

Una de las técnicas ampliamente utilizada en estudios espaciales es el Kriging, desarrollado por Matheron en la década de los 60, a diferencia de los LM, el Kriging considera la correlación entre las muestras. Este modelo se fundamenta en la minimización del error cuadrático medio de predicción, es decir se minimiza la expresión:

$$V(Z^*(x_o) - Z(x_o)) \quad (6)$$

Donde  $Z^*(x_o)$  es el mejor predictor lineal que se expresa como:

$$Z^*(x_o) = \lambda_1 Z(x_1) + \lambda_2 Z(x_2) + \dots + \lambda_n Z(x_n) = \sum_{i=1}^n \lambda_i Z(x_i) \quad (7)$$

Donde  $Z$  se refiere a las mediciones de la variable de interés en los puntos  $x_i$  y  $Z(x_o)$  corresponde al punto que se desea predecir.

Una característica de los métodos Kriging es que garantizan varianza mínima de predicción (Samper y Carrera, 1990); no así los métodos de interpolación determinísticos (e.g. método de la media aritmética, polígonos de Thissen, método de isoclinas, distancia inversa y triangulación de Delaunay). Algunos ejemplos de aplicación del Kriging los presenta Sun *et al.* (2003), donde se enfocaron en el estudio de la variabilidad espacial y temporal de las propiedades del suelo y los cambios de la calidad del suelo en una región montañosa de China. Su análisis demostró que los niveles de pH, materia orgánica disponible, P y K sufrieron cambios significativos de 1985 a 1997. Ibarra *et al.* (2009) probaron dos métodos de interpolación espacial: Kriging ordinario y Distancia Inversa Ponderada para la estimación de la distribución espacial del pH del suelo a partir de 105 muestras de suelo en la región agrícola del municipio de Zapopan, Jalisco. Concluyeron que el método de Kriging ordinario fue mejor, ya que el error cuadrado medio era menor. Marchetti *et al.* (2012) establecieron semivariogramas experimentales para determinar la dependencia espacial de la distribución de la textura del suelo y contenido de materia orgánica. Los resultados de 250 datos de puntos de muestreo de suelos fueron interpolados mediante Kriging ordinario a través de ArcGIS 9.2<sup>®</sup> para producir mapas. Un trabajo similar fue desarrollado por González *et al.* (2015) quienes delimitaron zonas aptas para establecer el cultivo de *Jatropha curcas* en el estado de Tabasco.

Respecto al estudio de metales pesados, Arauz *et al.* (2013) utilizó el método Kriging para determinar el nivel de contaminación y distribución espacial.

Los métodos Kriging asumen que la variable de respuesta tiene distribución normal, sin embargo la distribución de datos de suelo tales como: concentración de nutrientes y metales, se caracteriza por ser asimétrica (Figura 5). El Kriging se basa en el semivariograma experimental (gráfica que muestra la dependencia espacial entre dos muestras en función de la distancia entre ellas), sin embargo; estos no siempre coinciden con los teóricos (lineal, esférico, exponencial o Gaussiano), debido a que el primero no siempre presentan meseta, la cual caracteriza a los teóricos. Esta diferencia puede implicar una mala interpolación de predicción. Por otra parte, información de covariables, como el efecto de la dispersión de una plaga sobre la producción agrícola, no puede ser incorporada en el semivariograma. Por lo anterior, este método no es un buen candidato para estudiar datos espaciales asimétricos.

### 2.1.3 Modelos Simultáneos Autorregresivos

Una alternativa para incluir información sobre el comportamiento espacial de datos es a través de la inclusión de una matriz de pesos (Página 16) y del nivel de autocorrelación. La matriz de pesos es la representación de una matriz de vecindad, la cual a diferencia del semivariograma puede ser elaborada no solo en función de la distancia, sino de alguna característica en específico (e.g. color de pixel en imágenes satelitales, cantidad de radiación solar, rangos de producción de algún cultivo, presencia de enfermedades, etc.). El nivel de autocorrelación puede ser incluido a través de un conocimiento teórico, práctico o considerando algunos índices conocidos (e.g. Índice de Moran, Índice de Geary). La matriz de pesos y el nivel de autocorrelación fueron incluidos en el Modelo Simultáneo Autorregresivo (SAR, *Simultaneous Autoregressive Models*) y el Modelo Condicional Autorregresivo (CAR, *Conditional Autoregressive Models*).

De acuerdo a la notación propuesta por Anselin (1988); en los modelos SAR, tanto la matriz de pesos como el nivel de autocorrelación afectan a residuales y covariables:

$$y = \rho W y + X\beta + \varepsilon = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1} \varepsilon \quad (8)$$

En donde  $\varepsilon \sim N(0, V_C)$ , si  $\sigma_i^2 = \sigma^2$  para todas las  $i$  observaciones, la matriz de covarianza es  $V_C = \sigma^2(I - \rho W)^{-1}$ , donde  $W$  es la matriz de pesos espaciales y además es simétrica,  $\rho$  es el parámetro

de autocorrelación,  $I$  es la matriz identidad,  $X$  es la matriz diseño correspondientes a las covariables fijas y  $\beta$  es el vector de los coeficientes de regresión asociado a dichas covariables. Estos modelos están diseñados para variables de respuesta con distribución normal. Keitt *et al.* (2002) aplicaron los SAR en el estudio de la distribución de topillo rojo (roedor).

### 2.1.4 Modelo Condicional Autorregresivo

El modelo CAR fue propuesto por Keitt *et al.* (2002). En este modelo la autocorrelación (matriz de pesos y nivel de autocorrelación) únicamente se presenta en los residuales:

$$Y = X\beta + \rho W(Y - X\beta) + \varepsilon = X\beta + (I - \rho W)^{-1}\varepsilon \quad (9)$$

Donde  $\rho$ ,  $W$ ,  $I$  y  $\beta$  son como se definieron en el modelo 8. Los modelos CAR no son adecuados para procesos con dirección, esto es; cuando los datos están influenciados por condiciones naturales como vientos, mareas, corrientes, etc. Las aplicaciones de los modelos CAR se han producido en demografía, economía, epidemiología y geografía. De Oliveira (2012) aplica los modelos CAR bayesianos a un conjunto de datos de fosfato en una zona arqueológica de Grecia.

### 2.1.5 Modelos Lineales Generalizados

En comparación con los LM y el Kriging, los Modelos Lineales Generalizados (GLM, *Generalized Linear Models*) (Nelder y Wedderburn, 1972) permiten modelar variables de respuesta con distribuciones que pertenezcan a la familia exponencial. De acuerdo con Wood (2006), una distribución pertenece a la familia exponencial si su función de densidad se puede definir como:

$$f_{\theta}(y) = \exp \left[ \frac{y(\theta) - b(\theta)}{a(\phi)} + c(y, \phi) \right] \quad (10)$$

Donde  $a$ ,  $b$  y  $c$  son funciones arbitrarias,  $\theta$  es un parámetro natural (canónico) de la distribución y  $\phi$  es un parámetro de escala. Cuando el parámetro de escala no está presente en la distribución bajo estudio, se asume que  $\phi \equiv 1$ . Las funciones de  $a$  y  $c$  son tales que  $a(\phi) = \phi/w$  y  $c(y, \phi/w)$ , donde  $w$  es un peso conocido para cada observación, y  $b$  Es una función sólo de  $\theta$ .

Sea  $E[y] = \mu$ , y  $\mu$  es la media de la distribución de  $y$ . La relación entre la variable y las covariables se establece mediante la relación:

$$g(\mu) = \eta \quad (11)$$

Donde  $\eta = X\beta$  llamado predictor lineal,  $g$  es una función de respuesta conocida (función de vínculo o función liga) y  $X$  y  $\beta$  definidos en la expresión 1.

Por lo tanto:

$$E[y] = \mu = g^{-1}(X\beta) \quad (12)$$

Es posible obtener una expresión general para la media y la varianza de las distribuciones de la familia exponencial, en términos de  $a$ ,  $b$  y  $\phi$  (Wood, 2006).

$$E(Y) = b'(\theta) \quad (13)$$

$$Var(Y) = \frac{b''(\theta)\phi}{w} \quad (14)$$

Donde  $b'$  y  $b''$  denotan la primera y segunda derivada respectivamente con respecto a  $\theta$ .

La familia exponencial incluye tanto distribuciones continuas como discretas. El Cuadro 1 presenta las distribuciones continuas que se trabajaron en la presente investigación. La desventaja de estos modelos es que no son capaces de captar la dependencia espacial, ya que están diseñados para datos independientes. A pesar de lo anterior, han sido implementados en estudios geoestadísticos, como por ejemplo Carl *et al.* (2008) implementaron los GLM para modelar el comportamiento de la distribución de especies en una investigación ecológica. Oviedo *et al.* (2015) estimaron los casos de influenza en Cataluña (España), las covariables analizadas fueron: población, edad, datos del informe de gripe, y región sanitaria durante 2010-2014.

**Cuadro 1. Características de algunas distribuciones de la familia exponencial.**

	Normal	Gamma	Inversa Gaussiana
$f(y)$	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)$	$\frac{1}{\Gamma(v)y} \left(\frac{yv}{\mu}\right)^v \exp\left(-\frac{vy}{\mu}\right)$	$\sqrt{\frac{\gamma}{2\pi y^3}} \exp\left[\frac{-\gamma(y-\mu)^2}{2\mu^2 y}\right]$
Rango	$-\infty < y < \infty$	$y > 0$	$y > 0$
$\theta$	$\mu$	$-\frac{1}{\mu}$	$-\frac{1}{2\mu^2}$
$\phi$	$\sigma^2$	$\frac{1}{v}$	$\frac{1}{\gamma}$
$a(\phi)$	$\phi (= \sigma^2)$	$\phi \left(= \frac{1}{v}\right)$	$\phi \left(= \frac{1}{\gamma}\right)$
$b(\theta)$	$\frac{\sigma^2}{2}$	$-\log(-\theta)$	$-\sqrt{-2\theta}$
$c(y, \phi)$	$-\frac{1}{2} \left[ \frac{y^2}{\phi} + \log(2\pi\phi) \right]$	$v \log(vy) - \log(y\Gamma(v))$	$-\frac{1}{2} \left[ \log(2\pi y^3 \phi) + \frac{1}{\phi y} \right]$
$g$	$\mu$	$\frac{1}{\mu}$	$\frac{1}{\mu^2}$
$Var(y)$	$\phi$	$\phi\mu^2$	$\phi\mu^3$
$l(\mu_i, \phi; y_i, w_i)$	$-\frac{1}{2} \left[ \frac{w_i(y_i - \mu_i)^2}{\phi} + \log\left(\frac{\phi}{w_i}\right) + \log(2\pi) \right]$	$w_i \phi \log\left(\frac{w_i y_i \phi}{\mu_i}\right) - \frac{w_i y_i \phi}{\mu_i} - \log y_i - \log[\Gamma(w_i \phi)]$	$-\frac{1}{2} \left[ \frac{w_i(y_i - \mu_i)^2}{\phi} + \log\left(\frac{\phi y_i^3}{w_i}\right) + \log(2\pi) \right]$

Fuente: Wood (2006), Zwillinger y Kokoska (2000).

### Estimación por máxima verosimilitud

Tanto en los modelos LM como en los GLM, la estimación de los parámetros se puede obtener a través del método de máxima verosimilitud. Dado un vector  $y_0 = (y_1, y_2, \dots, y_N)$  de observaciones i.i.d. de la variable respuesta, la función de máxima verosimilitud para  $\beta$  así como su logaritmo es (Wood, 2006):

$$L(\beta) = \prod_{i=1}^N f_{\theta_i}(y_i) \quad (15)$$

$$l_{\beta} = \sum_{i=1}^N \log[f_{\theta_i}(y_i)] = \sum_{i=1}^N \frac{[y_i \theta_i - b_i(\theta_i)]}{a_i(\phi)} + c_i(\phi, y_i) \quad (16)$$

Donde  $f_\theta$ ,  $a$ ,  $b$  y  $c$  ya fueron definidas en la expresión 7 y si  $a(\phi) = \phi/w$ , entonces:

$$l_\beta = \sum_{i=1}^n w_i \frac{[y_i \theta_i - b_i(\theta_i)]}{a_i(\phi)} + c_i(\phi, y_i) \quad (17)$$

Al derivar la expresión anterior e igualar a cero, no existe una solución exacta (Durbán, 2014). Las estimaciones de máxima verosimilitud de los parámetros  $\beta$  en el predictor lineal  $\eta$  se obtienen por mínimos cuadrados ponderados iterativos (Venables y Ripley, 2002; Rodríguez, 2007; Carl *et al.*, 2008). Dado un  $\hat{\eta}_0$  inicial estimado, las estimaciones iniciales de los coeficientes de regresión y la varianza en función de los pesos puede ser calculada. El cálculo de regresión ponderada utiliza un vector  $z$  de trabajo construido como la respuesta y los pesos iterativos son dados por las siguientes expresiones (Venables y Dichmont, 2004; McCullagh y Nelder, 1989; Rodríguez, 2007):

$$z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i) \left( \frac{d\eta_i}{d\mu_i} \right) \quad (18)$$

$$w_i^{-1} = \left( \frac{d\eta_i}{d\mu_i} \right)^2 V_i \quad (19)$$

Donde la derivada de la función liga es evaluada en la estimación de prueba y  $V_i$  es la función de varianza evaluada en  $\hat{\mu}_i$ .

Finalmente, se obtiene una estimación de  $\beta$  mediante mínimos cuadrados ponderados como:

$$\hat{\beta} = (X'WX)^{-1}X'Wz \quad (20)$$

Donde  $X$  es la matriz diseño,  $W$  matriz diagonal de pesos con entradas  $w_i$  dadas por la expresión 14, y  $z$  es un vector de respuesta con las entradas  $z_i$ .

Durban (2014) menciona el método de Newton-Raphson (Fisher Scoring Algorithm) para la solución de las derivadas al igualar a cero.

### 2.1.6 Modelos Lineales Generalizados Mixtos

Los Modelos Lineales Generalizados Mixtos (GLMM, *Generalized Linear Mixed Models*) se obtienen a partir de los GLM con la incorporación de efectos aleatorios en los predictores lineales, esta idea fue desarrollada por Breslow y Clayton (1993). La estructura de un GLMM está dada por la expresión:

$$g(\mu) = \eta = X\beta + Z\gamma \quad (21)$$

Donde  $\mu$ ,  $g$ ,  $\eta$ ,  $X$  y  $\beta$  fueron definidos en la expresión 7,  $Z$  es la matriz diseño de los efectos aleatorios y  $\gamma \sim N(0, \psi)$  es el vector que contiene los efectos aleatorios. La matriz de covarianzas  $\psi$  es usualmente parametrizada en términos del parámetro  $\theta$ .

La importancia de los GLMM radica en que permiten modelar variables respuesta normales y no normales, continuas y discretas, siempre y cuando la distribución de la variable respuesta pertenezca a la familia exponencial (expresión 7), pueden modelar sobredispersión, heterocedasticidad y correlación mediante la incorporación de efectos aleatorios (McCulloch, 1997; Badiella, 2011; Torabi, 2015).

Las aplicaciones de los GLMM son diversas, por ejemplo Kleinschmidt *et al.* (2001) usaron estos modelos en el análisis espacial de áreas pequeñas con incidencia de malaria en KwaZulu-Natal, Sudáfrica, para identificar los factores que podrían explicar la heterogeneidad de incidencia entre poblados. Wu *et al.* (2006) predijeron un patrón espacial de la probabilidad de la nubosidad por hora en un Bosque Experimental de Luquillo (Puerto Rico), por medio de cuatro modelos diferentes. Eligieron los mejores modelos basados en el Criterio de Información de Akaike (AIC). Los modelos GLM y GLMM fueron los mejores modelos. Borda *et al.* (2012) aplicaron los GLMM para modelar datos de conteo (número de homicidios por barrio, número de habitantes por localidad, enfermos por municipio, entre otros), georreferenciados por municipios en el departamento de Antioquía (Colombia), con el objetivo de encontrar algún tipo de dependencia espacial a partir de la localización geográfica.

#### ***Estimación por máxima verosimilitud***

En un GLM la estimación de los parámetros del modelo a través de máxima verosimilitud resulta sencillo al considerar la independencia entre datos, lo cual no sucede con los GLMM. Algunos de los problemas debido a la correlación son: la dependencia entre la media y la varianza para datos

no normales. Aunque la distribución conjunta es posible matemáticamente, podría no serlo computacionalmente. El logaritmo de verosimilitud marginal puede ser construido mediante la integración de los efectos aleatorios de la distribución conjunta, bajo la condición de efectos aleatorios independientes. Sin embargo, la integración numérica es práctica sólo cuando el número de efectos aleatorios es pequeño. Debido a lo anterior, varios métodos de estimación se han propuesto en la literatura. Los dos enfoques básicos son: la aproximación a la función objetivo y aproximación al modelo. Los algoritmos para el segundo caso, se pueden expresar en términos de series de Taylor, también conocidos como métodos de linealización. Emplean expansiones para aproximarse al modelo, por otro basado en pseudo-datos con menos componentes no lineales. El proceso de cálculo de la aproximación lineal se debe repetir varias veces hasta que no se vea reflejado algún cambio significativo. El GLMM se aproxima mediante un Modelo Lineal Mixto (LMM) basado en los valores actuales de las estimaciones de los parámetros de covarianza. La ventaja de los métodos de aproximación integrales es proporcionar una función objetivo para la optimización. El proceso de estimación es iterativo. La desventaja de los métodos de aproximación integral es la dificultad de adaptar efectos aleatorios cruzados, múltiples efectos y estructuras de covarianza. Por lo que el número de efectos aleatorios debe ser pequeño para que los métodos de aproximación integral sean factibles en la práctica. Las ventajas de los métodos basados en la linealización es que incluyen una forma relativamente sencilla del modelo linealizado que por lo general puede estar basada sólo en la media y la varianza de la forma linealizada. Modelos en los que la distribución conjunta es difícil o imposible de determinar, se pueden ajustar considerando la linealización. Las desventajas es que no incluyen una verdadera función objetivo para el proceso global de optimización e incluyen estimaciones potencialmente sesgadas de los parámetros de covarianza. La función objetivo a ser optimizada después de cada actualización de linealización depende de los pseudo-datos actuales. El proceso puede fallar en ambos niveles del esquema de doble iteración (SAS, 2008).

Una de las técnicas de estimación es conocida como pseudo-verosimilitud restringida, propuesta por (Wolfinger y O'Connell 1993), al aplicar la primera serie de Taylor de  $\mu$  sobre  $\tilde{\beta}$  y  $\gamma$  se obtiene:

$$\mu = g^{-1}(\tilde{\eta}) + \tilde{\Delta}X(\beta - \tilde{\beta}) + \tilde{\Delta}Z(\gamma - \tilde{\gamma}) \quad (22)$$

Donde  $\tilde{\Delta} = \left( \frac{\partial g^{-1}(\eta)}{\partial \eta} \right)_{\tilde{\beta}, \tilde{\gamma}}$  es una matriz diagonal de las derivadas de la media condicional evaluada en el lugar de expansión. Reordenando términos se obtiene:

$$\tilde{\Delta}^{-1}(\mu - g^{-1}(\tilde{\eta})) + X\tilde{\beta} + Z\tilde{\gamma} \doteq X\beta + Z\gamma \quad (23)$$

El lado izquierdo es el valor esperado, condicionado a  $\gamma$ , de

$$\tilde{\Delta}^{-1}(y - g^{-1}(\tilde{\eta})) + X\tilde{\beta} + Z\tilde{\gamma} \equiv P \quad (24)$$

$$Var(P|\gamma) = \tilde{\Delta}^{-1}A^{1/2}RA^{1/2}\tilde{\Delta}^{-1} \quad (25)$$

Donde  $A$  es una matriz diagonal y contiene las funciones de la varianza del modelo. La función de la varianza expresa la varianza de una respuesta como una función de la media. La matriz  $R$  es una matriz de varianza correspondiente a los efectos aleatorios.

Entonces se puede considerar el modelo:

$$P = X\beta + Z\gamma + \varepsilon \quad (26)$$

Que es un modelo lineal mixto con pseudo-respuesta  $P$ , efectos fijos  $\beta$ , efectos aleatorios  $\gamma$  y error  $\varepsilon$ . Así,  $Var(y|\gamma) = A^{1/2}RA^{1/2}$  y  $Var(\varepsilon) = Var(P|\gamma)$ . Si se define a  $V(\theta) = Z\psi Z' + \tilde{\Delta}^{-1}A^{1/2}RA^{1/2}\tilde{\Delta}^{-1}$  como la varianza marginal en el modelo pseudo-lineal mixto, donde  $\theta$  es el vector de parámetros que contiene todas las incógnitas en  $\psi$  y  $R$ . Con base en este modelo linealizado, una función objetivo se puede definir, asumiendo que la distribución de  $P$  es conocida. El máximo del logaritmo de la pseudo-verosimilitud y el logaritmo restringido de la pseudo-verosimilitud:

$$l(\theta, P) = -\frac{1}{2}\log|V(\theta)| - \frac{1}{2}r'V(\theta)^{-1}r - \frac{f}{2}\log(2\pi) \quad (27)$$

$$l_R(\theta, P) = -\frac{1}{2}\log|V(\theta)| - \frac{1}{2}r'V(\theta)^{-1}r - \frac{1}{2}\log|X'V(\theta)^{-1}X| - \frac{f-k}{2}\log(2\pi) \quad (28)$$

Donde  $r = p - X(X'V^{-1}X)^{-1}X'V^{-1}p$ ,  $f$  denota la suma de las frecuencias usadas en el análisis y  $k$  el rango de  $X$ . La función objetivo para la minimización es  $-2l(\theta, p)$  o  $-2l_R(\theta, p)$ . En la convergencia, los parámetros perfilados son estimados y los efectos aleatorios se predicen como:

$$\hat{\beta} = \left( X'V(\hat{\theta})^{-1}X \right)^{-1} X'V(\hat{\theta})^{-1}p \quad (29)$$

$$\hat{\gamma} = \hat{\psi}Z'V(\hat{\theta})^{-1}\hat{r} \quad (30)$$

Existen otros métodos para la estimación de los parámetros antes mencionados en los modelos mixtos, como la cuasi-verosimilitud penalizada (PQL, *Penalized Quasi Likelihood*) y la marginal cuasi-verosimilitud (MQL, *Marginal Quasi Likelihood*) propuestos por Breslow y Clayton (1993). Bolker *et al.* (2009) presentan cuatro métodos de estimación para los GLMM: PQL, aproximación de Laplace, cuadratura de Gauss-Hermite y Cadena de Markov Monte Carlo.

## 2.2 Modelos No Lineales

### 2.2.1 Modelos Aditivos Generalizados

Estos modelos fueron desarrollados por Hastie y Tibshirani en 1986 (GAM, *Generalized Additive Models*). Son modelos lineales generalizados al igual que los GLM, la diferencia radica en que el predictor lineal puede implicar una suma de  $k$  funciones de suavizamiento de las covariables; se considerará un modelo de regresión no paramétrico. Su estructura general es:

$$g(\mu) = \eta = X\beta + f_1(x_1) + f_2(x_2) + \dots + f_k(x_k) \quad (31)$$

Donde  $\mu$ ,  $g$ ,  $\eta$ ,  $X$  y  $\beta$  fueron definidos en la expresión 11 y  $f_k$  son funciones de suavizamiento continuas.

El suavizamiento es una función que estima la tendencia de la variable de respuesta. Una de las técnicas de suavizamiento más empleada son los *splines* (Durbán, 2014), estas consisten en la unión de funciones polinómicas a través de puntos llamados nodos, estos puntos dividen el rango de  $x$  en regiones. Los *splines* dependen de tres elementos: grado del polinomio, número de nodos y localización de los nodos. Por lo anterior existen muchas combinaciones posibles, el más usado es el *spline* cúbico.

Un *spline* cúbico es una curva construida a partir de partes de polinomios de tercer grado que se ensamblan perfectamente de forma que la curva que forman es continua. Su utilidad radica en que son los polinomios de más pequeño grado que tienen segundas derivadas continuas y puntos de

inflexión. Para obtener más información acerca de las técnicas de suavizamiento se pueden consultar los trabajos de Wood (2006), Liu (2008), Mamouridis (2011) o Durbán (2014).

Seoane (2005) utiliza los modelos GAM para predecir la distribución de aves en las islas del Archipiélago Canario. Oviedo *et al.* (2015) implementaron los GAM para estimar la evolución de la transmisión de la influenza Cataluña (España). Brogniez *et al.* (2015) elaboraron mapas del contenido de carbono orgánico en Europa a partir de modelos GAM. Vaca y Golicher (2016) seleccionaron 258 especies de árboles mexicanos asociadas a cinco diferentes ecosistemas, que varían en su grado de exposición a los períodos de sequía y ajustaron modelos GAM para predecir su distribución.

### ***Estimación por penalización***

Una alternativa para controlar la suavidad es mantener fija la base de *splines*. Para controlar el grado de suavización del modelo, se emplea una penalización al momento de ajustar el modelo por minimización (Wood, 2006):

$$\|y - X\beta\|^2 + \lambda \int_0^1 [f''(x)]^2 dx \quad (32)$$

Donde  $\lambda$  es el parámetro de suavización, cuando  $\lambda = 0$  no existe penalización y a medida que  $\lambda$  aumenta se incrementa la intensidad de la penalización y cuando  $\lambda \rightarrow \infty$  se habla de un modelo de regresión lineal. Debido a que  $f$  es lineal en los parámetros  $\beta_i$ , la penalización puede ser escrita en forma cuadrática en  $\beta$ :

$$\int_0^1 [f''(x)]^2 dx = \beta^T S \beta \quad (33)$$

Donde  $S$  es una matriz de coeficientes conocidos que dependen de la base elegida y a la que se le conoce como matriz de penalización. Por lo tanto, el problema de ajuste de regresión *spline* penalizado es minimizar  $\|y - X\beta\|^2 + \lambda \beta^T S \beta$ . Por lo tanto el estimador de mínimos cuadrados penalizados de  $\beta$  dado  $\lambda$  y  $S$  es:

$$\hat{\beta} = (X^T X + \lambda S)^{-1} X^T y \quad (34)$$

### 2.2.2 Modelos Aditivos Generalizados Mixtos

Los Modelos Aditivos Generalizados Mixtos (GAMM, *Generalized Additive Mixed Models*) fueron propuestos por Lin y Zhang (1999). Son una extensión de los GAM mediante la incorporación de efectos aleatorios en los predictores lineales. El modelo general que los representa se muestra a continuación:

$$g(\mu) = \eta = X\beta + f_1(x_1) + f_2(x_2) + \dots + f_k(x_k) + Z\gamma \quad (35)$$

Donde  $\mu = E[y|b]$ ,  $y|b \sim$  alguna distribución de la familia exponencial,  $g$  es la función liga,  $\eta$  es el predictor lineal,  $X$  es la matriz diseño correspondientes a las covariables que definen los componentes paramétricos del modelo,  $\beta$  es el vector de los coeficientes de regresión de las covariables,  $\gamma \sim N(0, \psi)$  es el vector que contiene los efectos aleatorios y  $Z$  es la matriz diseño de los efectos aleatorios. En la actualidad existen infinidad de aplicaciones de estos modelos. Ma *et al.* (2013) evaluaron el efecto de la temperatura sobre la incidencia y pronóstico de la disentería bacilar. Polansky y Robbins (2013) estudiaron patrones temporales de estructuras vegetales que afectan la producción de fruta. Ciannelli *et al.* (2012) compararon varios modelos GAMM para estudiar la distribución de una población de peces.

#### *Estimación por DPQL*

La estimación de funciones no paramétricas y parámetros de la varianza en el contexto de modelos mixtos aditivos generalizados, se logra mediante el uso de Máxima Verosimilitud Restringida (REML, *Restricted Maximum Likelihood*) en el caso de las respuestas normales donde la función liga es la identidad se puede implementar PQL o DPQL (DPQL, *Double Penalized Quasi Likelihood*). Cuando la función de probabilidad exacta es difícil de calcular, la aproximación de Laplace puede ser una opción (Lin y Zhang, 1999; Wood, 2006, Mamouridis, 2011). De acuerdo a Lin y Zhang (1999), al aplicar DPQL, donde  $y = \beta_0 + \sum_{j=1}^k N_j f_j + Z\gamma$  se puede obtener:

$$\hat{\beta}_0 = S_0 \left( y - \sum_{j=1}^k N_j f_j - Z\gamma \right) \quad (36)$$

$$\hat{y} = S_y \left( y - \beta_0 - \sum_{j=1}^k N_j f_j \right) \quad (37)$$

Donde  $S_0 = (\mathbf{1}^T W \mathbf{1})^{-1} \mathbf{1}^T W$ ,  $S_y = (\mathbf{Z}^T W \mathbf{Z} + \psi^{-1})^{-1} \mathbf{Z}^T W$ ,  $\mathbf{1}$  es un vector de tamaño  $n$  y  $W$  definida anteriormente.

### 2.3 Matriz de vecindad y matriz de pesos

Como ya se mencionó, algunos modelos como el SAR y CAR hacen uso de la matriz de pesos y de vecindad para la incorporación de información espacial. La matriz de vecindad es aquella matriz que indica si una región (objeto, dato, etc.) es vecino espacial de otro. La matriz debe ser de tamaño  $n \times n$ , donde  $n$  es el número de observaciones, con elementos  $(t, r)$  que indican si el punto  $t$  es vecino del punto  $r$ .

Existe una amplia gama de maneras de construir la matriz de vecindad, todo depende del contexto que se desee estudiar, por ejemplo:

- $t$  y  $r$  son vecinos si comparten un borde, frontera (polígonos) o vértice.
- Dos regiones como vecinos si, sus centroides se encuentran dentro de una distancia determinada uno del otro.
- Todos son vecinos.
- Cada punto solo tiene un número determinado de vecinos y estos vecinos a su vez se eligen de acuerdo a una propiedad (distancia, característica, etc.).

La matriz de pesos se obtiene al asignar un valor numérico a la matriz de vecindades. De forma convencional se ha considerado que los elementos diagonales de la matriz tienen un valor de cero, es decir cuando el punto  $t$  es el punto  $r$ .

A continuación se presentan las cinco más representativas proporcionadas por la paquetería `spdep` del software R.

### **Matriz tipo W**

También llamada "fila de normalización", en dicha matriz 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:

$$w_{tr} = \frac{1}{\tilde{w}_t} \quad (38)$$

### **Matriz tipo B**

Conocida como "matriz binaria", consiste en dar el valor de la unidad cuando entre  $t$  y  $r$  son vecinos y cero a los elementos de la matriz no vecinos:

$$\begin{aligned} w_{tr} &= 1 \text{ si es vecino} \\ w_{tr} &= 0 \text{ de otro modo} \end{aligned} \quad (39)$$

Por lo tanto la suma de los elementos de la matriz  $w_{tr}$  es el número total de vecindades.

### **Matriz tipo C**

Sea  $n$  es el número de regiones y  $\tilde{w}_{tr}$  el número total de vecindades, entonces los elementos de la matriz de pesos se definen como:

$$w_{tr} = \frac{n}{\tilde{w}_{tr}} \quad (40)$$

### **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:

$$w_{tr} = \frac{1}{\tilde{w}_{tr}} \quad (41)$$

## Matriz tipo S

Es un esquema de codificación propuesto por Tiefelsdorf *et al.* (1999), en el cual 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 da como resultado el número total de regiones. Además se cumple que filas con mayor número de vecindades tienen mayor peso que aquellas con menor número. Es decir:

$$w_{1r} = \frac{a_1}{\tilde{w}_{1r}}, w_{2r} = \frac{a_2}{\tilde{w}_{2r}}, \dots, w_{nr} = \frac{a_n}{\tilde{w}_{nr}} \quad (42)$$

Donde las  $a_t$  son las ponderaciones por fila tal que:  $a_1 + a_2 + \dots + a_n = n$ .

## 2.4 Índice de Moran

Desarrollado por Patrick Moran (1948), indica el grado de asociación espacial entre un conjunto de datos. Sus valores oscilan entre -1 (perfecta dispersión) a 1 (correlación positiva perfecta), donde un valor de cero es indicativo de un modelo espacial aleatorio. Es decir, los valores altos o bajos para una variable aleatoria tienden a agruparse en el espacio (autocorrelación positiva) o tienden a estar rodeado de vecinos con valores muy disímiles (autocorrelación negativa) (Anselin y Bera, 1998). El Índice de Moran se expresa por:

$$I = \frac{n \sum_{t=1}^n \sum_{r=1}^n W_{tr} (y_t - \bar{y})(y_r - \bar{y})}{\sum_{t=1}^n \sum_{r=1}^n W_{tr} \sum_{t=1}^n (y_t - \bar{y})^2} \quad (43)$$

Donde  $n$  es el número de observaciones,  $y_t$  es la variable de respuesta en el punto  $t$ ,  $y_r$  es la variable de respuesta en el punto  $r$ ,  $\bar{y}$  es la media de  $y$  y  $W_{tr}$  es el valor de la matriz de pesos asociado al punto  $t$  con respecto al punto  $r$ .

### III. JUSTIFICACIÓN

Cada uno de los modelos presentados tiene ventajas y desventajas. Los modelos GLMM y GAMM son extensiones de los GLM y GAM, respectivamente, los cuales gozan de mayor flexibilidad para modelar datos correlacionados, heterocedásticos y con componentes aleatorios. Los GLMM son modelos paramétricos mientras que los GAMM pueden ser paramétricos, semiparamétricos o no paramétricos. Aunque los GLM y GAM están diseñados para modelar datos independientes y homocedásticos, Carl *et al.* (2008) y Oviedo *et al.* (2015) aplicaron estos modelos a datos espaciales de manera exitosa. Por lo anterior en la presente investigación se realizó un análisis comparativo de los modelos GLM, GLMM, GAM y GAMM, dado que la concentración de metales pesados se caracteriza por ser positiva, correlacionada, distribución asimétrica, sesgada por la derecha y heterocedástica. Se decidió trabajar con simulación estadística para representar diferentes comportamientos de concentraciones de metales en suelo, mediante la expresión general  $y = X\beta + \lambda Wy + \varepsilon + v$ , donde  $\varepsilon = (I - \rho W)^{-1}u$  cuando el error es de tipo SAR y  $\varepsilon = \rho Wu + u$  cuando es de tipo SMA, donde  $y, X, \beta, \lambda, W$  y  $v$  fueron definidas en la sección (5.1). Por lo tanto los objetivos de esta investigación se presentan a continuación.

## IV. OBJETIVOS

### Objetivo general

Analizar y comparar los modelos de regresión GLM, GLMM, GAM y GAMM ajustados a datos autocorrelacionados espacialmente a través de simulación estadística.

### Objetivos particulares

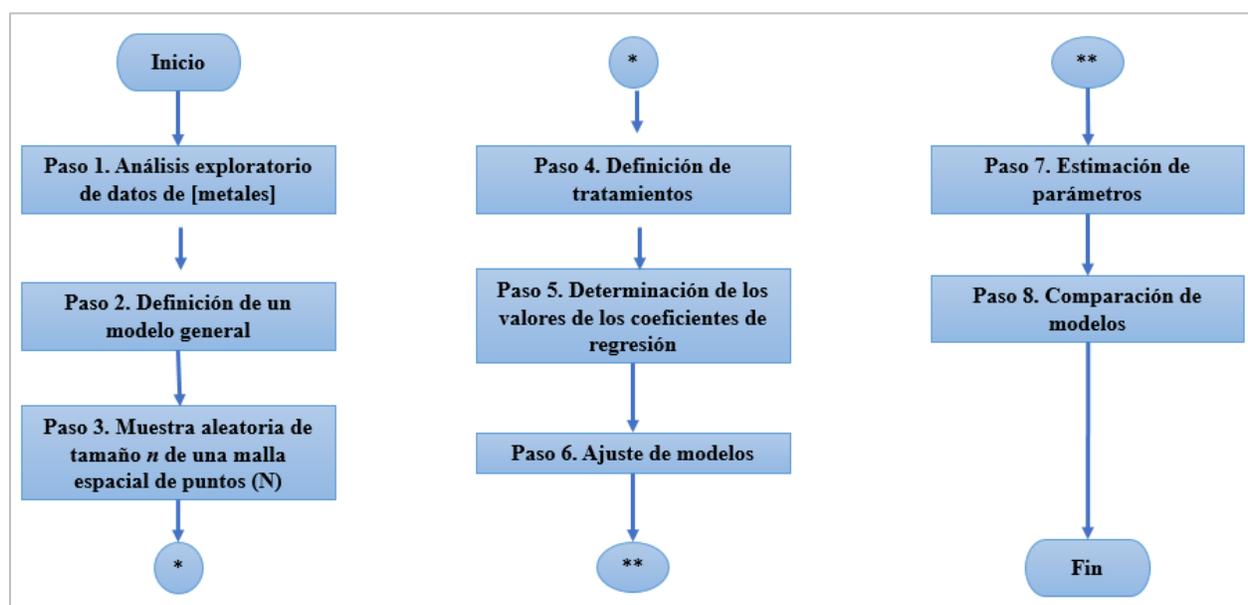
1. Comparar el comportamiento de los estimadores de los modelos seleccionados a través del sesgo y Error Cuadrado Medio (ECM), dada la generación de la variable de respuesta con diferentes tipos de matrices de pesos, niveles de autocorrelación y distribuciones asimétricas.
2. Identificar los mejores modelos de acuerdo al sesgo y el Error Cuadrado Medio (ECM) de los parámetros estimados.
3. Aplicar los modelos seleccionados a datos de concentración de metales pesados de los Distritos de Riego 003 Tula, 100 Alfajayucan y 112 Ajacuba, proporcionados por la Comisión Nacional del Agua (CONAGUA).
4. Elaborar mapas de predicción de concentración de metales pesados en función de los modelos estimados.
5. Elaborar mapas de probabilidad de exceder los límites máximos de contaminantes permisibles estipulados en la NOM-001-SEMARNAT-1996.
6. Proponer un tamaño de muestra óptimo en función del nivel de autocorrelación de los diferentes metales pesados a estudiar.

## V. MATERIALES Y MÉTODOS

El propósito del estudio de simulación es identificar los mejores modelos que sean capaces de detectar características propias de datos espaciales como lo es la autocorrelación en la variable respuesta, distribuciones diferentes a la normal y heteroscedasticidad.

### 5.1 Estudio de simulación

En la simulación se trató de imitar el comportamiento que podría llegar a tener la concentración de metales pesados en suelo, partiendo de la concentración de seis contaminantes en el Valle del Mezquital, Hidalgo. En la Figura 1 se resumen los pasos que se llevaron a cabo para la simulación de la variable respuesta, el ajuste de modelos y la evaluación de los mismos. Toda la programación tanto de la simulación como de la aplicación de los modelos a datos reales se llevó a cabo en el Software R.



**Figura 1. Diagrama de bloques del proceso de simulación.**

A continuación se describen las condiciones del análisis de simulación.

#### **Paso 1. Análisis exploratorio de datos de concentración de metales**

En primer lugar se realizó el análisis exploratorio de los seis metales proporcionados y se realizó la prueba de Shapiro–Wilk para probar normalidad. En todos los casos se rechazó la hipótesis nula con un nivel de significancia de 0.05 (Cuadro 11).

## Paso 2. Definición de un modelo general

Como modelo general se utilizó el propuesto por Viton (2010) donde la variable de respuesta autocorrelacionada depende de covariables y de errores:

$$y = X\beta + \lambda Wy + \varepsilon + v \quad (44)$$

Donde  $X$  es la matriz diseño,  $\beta$  es el vector de los coeficientes de regresión,  $\lambda$  es el nivel de autocorrelación de la variable respuesta,  $W$  es la matriz de pesos espaciales,  $\varepsilon$  es el error autocorrelacionado (definido en el paso 4) y  $v$  es ruido blanco.

## Paso 3. Selección de muestra aleatoria de puntos

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  $n = 100$ .

## Paso 4. Definición de tratamientos

De acuerdo con el modelo general, se generaron variables respuesta bajo diferentes tratamientos. En primer lugar se consideraron tres distribuciones: Gamma, Inversa Gaussiana y Normal (como de referencia) para representar el comportamiento de Cromo (Cr) en el Valle del Mezquital. A los datos reales se les adicionó 4 unidades 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  $y \sim IG(\mu, \gamma)$ .

Referente al error autocorrelacionado, se consideraron dos tipos (Anselin, 2005), el Espacial Autorregresivo Aleatorio (SAR, *Spatial Autoregressive Random*), que se expresa como:  $\varepsilon = (I - \rho W)^{-1}u$  y Espacial de Promedios Móviles (SMA, *Spatial Moving Average*), expresado como:  $\varepsilon = \rho W u + u$ . Donde  $W$  es la matriz de pesos espaciales,  $\rho$  es el nivel de autocorrelación correspondiente a los residuales y  $u$  es el error autocorrelacionado. Para la matriz de vecindades, se implementaron cinco tipos de matrices de pesos:  $W, B, C, U$  y  $S$ . Los niveles de autocorrelación para  $\lambda$  y  $\rho$ , fueron los mismos:  $\{0, 0.2, 0.5, 0.7, 0.9\}$ .

En total se obtuvieron 150 combinaciones bajo los diferentes tratamientos con los que fueron generadas las variables respuesta (Anexo 1) para probar la efectividad de estimación y precisión de los modelos GLM, GLMM, GAM y GAMM, de manera resumida se presentan en el Cuadro 2.

**Cuadro 2. Tratamientos para la generación de variable respuesta.**

Combinación	Distribución de $y$	Tipo de error	Tipo de $W$	Nivel de $\lambda, \rho$
1	Normal	SAR	W	0.00
2	Normal	SAR	W	0.20
3	Normal	SAR	W	0.50
4	Normal	SAR	W	0.70
5	Normal	SAR	W	0.90
6	Normal	SAR	B	0.00
⋮	⋮	⋮	⋮	⋮
26	Normal	SMA	W	0.00
⋮	⋮	⋮	⋮	⋮
50	Normal	SMA	S	0.90
51	Gamma	SAR	W	0.00
⋮	⋮	⋮	⋮	⋮
101	Inv. Gau.	SAR	W	0.00
⋮	⋮	⋮	⋮	⋮
150	Inv. Gau.	SMA	S	0.90

**Nota:** Inv. Gau: Inversa Gaussiana.

### Paso 5. Determinación de valores de los coeficientes de regresión

Los valores de los coeficientes de la regresión  $y = \beta_{00} + \beta_{01}x_1 + \beta_{02}x_2 + \beta_{03}x_3 + \beta_{04}x_4 + \lambda W y + \varepsilon + v$ , fueron  $\beta_{00} = 1$ ,  $\beta_{01} = 5$ ,  $\beta_{02} = 10$ ,  $\beta_{03} = 1$  y  $\beta_{04} = 1$ . Se tomaron en cuenta

cuatro covariables; 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 ( $x_3, x_4$ ) corresponden a las coordenadas de longitud y latitud del punto seleccionado en la malla,  $\varepsilon$  se simuló como se definió en el paso 3 y  $v$  como  $N(0,1)$ , para cada punto de la malla. De acuerdo al modelo general (44), se simuló un dato en cada punto de la malla en cada una de las 1000 repeticiones.

### **Paso 6. Ajuste de modelos**

Para cada una de las combinaciones mencionadas en el paso 3, fueron ajustados los siguientes modelos lineales y no lineales en cada repetición.

#### **Modelos lineales:**

Modelo Lineal Generalizado:  $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ ,

Modelo Lineal Generalizado Mixto:  $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + Z\gamma$ ,

#### **Modelos no lineales:**

Modelo Aditivo Generalizado:  $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + f_3 x_3 + f_4 x_4$  y

Modelo Aditivo Generalizado Mixto:  $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + f_3 x_3 + f_4 x_4 + Z\gamma$

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  $\gamma$  el vector que contiene los efectos aleatorios.

### **Paso 7. Estimación de parámetros**

Para cada uno de los modelos ajustados y en cada repetición fueron estimados los parámetros  $\beta_0, \beta_1, \beta_2$  mediante simulación bootstrap. Los parámetros  $\beta_3, \beta_4$  corresponden 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.

## **Paso 8. Comparación de los modelos**

Las propiedades de los estimadores producidos de los diferentes modelos ajustados para cada uno de los tratamientos, se pueden comparar de varias maneras, para la presente investigación se consideraron las mencionadas por Anselin (2005):

### ***Sesgo***

El cual se calcula como la diferencia entre el promedio a través de las repeticiones y el valor inicial del parámetro.

$$Sesgo(\hat{\beta}) = \sum_{r=1}^R \frac{\hat{\beta}_r}{R} - \beta \quad (45)$$

Donde  $R$  es el número total de repeticiones del proceso de simulación,  $\beta$  es el valor inicial del coeficiente.

### ***Varianza***

Definida por la siguiente expresión.

$$Var(\hat{\beta}) = \left(\frac{1}{R}\right) \sum_{r=1}^R (\hat{\beta}_r - \bar{\beta})^2 \quad (46)$$

### ***Error Cuadrado Medio (ECM)***

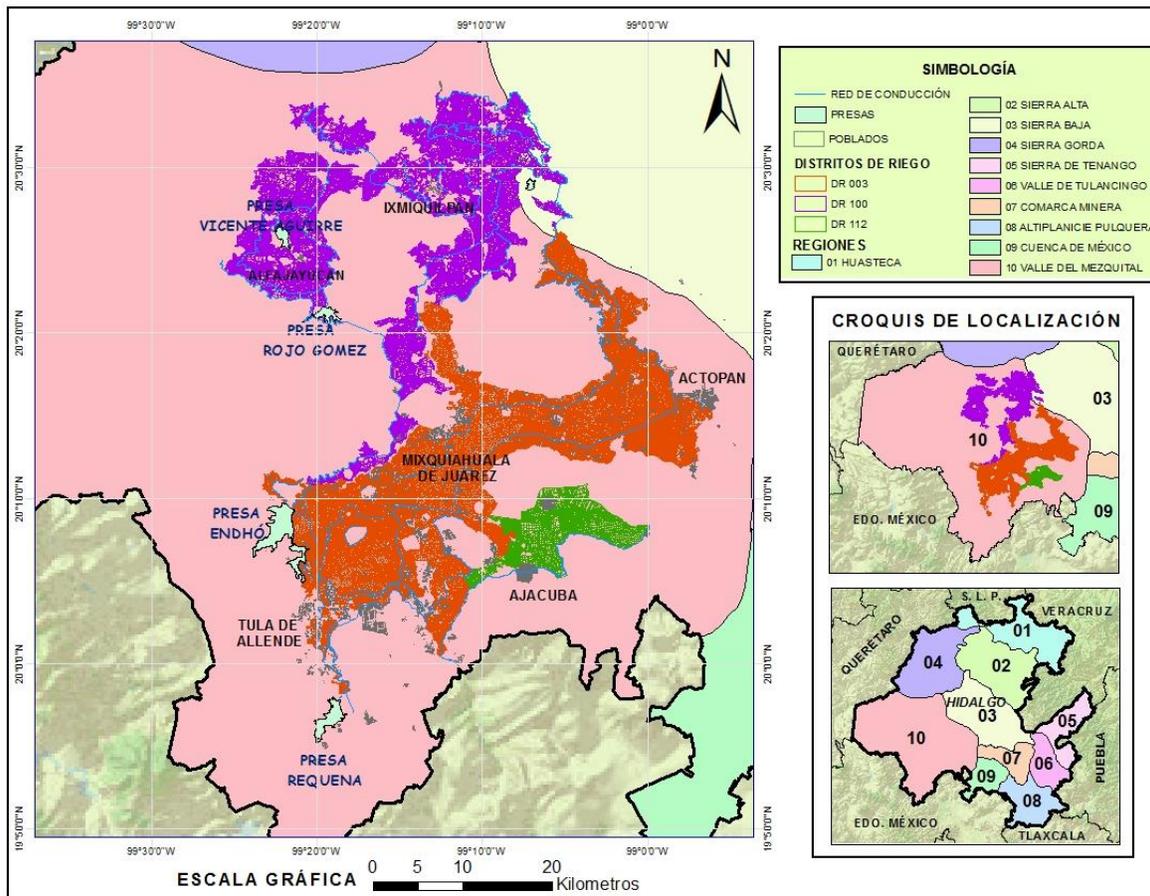
El ECM evalúa la calidad de un estimador o conjunto de predicciones en cuanto a su variación y el grado de sesgo, definido por la siguiente expresión:

$$ECM(\hat{\beta}) = \left(\frac{1}{R}\right) \sum_{r=1}^R (\hat{\beta}_r - \beta)^2 \quad (47)$$

## 5.2 Aplicación

### 5.2.1 Ubicación del área de estudio

El Valle del Mezquital se encuentra en la parte sureste del estado de Hidalgo, a 50 km al noreste de la Ciudad de México (Prieto *et al.*, 2007). La zona de estudio se reduce a los Distritos de Riego (DR) 003 Tula, 100 Alfajayucan y 112 Ajacuba (Figura 2), ubicados en este valle.



**Figura 2. Localización de los DR 003 Tula, DR 100 Alfajayucan y DR 112 Ajacuba.**

**Fuente:** Elaboración propia, con datos de CONAGUA y SGM.

El clima árido, en la mayor parte del estado de Hidalgo y en específico del Valle del Mezquital aunado a la poca precipitación anual, origina la necesidad de utilizar aguas residuales para el riego que propician el incremento de las tierras cultivables con una mayor capacidad de producción (Mendoza, 1981). Las aguas residuales se aplican sin tratamiento alguno, de los principales contaminantes incorporados a suelos a través del riego con agua residual están los metales pesados.

### **5.2.2 Muestras de suelo y obtención de datos**

Los datos fueron facilitados por la Comisión Nacional del Agua (CONAGUA), los cuales consisten en 72 observaciones de un muestreo realizado en los Distritos de Riego 003 Tula, 100 Alfajayucan y 112 Ajacuba durante los meses de septiembre y octubre de 2013. Las muestras se colectaron a 0-20 cm de profundidad. Los procedimientos seguidos para el análisis de las muestras de suelo se realizaron con base en la norma NOM-147-SEMARNAT/SSA1-2004 que establece criterios para determinar las concentraciones de remediación de suelos contaminados por arsénico, cadmio, cromo, níquel y plomo, así como en la norma NOM-021-SEMARNAT-2000 que establece las especificaciones de fertilidad, salinidad y clasificación de suelos, estudio, muestreo y análisis. Los metales pesados que se analizaron fueron: Cadmio (Cd), Cobre (Cu), Cromo (Cr), Níquel (Ni), Plomo (Pb) y Zinc (Zn), las unidades en las que se presentan las concentraciones de metales pesados son mg/L.

## VI. RESULTADOS

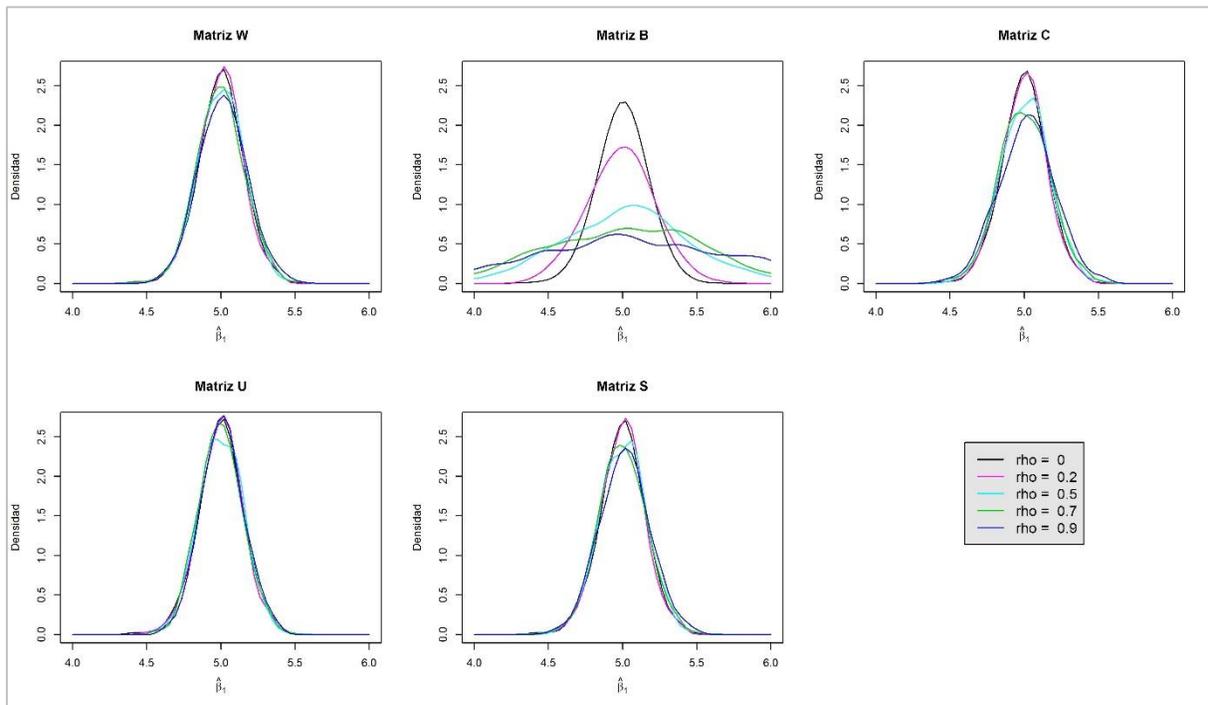
En una primera parte se analizan los resultados obtenidos del estudio de simulación; y en la segunda parte, se aplican los modelos de regresión a un conjunto de datos.

### 6.1 Estudio de simulación

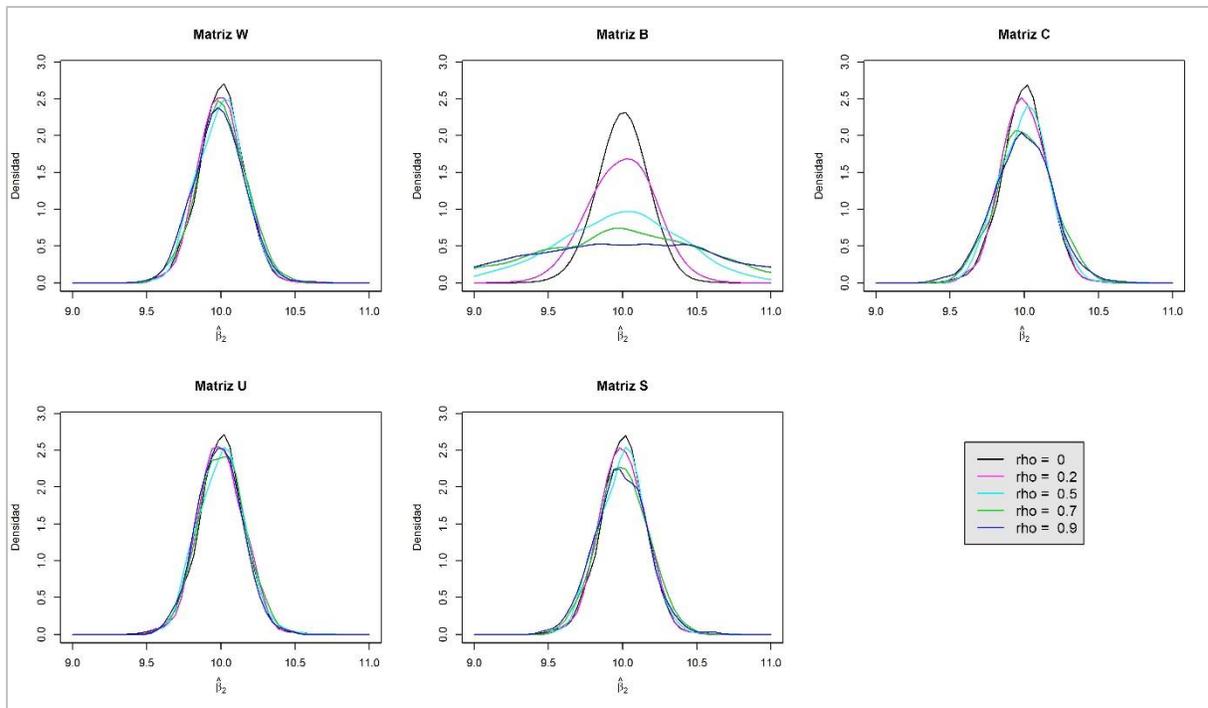
#### 6.1.1 Efecto del tipo de matriz

Al obtener las gráficas de la función de distribución de los estimadores (Figura 3 y Figura 4) para cada uno de los tratamientos en función del tipo de matriz de vecindad y el nivel de autocorrelación, fue notorio el efecto que tuvo el tipo de matriz en la estimación de los parámetros.

Cuando la variable respuesta de la expresión 44 es generada a través de la matriz de pesos de tipo U, la estimación de los parámetros  $\beta_1$  y  $\beta_2$  de los modelos de regresión (45, 46, 47 y 48) es más robusta al nivel de autocorrelación dado que la estimación se centra más en el valor inicial de los parámetros, presentando menor valor de ECM (Cuadro 3 y Cuadro 4). Para el análisis de los resultados no se considera  $\beta_0$  porque es un factor de ajuste.



**Figura 3. Gráficas de densidad para  $\hat{\beta}_1$ , generadas a partir del ajuste de un modelo GLM donde  $y \sim IG(4.622, 27.561)$  con error de tipo SMA.**



**Figura 4. Gráficas de densidad para  $\hat{\beta}_2$ , generadas a partir del ajuste de un modelo GLM donde  $y \sim IG(4.622, 27.561)$  con error de tipo SMA.**

Es evidente que cuando se implementó la matriz de tipo B, la estimación de los coeficientes presenta mayor variabilidad independientemente del nivel de autocorrelación.

**Cuadro 3. Efecto del tipo de matriz en la estimación de  $\beta_1$  y  $\beta_2$  con modelos GLM y GAM.**

Dist. $y$	Error	$W$	$\lambda, \rho$	GLM				GAM			
				Sesgo		ECM		Sesgo		ECM	
				$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
Normal	SMA	W	0.50	0.000	0.004	0.024	0.022	-0.001	0.003	0.024	0.023
Normal	SMA	B	0.50	0.000	0.027	0.173	0.173	-0.002	0.023	0.169	0.174
Normal	SMA	C	0.50	-0.002	0.006	0.027	0.025	-0.002	0.005	0.027	0.026
Normal	SMA	U	0.50	-0.002	0.002	0.022	0.021	-0.002	0.002	0.023	0.022
Normal	SMA	S	0.50	-0.001	0.005	0.025	0.022	-0.002	0.004	0.025	0.023
Gamma	SMA	W	0.50	-0.004	0.003	0.024	0.023	-0.003	0.003	0.024	0.023
Gamma	SMA	B	0.50	0.019	-0.001	0.179	0.172	0.022	0.000	0.176	0.169
Gamma	SMA	C	0.50	-0.001	0.001	0.028	0.026	0.000	0.001	0.028	0.026
Gamma	SMA	U	0.50	-0.006	0.002	0.023	0.022	-0.005	0.003	0.023	0.022
Gamma	SMA	S	0.50	-0.003	0.002	0.025	0.024	-0.002	0.002	0.025	0.024
Inv. Gau.	SMA	W	0.50	0.003	-0.001	0.023	0.025	0.004	0.000	0.023	0.026
Inv. Gau.	SMA	B	0.50	0.029	-0.025	0.188	0.175	0.026	-0.017	0.185	0.172
Inv. Gau.	SMA	C	0.50	0.005	-0.002	0.027	0.027	0.006	-0.001	0.027	0.028
Inv. Gau.	SMA	U	0.50	0.001	0.003	0.022	0.024	0.002	0.003	0.022	0.025
Inv. Gau.	SMA	S	0.50	0.004	-0.001	0.024	0.025	0.005	0.000	0.024	0.026

**Nota:** Dist.: Distribución; Inv. Gau: Inversa Gaussiana.

**Cuadro 4. Efecto del tipo de matriz en la estimación de  $\beta_1$  y  $\beta_2$  con modelos GLMM y GAMM.**

Dist. y	Error	W	$\lambda, \rho$	GLMM				GAMM			
				Sesgo		ECM		Sesgo		ECM	
				$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
Normal	SMA	W	0.50	0.000	0.004	0.024	0.022	0.000	0.005	0.024	0.022
Normal	SMA	B	0.50	0.000	0.027	0.173	0.173	0.001	0.026	0.173	0.174
Normal	SMA	C	0.50	-0.002	0.006	0.027	0.025	-0.002	0.008	0.027	0.025
Normal	SMA	U	0.50	-0.002	0.002	0.022	0.021	-0.002	0.003	0.022	0.021
Normal	SMA	S	0.50	-0.001	0.005	0.025	0.022	-0.001	0.006	0.025	0.023
Gamma	SMA	W	0.50	-0.004	0.003	0.024	0.023	0.003	0.002	0.024	0.022
Gamma	SMA	B	0.50	0.019	-0.001	0.179	0.172	0.019	-0.003	0.185	0.157
Gamma	SMA	C	0.50	-0.001	0.001	0.028	0.026	0.009	0.007	0.028	0.024
Gamma	SMA	U	0.50	-0.006	0.002	0.023	0.022	0.000	0.004	0.022	0.021
Gamma	SMA	S	0.50	-0.003	0.002	0.025	0.024	0.009	0.000	0.025	0.023
Inv. Gau.	SMA	W	0.50	0.003	-0.001	0.023	0.025	0.003	-0.001	0.023	0.025
Inv. Gau.	SMA	B	0.50	0.029	-0.025	0.188	0.175	0.028	-0.024	0.188	0.176
Inv. Gau.	SMA	C	0.50	0.005	-0.002	0.027	0.027	0.005	-0.002	0.027	0.027
Inv. Gau.	SMA	U	0.50	0.001	0.003	0.022	0.024	0.001	0.003	0.022	0.024
Inv. Gau.	SMA	S	0.50	0.004	-0.001	0.024	0.025	0.004	-0.002	0.024	0.025

**Nota:** Dist.: Distribución; Inv. Gau: Inversa Gaussiana.

Al analizar el ECM de los tratamientos que corresponden a la estimación de  $\beta_1$  y  $\beta_2$  generados a partir del ajuste de modelos lineales (GLM y GLMM) y no lineales (GAM y GAMM) donde la variable de respuesta proviene de las tres diferentes distribuciones y diferentes matrices de vecindad pero con el mismo tipo de error (SMA) y mismo nivel de autocorrelación (0.50); se aprecia que en todos los casos la matriz de tipo U generó menor valor de ECM aunque no de sesgo.

### 6.1.2 Efecto del nivel de autocorrelación

Como se pudo observar en las figuras anteriores, el nivel de autocorrelación también afecta la estimación de los coeficientes  $\beta_1$  y  $\beta_2$ , en este caso se analizó a partir de la matriz de tipo W, dado que la matriz de tipo U es más robusta al cambio del nivel de autocorrelación. De manera general se observó en el Cuadro 5 y

Cuadro 6 que los valores del ECM tienden a aumentar ligeramente, al aumentar el nivel de autocorrelación cuando la variable de respuesta fue generada a través de un error de tipo SAR y con las tres diferentes distribuciones, aunque a manera de ejemplificar solo se presentan los datos correspondientes a la distribución Normal e Inversa Gaussiana.

**Cuadro 5. Efecto del nivel de autocorrelación en la estimación de  $\beta_1$  y  $\beta_2$  con modelos GLM y GAM.**

Dist. y	Error	W	$\lambda, \rho$	GLM				GAM			
				Sesgo		ECM		Sesgo		ECM	
				$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
Normal	SAR	W	0.00	-0.005	0.002	0.023	0.021	-0.005	0.002	0.023	0.021
Normal	SAR	W	0.20	-0.001	-0.005	0.022	0.022	-0.001	-0.005	0.022	0.022
Normal	SAR	W	0.50	0.000	0.004	0.026	0.023	-0.001	0.003	0.025	0.024
Normal	SAR	W	0.70	-0.003	0.001	0.028	0.030	-0.002	-0.001	0.027	0.029
Normal	SAR	W	0.90	0.004	-0.008	0.038	0.038	0.005	-0.004	0.034	0.034
Inv. Gau.	SAR	W	0.00	0.001	0.002	0.023	0.022	0.001	0.003	0.024	0.023
Inv. Gau.	SAR	W	0.20	-0.001	0.001	0.022	0.022	-0.001	0.001	0.022	0.023
Inv. Gau.	SAR	W	0.50	0.004	-0.001	0.024	0.026	0.004	0.000	0.024	0.027
Inv. Gau.	SAR	W	0.70	-0.005	0.002	0.029	0.032	0.002	0.004	0.027	0.030
Inv. Gau.	SAR	W	0.90	0.005	-0.009	0.037	0.041	-0.002	0.007	0.038	0.035

Nota: Dist.: Distribución; Inv. Gau: Inversa Gaussiana.

**Cuadro 6. Efecto del nivel de autocorrelación en la estimación de  $\beta_1$  y  $\beta_2$  con modelos GLMM y GAMM.**

Dist. y	Error	W	$\lambda, \rho$	GLMM				GAMM			
				Sesgo		ECM		Sesgo		ECM	
				$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
Normal	SAR	W	0.00	-0.005	0.002	0.023	0.021	-0.005	0.001	0.023	0.021
Normal	SAR	W	0.20	-0.001	-0.005	0.022	0.022	-0.002	-0.005	0.022	0.022
Normal	SAR	W	0.50	0.000	0.004	0.026	0.023	0.000	0.004	0.025	0.023
Normal	SAR	W	0.70	-0.003	0.001	0.028	0.030	-0.002	0.001	0.028	0.028
Normal	SAR	W	0.90	0.004	-0.008	0.038	0.038	0.006	-0.002	0.034	0.036
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

Nota: Dist.: Distribución; Inv. Gau: Inversa Gaussiana.

### 6.1.3 Efecto del tipo de error

Al comparar los resultados de sesgo y ECM generados a partir de la matriz tipo S y con un nivel de correlación de 0.2 y 0.5 pero con diferente tipo de error se pudieron observar ciertos patrones. Para las tres distribuciones, el sesgo de los estimadores es el mismo para los mismos niveles de autocorrelación en los modelos GLM, GAM y GLMM. En lo que se refiere al ECM, los valores más pequeños se presentaron con el error de tipo SMA cuando el nivel de autocorrelación es de 0.50; a nivel de 0.20 no se identificaron cambios en los estimadores (Cuadro 7 y Cuadro 8).

**Cuadro 7. Efecto del tipo de error en la estimación de  $\beta_1$  y  $\beta_2$  con modelos GLM y GAM.**

Dist. $y$	Error	$W$	$\lambda, \rho$	GLM				GAM			
				Sesgo		ECM		Sesgo		ECM	
				$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
Normal	SAR	S	0.20	-0.001	-0.005	0.022	0.022	0.000	-0.005	0.022	0.022
Normal	SAR	S	0.50	-0.001	0.005	0.026	0.024	-0.002	0.004	0.026	0.024
Normal	SMA	S	0.20	-0.001	-0.005	0.022	0.022	0.000	-0.006	0.022	0.022
Normal	SMA	S	0.50	-0.001	0.005	0.025	0.022	-0.002	0.004	0.025	0.023
Gamma	SAR	S	0.20	-0.006	-0.005	0.022	0.023	-0.006	-0.004	0.022	0.024
Gamma	SAR	S	0.50	-0.003	0.002	0.027	0.025	-0.002	0.002	0.026	0.025
Gamma	SMA	S	0.20	-0.006	-0.005	0.022	0.023	-0.006	-0.004	0.022	0.024
Gamma	SMA	S	0.50	-0.003	0.002	0.025	0.024	-0.002	0.002	0.025	0.024
Inv. Gau.	SAR	S	0.20	-0.002	0.001	0.022	0.022	-0.002	0.001	0.022	0.023
Inv. Gau.	SAR	S	0.50	0.004	-0.002	0.025	0.026	0.005	0.000	0.025	0.027
Inv. Gau.	SMA	S	0.20	-0.002	0.001	0.022	0.022	-0.002	0.001	0.022	0.023
Inv. Gau.	SMA	S	0.50	0.004	-0.001	0.024	0.025	0.005	0.000	0.024	0.026

**Nota:** Dist.: Distribución; Inv. Gau: Inversa Gaussiana.

**Cuadro 8. Efecto del tipo de error en la estimación de  $\beta_1$  y  $\beta_2$  con modelos GLMM y GAMM.**

Dist. $y$	Error	$W$	$\lambda, \rho$	GLMM				GAMM			
				Sesgo		ECM		Sesgo		ECM	
				$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
Normal	SAR	S	0.20	-0.001	-0.005	0.022	0.022	-0.001	-0.005	0.022	0.022
Normal	SAR	S	0.50	-0.001	0.005	0.026	0.024	-0.001	0.006	0.026	0.024
Normal	SMA	S	0.20	-0.001	-0.005	0.022	0.022	-0.002	-0.004	0.022	0.022
Normal	SMA	S	0.50	-0.001	0.005	0.025	0.022	-0.001	0.006	0.025	0.023
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
Inv. Gau.	SAR	S	0.20	-0.002	0.001	0.022	0.022	0.000	0.003	0.022	0.022
Inv. Gau.	SAR	S	0.50	0.004	-0.002	0.025	0.026	-0.001	0.003	0.027	0.025
Inv. Gau.	SMA	S	0.20	-0.002	0.001	0.022	0.022	-0.002	0.001	0.022	0.022
Inv. Gau.	SMA	S	0.50	0.004	-0.001	0.024	0.025	0.004	-0.002	0.024	0.025

**Nota:** Dist.: Distribución; Inv. Gau: Inversa Gaussiana.

#### 6.1.4 Efecto de los modelos

Como se identificó en el apartado 6.1.1, la matriz de tipo U presentó características de robustez al nivel de autocorrelación por lo que se decidió hacer la comparación de los diferentes modelos de ajuste con dicho tipo de matriz, con niveles de 0.5 y 0.9.

No se identificaron diferencias entre el modelo lineal GLM y el modelo mixto GLMM, pero si en los tratamientos 120 y 145 el sesgo del estimador  $\hat{\beta}_1$  resultó ser cero.

En el Cuadro 9, son más notorios los cambios entre los modelos GAM y GAMM, siendo los segundos los que presentaron menores valores de varianza y ECM para las tres distribuciones, los diferentes tipos de error y nivel de autocorrelación en ambos estimadores. El índice de autocorrelación afecta la estimación de los parámetros, el cual es captado mejor por los modelos mixtos.

Debido a que la matriz de tipo U produjo estimadores más robustos e insesgados, no es posible identificar el efecto que tienen los modelos al momento de la estimación de los parámetros, pero al emplear la matriz de tipo C, con un error SMA y con un nivel de autocorrelación de 0.5, también se observó que la menor variabilidad y ECM es producida por el modelo GAMM (Cuadro 10).

Con el fin de ampliar la investigación y ver cómo se comportan los modelos al ajuste de datos reales se procedió a ajustar los mismos modelos a un conjunto de datos de la zona de estudio.

**Cuadro 9. Efecto del modelo GAM y GAMM en la estimación de  $\beta_1$  y  $\beta_2$  con matriz de tipo U.**

Distribución de y	Error	$\lambda, \rho$	GAM						GAMM					
			Sesgo		Varianza		ECM		Sesgo		Varianza		ECM	
			$\hat{\beta}_1$	$\hat{\beta}_2$										
Normal	SAR	0.50	-0.001	-0.005	0.022	0.022	0.022	0.022	-0.001	-0.005	0.022	0.022	0.022	0.022
Normal	SAR	0.90	-0.002	0.002	0.023	0.022	0.023	0.022	-0.002	0.002	0.022	0.021	0.022	0.021
Normal	SMA	0.50	-0.001	-0.005	0.022	0.022	0.022	0.022	-0.001	-0.005	0.022	0.022	0.022	0.022
Normal	SMA	0.90	-0.002	0.002	0.023	0.022	0.023	0.022	-0.002	0.002	0.022	0.021	0.022	0.021
Gamma	SAR	0.50	0.001	-0.002	0.022	0.023	0.022	0.023	0.001	-0.004	0.021	0.022	0.021	0.022
Gamma	SAR	0.90	-0.003	-0.003	0.022	0.022	0.022	0.022	-0.004	-0.004	0.021	0.021	0.021	0.021
Gamma	SMA	0.50	-0.006	0.001	0.022	0.022	0.022	0.022	-0.006	-0.001	0.021	0.021	0.021	0.021
Gamma	SMA	0.90	0.006	0.001	0.023	0.022	0.023	0.022	0.006	0.001	0.022	0.021	0.023	0.021
Inv. Gau.	SAR	0.50	0.000	0.003	0.022	0.023	0.022	0.023	0.001	0.002	0.021	0.022	0.021	0.022
Inv. Gau.	SAR	0.90	0.002	0.002	0.022	0.025	0.022	0.025	0.000	0.002	0.021	0.024	0.021	0.024
Inv. Gau.	SMA	0.50	0.001	0.003	0.022	0.023	0.022	0.023	0.001	0.003	0.021	0.022	0.021	0.022
Inv. Gau.	SMA	0.90	0.002	0.001	0.022	0.025	0.022	0.025	0.000	0.001	0.021	0.024	0.021	0.024

Nota: Inv. Gau: Inversa Gaussiana.

**Cuadro 10. Efecto de los modelos en la estimación de  $\beta_1$  y  $\beta_2$  con matriz de tipo C, error tipo SMA y nivel de autocorrelación de 0.5.**

Dist. y	Modelos no mixtos								Modelos mixtos							
	GLM				GAM				GLMM				GAMM			
	Varianza		ECM		Varianza		ECM		Varianza		ECM		Varianza		ECM	
$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	
Normal	0.027	0.025	0.027	0.025	0.027	0.026	0.027	0.026	0.027	0.025	0.027	0.025	0.027	0.025	0.027	0.025
Gamma	0.028	0.026	0.028	0.026	0.028	0.026	0.028	0.026	0.028	0.026	0.028	0.026	0.028	0.024	0.028	0.024
Inv. Gau.	0.027	0.027	0.027	0.027	0.027	0.028	0.027	0.028	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027

Nota: Inv. Gau: Inversa Gaussiana.

## 6.2 Aplicación

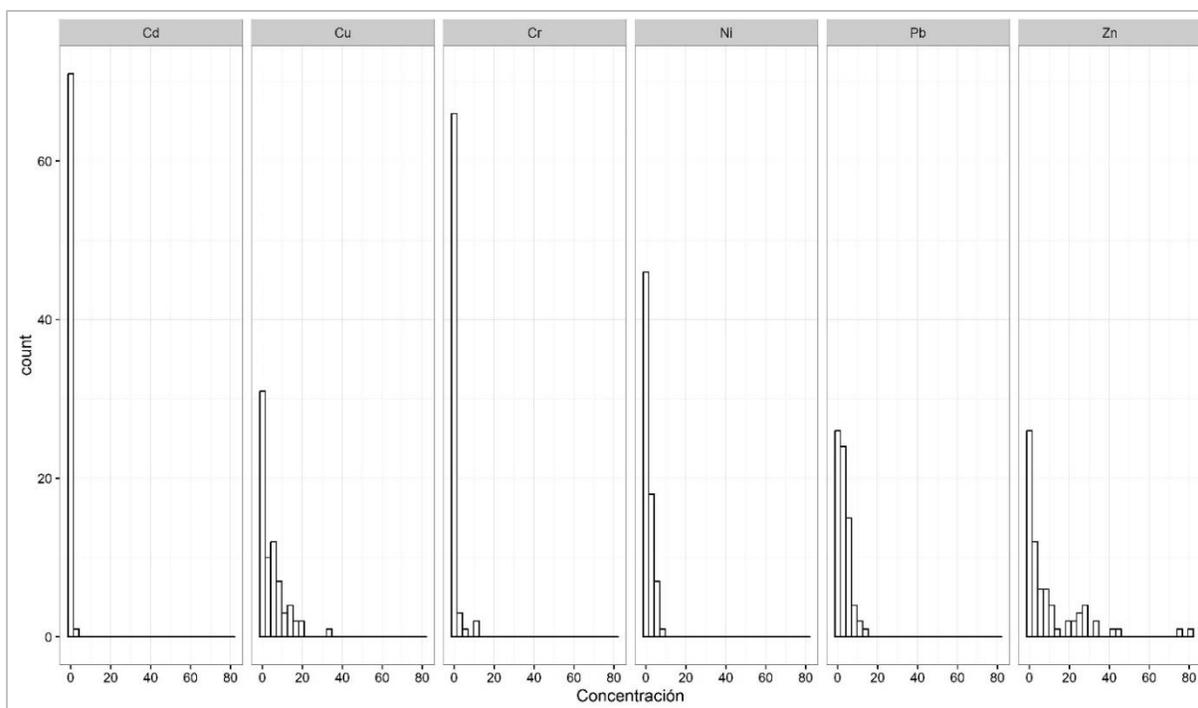
### 6.2.1 Análisis exploratorio

El Cuadro 11 detalla los resultados correspondientes al análisis exploratorio de los datos. Como se aprecia, las concentraciones de los metales son mayor o igual a cero. De acuerdo a la prueba de Shapiro–Wilk, en todos los casos la hipótesis nula de normalidad se rechazó con un nivel de significancia de 0.05, es decir, ninguno de los metales se distribuye normal. La concentración de metales pesados se caracteriza por ser positiva y sesgada por la derecha (Figura 5).

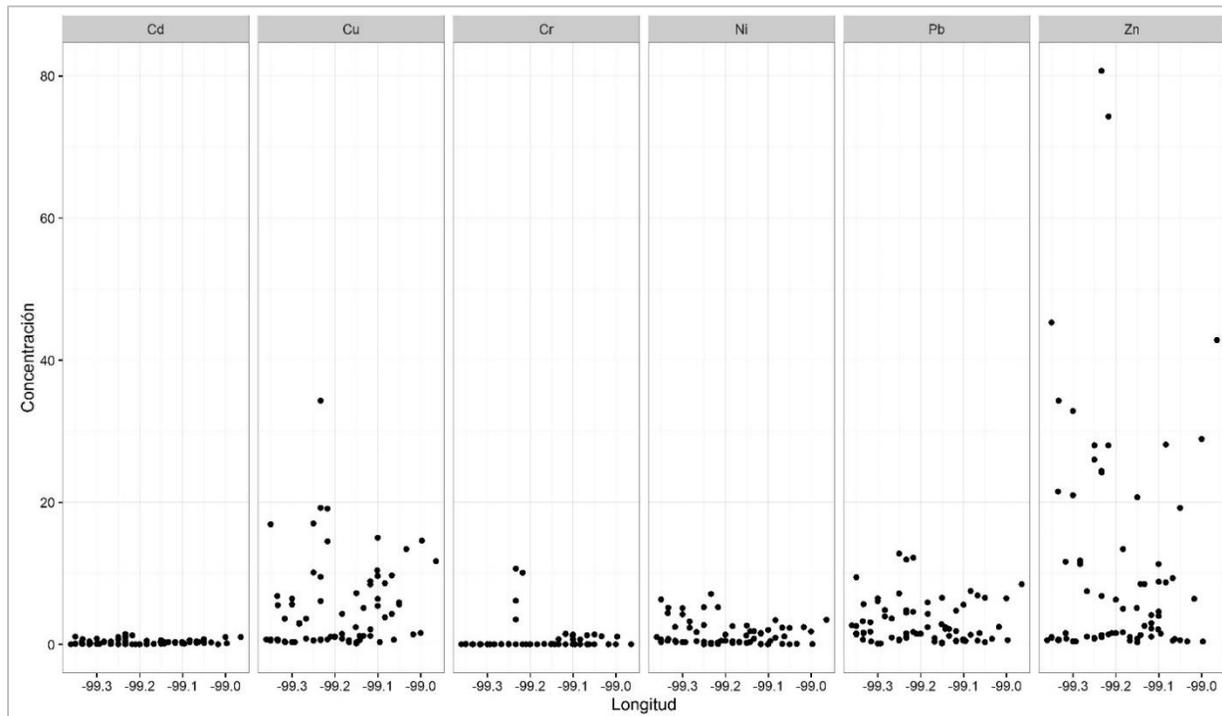
**Cuadro 11. Estadísticas descriptivas de la concentración de los metales.**

<b>Metal</b>	<b>Mínimo</b>	<b>Mediana</b>	<b>Media</b>	<b>Máximo</b>	<b>Prueba Shapiro-Wilks (p-value)</b>
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

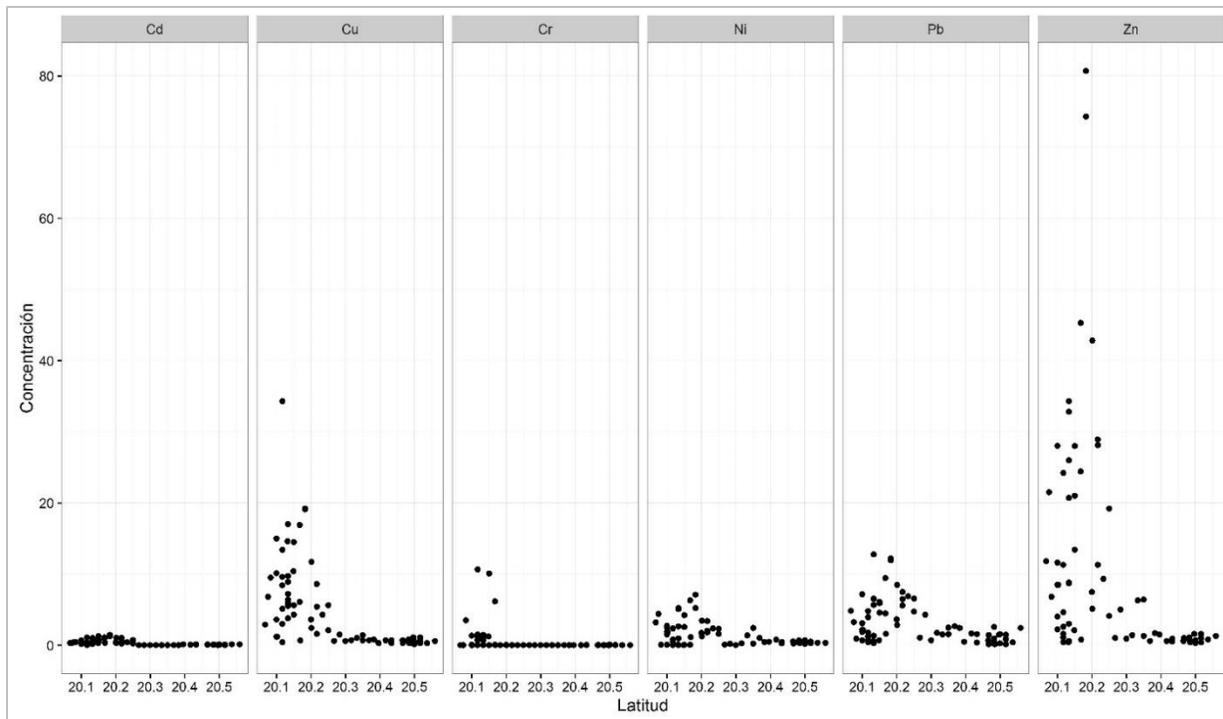
La dispersión de los metales es aleatoria sobre la longitud (Figura 6), es decir no se presentó tendencia alguna, en cambio, se observaron tendencias en latitud y altitud; a menores latitudes (hacia el sur), se notan valores más altos en las concentraciones de los metales (Figura 7). También los valores más altos se observa a mayor altitud (Figura 8). Los valores más altos se observan en áreas con mayor altitud, dado que el agua contaminada inicia su distribución por gravedad desde las áreas agrícolas altas hacia las zonas bajas.



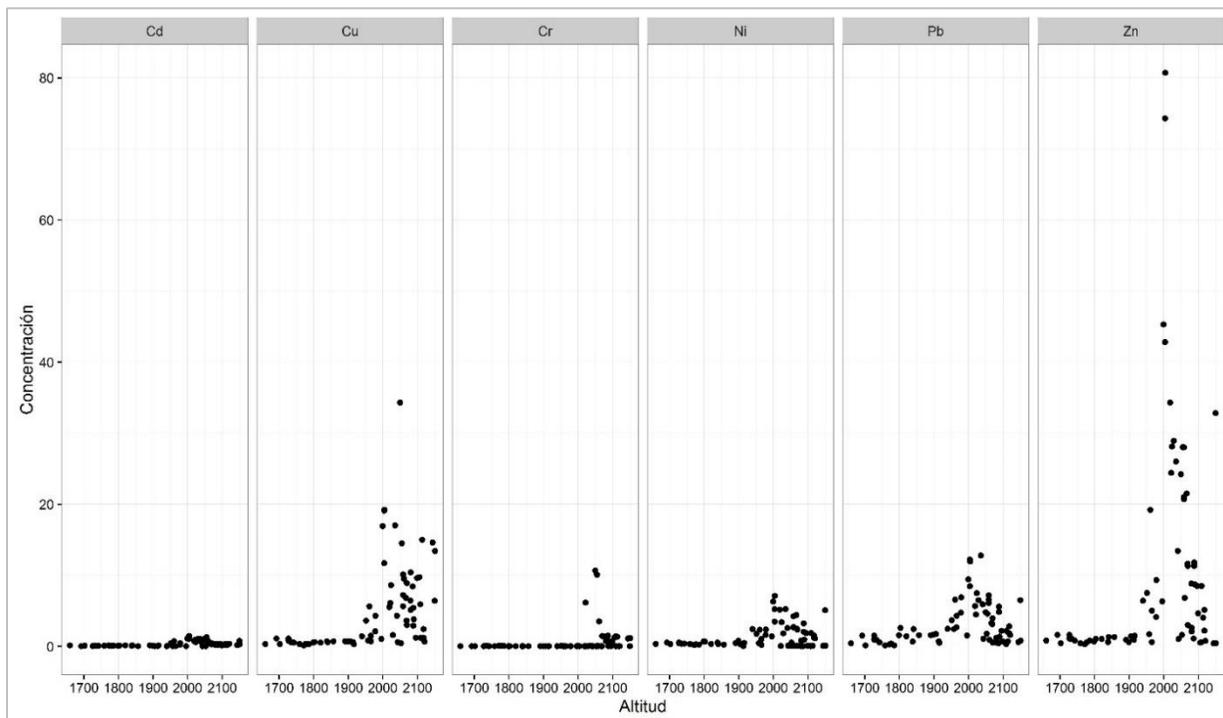
**Figura 5. Histogramas de la concentración de los metales.**



**Figura 6. Longitud vs. concentración de metales.**



**Figura 7. Latitud vs. concentración de metales.**



**Figura 8. Altitud vs. concentración de metales.**

### 6.2.2 Ajuste de modelos

Para cada uno de los metales se ajustaron los modelos: GLM, GAM, GLMM y GAMM. El modelo lineal (LM) se ajustó como “testigo” para comparar los modelos. Las covariables consideradas fueron las tres coordenadas (longitud, latitud y altitud). Dado que la longitud no fue significativa, esta fue excluida de los modelos. Los modelos ajustados para cada metal fueron los siguientes:

**Modelo Lineal y Lineal Generalizado:**  $\eta = \beta_0 + \beta_1 lon + \beta_2 lat,$

**Modelo Lineal Generalizado Mixto:**  $\eta = \beta_0 + \beta_1 lon + \beta_2 lat + Z\gamma,$

**Modelo Aditivo Generalizado:**  $\eta = f_1 lon + f_2 lat, y$

**Modelo Aditivo Generalizado Mixto:**  $\eta = f_1 lon + f_2 lat + Z\gamma$

Donde  $\beta_0, \beta_1, \beta_2$  son los coeficientes de regresión de la covariables,  $lon$  es longitud,  $lat$  es latitud de los puntos muestreados,  $f_1$  y  $f_2$  son funciones de suavizamiento,  $g$  la función liga igual a la identidad,  $Z$  es la matriz diseño de los efectos aleatorios y  $\gamma \sim N(0, \psi)$  es el vector que contiene los efectos aleatorios, como efecto aleatorio se consideró el DR.

Los resultados del ajuste de cada modelo se presentan en el Cuadro 12. De manera general se observa que independientemente de la distribución asociada a cada metal (Normal, Gamma e Inversa Gaussiana), el valor del AIC y del ECM son menores en los modelos no lineales. Al analizar los valores de ECM, para cada metal, se aprecia que los modelos no lineales (GAM) presentaron el mejor ajuste para cada metal.

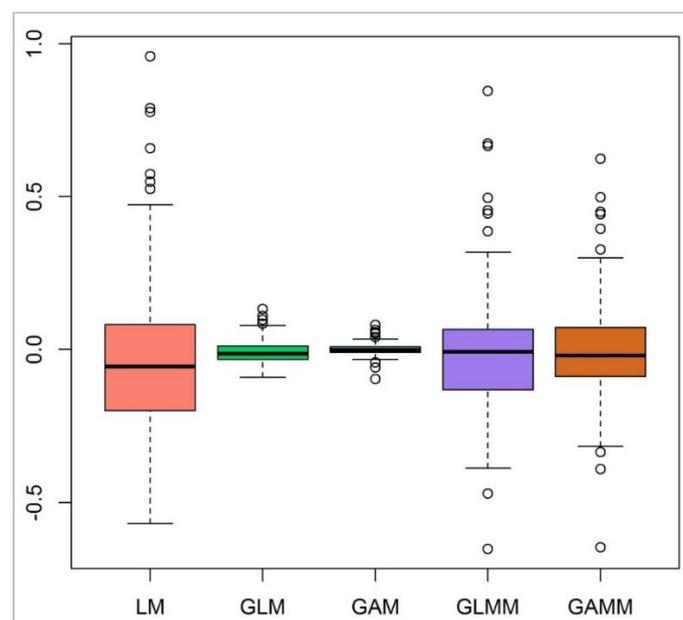
**Cuadro 12. Comparación de modelos ajustados para concentración de metales.**

<b>Metal</b>	<b>Dist.</b>	<b>Modelo</b>	<b>AIC</b>	<b>ECM</b>	<b>Ind. Mor.</b>	<b>p-value</b>
Cadmio	Normal	LM	40.141	0.091	0.342	1.026E-07
		Gamma	<b>30.859</b>	0.007	0.340	1.279E-07
	Inv. Gau.	GLMM	NA	0.068	0.251	7.162E-05
		GAM	<b>-23.582</b>	0.003	<b>0.024</b>	5.635E-01
		GAMM	6.030	0.037	0.121	4.098E-02
		GLM	26.623	0.002	0.339	1.443E-07
		GLMM	NA	0.068	0.253	6.269E-05
		GAM	-28.951	<b>0.001</b>	<b>0.026</b>	5.465E-01
		GAMM	2.310	0.040	0.156	1.035E-02
		GAMM	2.310	0.040	0.156	1.035E-02
Cobre	Normal	LM	447.927	26.370	0.100	7.419E-02
		Gamma	379.569	0.229	0.142	1.994E-02
	Inv. Gau.	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
		GLM	<b>360.911</b>	0.028	0.175	4.931E-03
		GLMM	NA	27.909	0.127	2.562E-02
		GAM	<b>339.061</b>	<b>0.018</b>	0.049	3.425E-01
		GAMM	378.953	26.372	0.099	7.546E-02
		GAMM	378.953	26.372	0.099	7.546E-02
Cromo	Normal	LM	297.171	3.249	0.150	4.692E-03
		Gamma	229.177	0.108	0.146	9.099E-03
	Inv. Gau.	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
		GLM	<b>203.588</b>	0.023	0.150	8.564E-03
		GLMM	NA	3.255	0.157	3.051E-03
		GAM	<b>201.572</b>	<b>0.020</b>	0.106	5.566E-02
		GAMM	278.751	3.255	0.157	3.051E-03
		GAMM	278.751	3.255	0.157	3.051E-03
Niquel	Normal	LM	275.425	2.402	0.231	2.455E-04
		Gamma	<b>249.460</b>	0.095	0.254	6.959E-05
	Inv. Gau.	GLMM	NA	1.659	<b>0.025</b>	5.529E-01
		GAM	<b>233.424</b>	0.057	0.098	9.502E-02
		GAMM	236.018	1.659	<b>0.025</b>	5.529E-01
		GLM	239.419	0.020	0.263	4.081E-05
		GLMM	NA	1.666	0.032	4.914E-01
		GAM	223.939	<b>0.012</b>	0.103	8.172E-02
		GAMM	244.692	1.666	0.026	5.477E-01
		GAMM	244.692	1.666	0.026	5.477E-01
Plomo	Normal	LM	359.832	7.758	0.320	5.736E-07
		Gamma	328.987	0.163	0.367	1.692E-08
	Inv. Gau.	GLMM	NA	4.243	<b>0.008</b>	7.400E-01
		GAM	277.782	0.058	0.035	4.607E-01
		GAMM	293.255	4.243	<b>0.008</b>	7.400E-01

Metal	Dist.	Modelo	AIC	ECM	Ind. Mor.	p-value	
Zinc	Inv. Gau.	GLM	<b>317.864</b>	0.026	0.382	5.166E-09	
		GLMM	NA	4.281	<b>0.010</b>	7.133E-01	
		GAM	<b>268.903</b>	<b>0.009</b>	0.038	4.387E-01	
		GAMM	314.233	3.311	<b>-0.103</b>	1.739E-01	
	Normal	LM	594.723	202.562	0.233	9.682E-05	
		Gamma	GLM	<b>495.104</b>	0.613	0.319	5.909E-07
			GLMM	NA	153.538	0.139	1.397E-02
			GAM	<b>444.220</b>	<b>0.217</b>	<b>-0.035</b>	7.540E-01
			GAMM	502.633	149.639	0.113	3.884E-02

**Nota:** Dist: Distribución; Ind. Mor.: Índice de Moran de los residuales; Inv. Gau.: Inversa Gaussiana.

Al analizar los 9 Índices de Moran de los residuales más pequeños, 6 de ellos corresponden a modelos mixtos los cuales presentan índices no significativos; esto indica que son capaces de describir la correlación espacial. Tres de los 6 valores mencionados corresponden a modelos GAMM y los otros tres a modelos GAM. Demostrándose en esta aplicación, los modelos no lineales son mejores los lineales. Aunque los modelos GAM no están diseñados para datos correlacionados las funciones de suavizamiento pueden describir dicho comportamiento. La Figura 10 muestra que el mejor modelo para cadmio fue el GAM, ajustado con la distribución Inversa Gaussiana de acuerdo al comportamiento de los residuales. Es evidente que sería erróneo considerar una distribución Normal.



**Figura 9. Gráfica boxplot para residuales del ajuste de modelos para [Cd] con distribución Normal en LM e Inversa Gaussiana para el resto de los modelos.**

### 6.2.3 Predicción y análisis de riesgo

Con el Sistema Generador de Modelos Altimétricos (SIGMA) (Pedraza, 2000), se elaboró una malla de puntos del cuadrante de la zona de estudio, posteriormente con el módulo ArcMap de ArcGis (ESRI, 2014) se realizó un recorte de dichos puntos con la capa vectorial de parcelas de los DR, a fin de obtener los puntos dentro de cada DR, obteniendo 112,728 puntos. Esta malla de puntos fue utilizada para la elaboración de mapas de predicción.

En la Figura 10 se presentan los mapas de predicción bajo los diferentes modelos, correspondientes a la concentración de cadmio considerando una distribución Inversa Gaussiana. En el Anexo 2 se presentan los mapas de predicción para cada uno de los metales.

En la NOM-001-SEMARNAT-1996, se consideran los siguientes límites de concentración en el suelo para que éstos puedan tener uso agrícola (Cuadro 13).

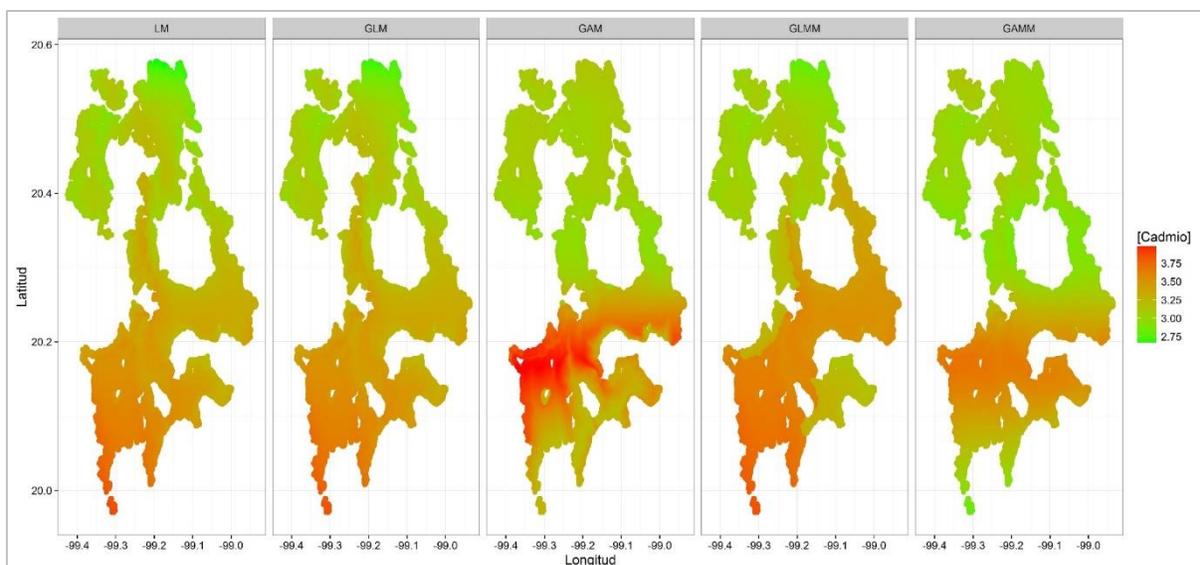
**Cuadro 13. Límites máximos permisibles de metales pesados.**

<b>Metal</b>	<b>Límites máximos permisibles (mg L<sup>-1</sup>)</b>	<b>Límites máximos permisibles (mg L<sup>-1</sup>)*</b>
Cadmio	0.05	3.05
Cobre	4.00	7.00
Cromo	0.50	3.50
Níquel	2.00	5.00
Plomo	5.00	8.00
Zinc	10.00	13.00

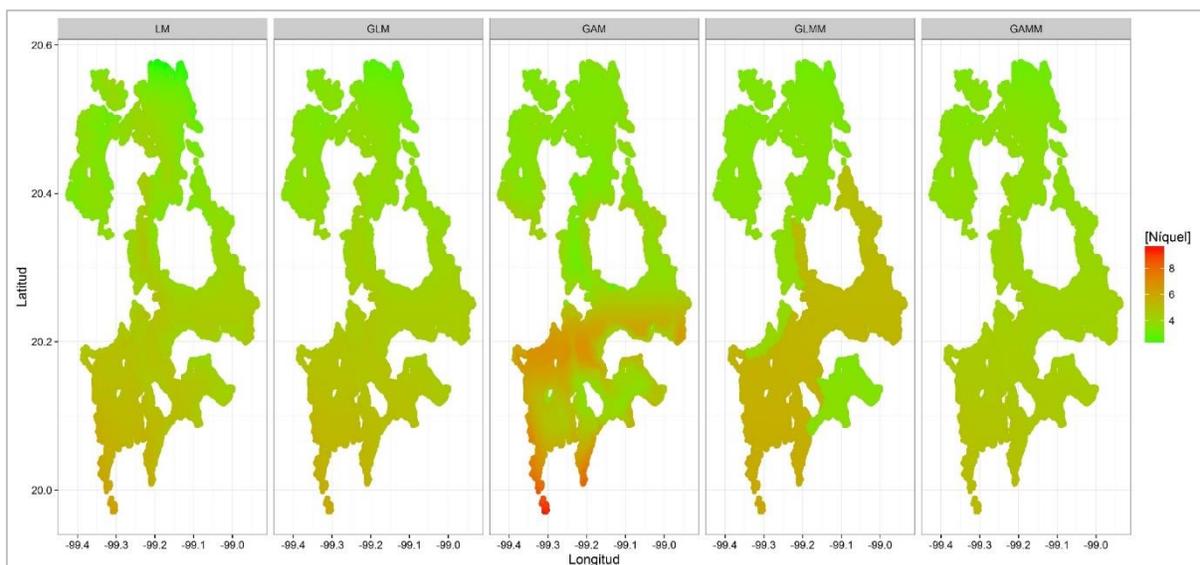
**Nota:** \* Límites +3. **Fuente:** NOM-001-SEMARNAT-1996.

De acuerdo a los límites permisibles, en el Cuadro 14 se muestra el porcentaje de puntos predichos que sobrepasan el límite permisible. Como se puede apreciar, en algunos metales (Ni y Pb) es notorio el cambio que se presenta de acuerdo al tipo de modelo a implementar. Por lo que los mapas de predicción a implementar dependen directamente del tipo de modelo ajustado, para poder identificar adecuadamente las zonas más contaminadas.

Analizando los mapas de predicción para Cd (Figura 10) y Ni (Figura 11), se aprecia que el modelo GAM enfatiza con mayor peso zonas de riesgo (color rojo) que el resto de los modelos. La Figura 12 muestra la probabilidad de que el valor predicho sobrepase el límite permisible para cada uno de los metales, dadas las distribuciones antes mencionadas. Se observa que las concentraciones de Cd y Cr presentan mayor probabilidad (casi 1) de exceder los límites permisibles observándose colores rojos en todo el mapa.



**Figura 10. Mapas de predicción para [Cd] con distribución Normal para LM e Inversa Gaussiana para el resto de los modelos.**

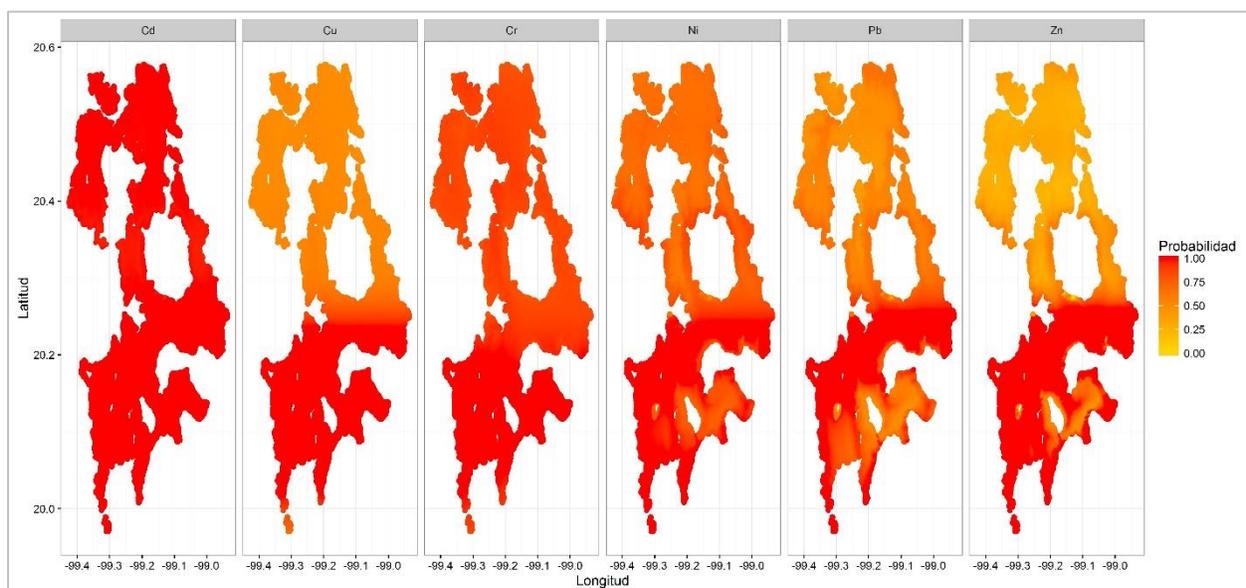


**Figura 11. Mapas de predicción para [Ni] con distribución Normal para LM y Gamma para el resto de los modelos.**

**Cuadro 14. Porcentaje de puntos que exceden los límites permisibles.**

Distribución	Modelo	Metal					
		Cd	Cu	Cr	Ni	Pb	Zn
Normal	LM	86.71	58.13	55.64	18.01	1.61	48.43
Gamma	GLM	87.56	58.47	56.17	15.14	0.54	50.62
	GAM	83.98	45.82	35.59	30.19	28.54	42.42
	GLMM	81.15	58.47	56.17	50.27	44.88	53.24
Inv. Gau.	GAMM	63.97	49.19	56.17	0.08	0.00	28.78
	GLM	88.02	57.28	56.17	14.16	0.41	
	GAM	84.22	44.58	33.68	30.04	28.59	
	GLMM	81.75	57.28	56.17	51.81	47.71	
	GAMM	60.82	46.09	56.17	0.46	0.03	

Nota: Inv. Gau.: Inversa Gaussiana.



**Figura 12. Probabilidad de que el valor predicho sobrepase el límite permisible.**

En el Anexo 3 se presentan los efectos conocidos por altas concentraciones de metales pesados en las plantas, animales y seres humanos. El Anexo 4 menciona algunos métodos de tratamiento de aguas residuales y recuperación de suelos.

### 6.2.4 Índice de Moran y tamaño de muestra óptimo

El Índice de Moran que se obtuvo para cada uno de los metales implementando las cinco diferentes tipos de matrices de pesos se presentan a continuación.

**Cuadro 15. Índices de Moran empleando diferente matriz de pesos.**

Metal	Matriz W		Matriz B		Matriz C		Matriz U		Matriz S	
	I. M.	p-value								
Cd	0.520	4.21E-15	0.518	2.06E-15	0.518	2.06E-15	0.518	2.06E-15	0.520	2.15E-15
Cu	0.358	1.46E-08	0.347	2.49E-08	0.347	2.49E-08	0.347	2.49E-08	0.353	1.63E-08
Cr	0.234	2.08E-05	0.232	1.96E-05	0.232	1.96E-05	0.232	1.96E-05	0.234	1.80E-05
Ni	0.349	8.16E-08	0.332	2.03E-07	0.332	2.03E-07	0.332	2.03E-07	0.341	1.19E-07
Pb	0.414	2.50E-10	0.395	8.27E-10	0.395	8.27E-10	0.395	8.27E-10	0.405	3.91E-10
Zn	0.356	1.15E-08	0.336	4.35E-08	0.336	4.35E-08	0.336	4.35E-08	0.347	1.97E-08

Bajo el siguiente juego de hipótesis:  $H_0$ : No existe autocorrelación vs.  $H_a$ : Existe autocorrelación, en todos los casos se rechaza  $H_0$  con un intervalo de confianza de 95%, es decir, todos los metales presentan un nivel de autocorrelación mayor a cero, lo que indica una autocorrelación positiva. Como se puede apreciar, el uso de diferentes tipos de matriz no afecta significativamente el valor del Índice de Moran.

De acuerdo a Griff (2005), si existe una variable georreferenciada la cual tiene distribución normal o aproximadamente normal y  $\rho$  es la estimación de un parámetro de autocorrelación, el tamaño óptimo de la muestra ( $n^*$ ) esta definido por:

$$n^* = n \left[ 1 - \frac{1}{1 - e^{-1.92369}} \frac{n-1}{n} \left( 1 - e^{-2.12373*\rho + 0.20024*\sqrt{\rho}} \right) \right] \quad (48)$$

Donde  $n$  es el tamaño de muestra inicial.

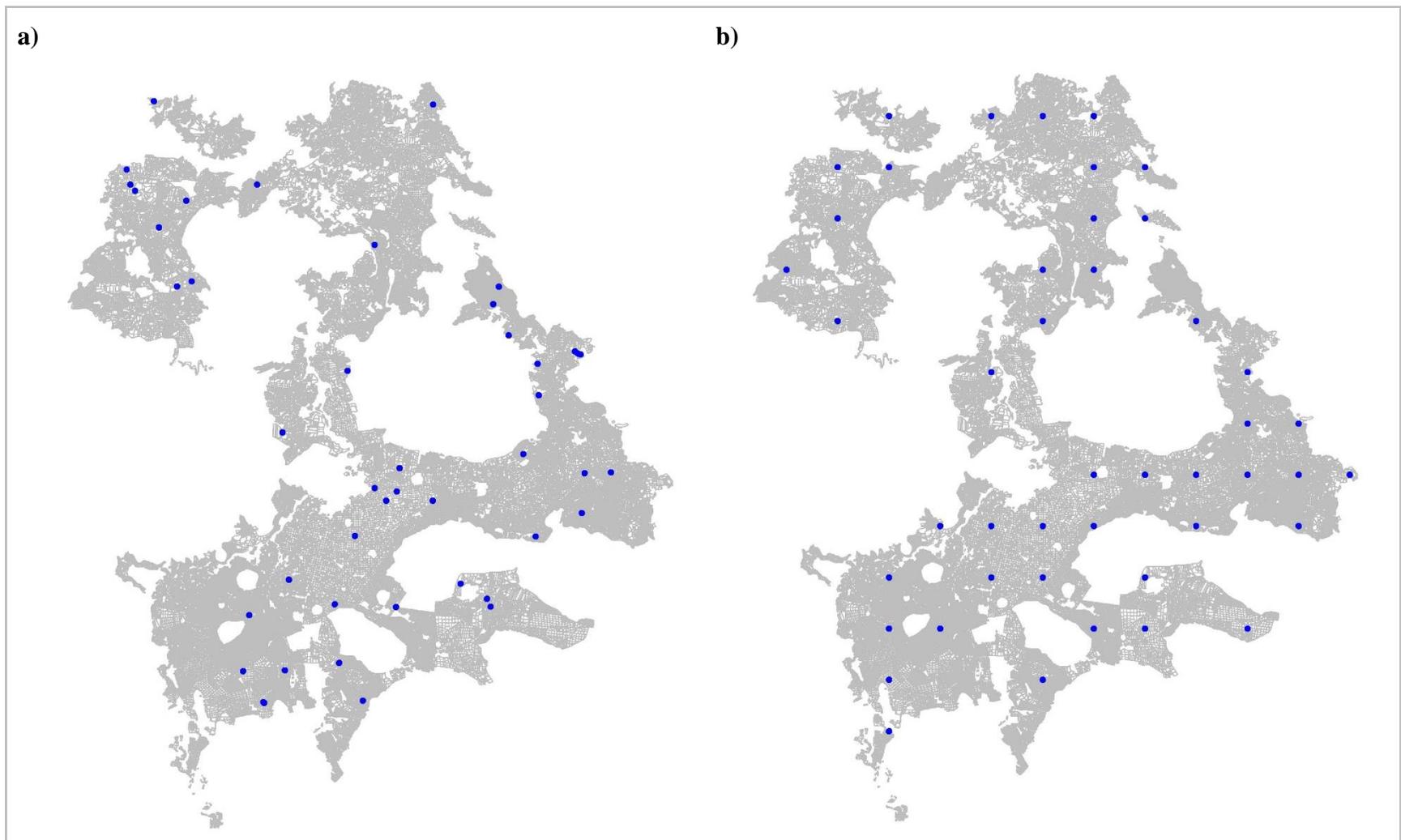
Para la estimación del tamaño de muestra óptimo, a los datos de la concentración de metales se les aplicó una transformación Box-Cox con el fin de aproximarlos a la distribución normal. Para el valor de  $\rho$ , se consideró el valor del Índice de Moran utilizando los cinco tipos de matriz (Página 16).

En el Cuadro 16 se indica el valor de lambda que fue empleado para la normalización y el tamaño de muestra óptimo recomendado para estudios posteriores, calculado a partir del Índice de Moran bajo diferentes tipos de matriz de pesos.

**Cuadro 16. Tamaño de muestra empleando diferente matriz de pesos.**

<b>Metal</b>	<b>Lambda</b>	<b>Matriz W</b>	<b>Matriz B</b>	<b>Matriz C</b>	<b>Matriz U</b>	<b>Matriz S</b>
Cadmio	-2.205	21	21	21	21	21
Cobre	-0.280	33	34	34	34	33
Cromo	-3.468	45	45	45	45	45
Níquel	-0.742	33	35	35	35	34
Plomo	-0.289	28	30	30	30	29
Zinc	-0.297	33	35	35	35	34

Dado que el Índice de Moran es diferente para cada uno de los metales y en algunos casos para diferente tipo de matriz. De los tamaños de muestra se eligió el máximo, y con dicho valor se identificaron los puntos de muestreo de suelos, en la zona de estudio, bajo dos condiciones: muestreo al azar y muestreo sistemático (Figura 13).



**Figura 13. Muestreo de suelo a) aleatorio y b) sistemático para el DR 003, DR 100 y DR 112.**

## VII. CONCLUSIONES Y RECOMENDACIONES

Cuando la variable respuesta simulada es generada a través de la matriz de pesos de tipo U, la estimación de los parámetros  $\beta_1$  y  $\beta_2$  es más robusta al variar el nivel de autocorrelación, obteniendo valores más cercanos a los respectivos valores previamente definidos para los parámetros.

Los mejores resultados se obtuvieron con los modelos mixtos y los no lineales debido a la integración de efectos aleatorios y funciones de suavizamiento respectivamente.

La incorporación de nuevas condiciones como las distribuciones asimétricas, funciones de suavizamiento y efectos aleatorios describe con mayor confiabilidad estadística el comportamiento espacial de los metales pesados.

La zona más contaminada corresponde al DR 003, por lo que se recomienda implementar técnicas de recuperación de suelos. La mayor contaminación se presenta por cadmio y en menor proporción por plomo.

La implementación de medidas de autocorrelación y matrices de pesos de vecindad en el Índice de Moran permite optimizar recursos (económicos), sin perder información estadística importante.

Se recomienda la implementación de la matriz U en lugar de las convencionales W y B, con objeto de maximizar la robustez en la estimación de los parámetros en el ajuste de modelos.

## VIII. BIBLIOGRAFÍA CONSULTADA

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.

Anselin, L. and Bera, A.K. 1998. *Spatial dependence in linear regression models with an introduction to spatial econometrics*. Statistics Textbooks and Monographs, 155: 237-290.

Araúz, D; García, A. L; Rodríguez, F. y Zárata, M. F. 2013. *Nivel de Contaminación y Distribución Espacial de Metales Pesados en Sedimentos Superficiales de Bahía Damas, Isla Coiba*. RIDTEC, Vol. 9, No. 2: 14-23.

Ayres, R. M. y Mara, D. D. 1997. *Análisis de las aguas residuales para su uso en la agricultura: manual de técnicas parasitológicas y bacteriológicas de laboratorio*. Ginebra: Organización Mundial de la Salud.

Badiella, L. 2011. *Modelos lineales generalizados mixtos, algunos casos prácticos*. Memorias del X Congreso Galego de Estatística e Investigación de Operacións. España.

Basso, M. C; Cerrella, E. G. y Cukierman, A. L. 2002. *Empleo de algas marinas para la biosorción de metales pesados de aguas contaminadas*. Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales. Avances en Energías Renovables y Medio Ambiente, Vol. 6, No. 1. Argentina.

Bolker, B. M; Brooks, M. E; Clark, C. J; Geange, S. W; Poulsen, J. R; Stevens, M. H. y White, J. S. 2009. *Generalized linear mixed models: a practical guide for ecology and evolution*. Trends in Ecology and Evolution, Vol. 24, No.3.

Borda, H. R. A; Iral, P. R. y Roy, C. K. 2012. *Application of generalized linear mixed models in the modeling of georeferenced count data in the municipalities in the department of Antioquia*. Cuadernos Latinoamericanos de Administración, Vol. VIII, No. 15: 69-76.

Breslow, N. E. y Clayton, D. G. 1993. *Approximate inference in generalized linear mixed models*. Journal of the American Statistical Association, Vol. 88: 9-25.

Brogniez, D; Ballabio, C; Stevens, A; Jones, R. J. A. y Montanarella, L. 2015. *A map of the topsoil organic carbon content of Europe generated by a generalized additive model*. European Journal of Soil Science, Vol. 66: 121-134.

Carl, G; Dormann, C. F. y Kühn, I. 2008. *A wavelet-based method to remove spatial autocorrelation in the analysis of species distributional data*. Web Ecology, Vol. 8: 22-29.

Chicas, S. R. A; Vanegas, C. E. A. y García, A. N. 2014. *Determinación indirecta de la capacidad de retención de humedad en suelos de la subcuenca del río Torjá, Chiquimula, Guatemala*. Revista Ciencias Técnicas Agropecuarias, Vol. 23, No. 1: 41-46.

Ciannelli, L; Bartolino, V. y Chan, K. S. 2012. *Non-additive and non-stationary properties in the spatial distribution of a large marine fish population*. Proceedings of the Royal Society, Vol. 279: 3635-3642.

Cortés, D. D; Pérez, B. J. y Camacho, T. J. 2013. *Relación espacial entre la conductividad eléctrica y algunas propiedades químicas del suelo*. Universidad Nacional de Colombia, Facultad de Ingeniería, Programa de Ingeniería Agrícola. Revista U.D.C.A Actualidad & Divulgación Científica, Vol. 16, No. 2: 401-408.

De Oliveira, V. 2012. *Bayesian analysis of conditional autoregressive models*. The Institute of Statistical Mathematics, Vol. 64:107-133.

Dormann, C. F; McPherson, J. M; Araújo, M. B; Bivand, R; Bolliger, J; Carl, G; Davies, R. G; Hirzel, A; Jetz, W; Kissling, W. D; Kühn, I; Ohlemüller, R; Peres, N. P. R; Reineking, B; Schröder, B; Schurr, F. M. y Wilson, R. 2007. *Methods to account for spatial autocorrelation in the analysis of species distributional data: a review*. Ecography, Vol. 30: 609-628.

Durbán, M. 2014. *Modelos Aditivos Generalizados con P-splines*. Universidad Carlos III de Madrid. España. Pp. 77.

Griffith, D. A. 2005. *Effective Geographic Sample Size in the Presence of Spatial Autocorrelation*. Annals of the Association of American Geographers, Vol. 95, No. 4: 740-760.

González, M. R; Juárez, L. J. F; Aceves, N. L. A; Rivera, H. B. y Guerrero, P. A. 2015. *Zonificación edafoclimática para el cultivo de *Jatropha curcas* L., en Tabasco, México*. Universidad Nacional Autónoma de México. Investigaciones Geográficas. Boletín del Instituto de Geografía, No. 86: 25-37.

Hastie, T. y Tibshirani, R. 1986. *Generalized additive models (with discussion)*. Statistical Science Vol. 1: 297-318.

- Ibarra, C. D; Ruiz, C. J. A; González, E. D. R. y Flores, G. J. G. 2009. *Distribución espacial del pH de los suelos agrícolas de Zapopan, Jalisco, México*. Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias. Agricultura Técnica en México, Vol. 35, No. 3: 267-276.
- Keitt, T. H; Bjørnstad, O. N; Dixon, P. M. y Citron P. S. 2002. *Accounting for spatial pattern when modeling organism-environment interactions*. Ecography, Vol. 25: 616-625.
- Kleinschmidt, I; Sharp, B. L; Clarke, G. P. Y; Curtis, B. y Fraser, C. 2001. *Use of Generalized Linear Mixed Models in the Spatial Analysis of Small-Area Malaria Incidence Rates in KwaZulu Natal, South Africa*. American Journal of Epidemiology, Vol. 153, No. 12: 1213-1221.
- Lin, X. y Zhang, D. 1999. *Inference in generalized additive mixed models using smoothing splines*. Journal of the Royal Statistical Society, Serie B, Vol. 61: 381-400.
- Liu, H. 2008. *Generalized Additive Model*. University of Minnesota Duluth. Department of Mathematics and Statistics.
- Ma, W; Sun, X; Song, Y; Tao, F; Feng, W; He, Y; Zhao, N. y Yuan, Z. 2013. *Applied Mixed Generalized Additive Model to Assess the Effect of Temperature on the Incidence of Bacillary Dysentery and Its Forecast*. PLoS One, Vol. 8, No. 4: e62122.
- Mamouridis, V. 2011. *Additive Mixed Models applied to the study of red shrimp landings: comparison between frequentist and Bayesian perspectives*. Universidade da Coruña. Departamento de Matemáticas. España.
- Marchetti, A; Piccini, C; Francaviglia, R. y Mabit, L. 2012. *Spatial distribution of soil organic matter using geostatistics: A key indicator to assess soil degradation status in central Italy*. Pedosphere, Vol. 22, No. 2: 230-242.
- Matheron, G. 1962. *Traite de Geostatistique Apliquee*. Tomo I. Memoires de Investigación Geológica y Minera. Editorial Technip. Paris, Francia.
- McCullagh, P. y Nelder, J. A. 1989. *Generalized Linear Models*. Editorial Chapman & Hall. Segunda Edición. Londres, Inglaterra.
- McCulloch, C. E. 1997. *Maximum likelihood algorithms for generalized linear mixed models*. Journal of The American Statistical Association, Vol. 92, No. 437: 162-170.

- McCulloch, C. E. y Searle, S. R. 2001. *Generalized, Linear and Mixed Models*. Editorial John Wiley & Sons.
- Mendoza, H. 1981. *An evaluation of the use of Mexico city Wastewater on the irrigation of Crops*. Water Reuse Symposium II. Washungton D. C.
- Meyer, P. 1992. *Probabilidad y aplicaciones estadísticas*. Washington Slate University. Departamento de Matemáticas.
- Moran, P. A. P. 1948. *The Interpretation of Statistical Maps*. Journal of the Royal Statistical Society, Serie B, Vol. 10, No. 2: 243-251.
- Morton, B. O. 2006. *Contenido de metales pesados en suelos superficiales de la Ciudad de México*. Universidad Nacional Autónoma de México. Instituto de Geofísica. Revista Especializada en Ciencias Químico-Biológicas. Nota Científica, Vol. 9, No. 1:45-47.
- Mulligan, C. N. 2001. *An overview of in situ bioremediation processes*. Proceedings of the 29th Annual Conference of the Canadian Society for Civil Engineering. British Columbia, Canada.
- Navarro, A. J. P; Aguilar, A. I. y López, M. J. R. 2007. *Aspectos bioquímicos y genéticos de la tolerancia y acumulación de metales pesados en plantas*. Universidad Politécnica de Valencia. Escuela Superior de Ingenieros Agrónomos. Departamento de Biología Vegetal. Ecosistemas, Vol. 16, No. 2: 10-25.
- Nelder, J. A. y Wedderburn, R. W. M. 1972. *Generalized linear models*. Journal of the Royal Statistical Society, Series A, Vol. 135, No. 3: 370-384.
- Norma Oficial Mexicana NOM-001-SEMARNAT-1996, que establece los límites máximos permisibles de contaminantes en las descargas de aguas residuales en aguas y bienes nacionales. Diario Oficial de la Federación, 6 de enero de 1997.
- Ortiz, B. I; Sanz, G. J; Dorado, V. M. y Villar, F. S. 2007. *Técnicas de recuperación de suelos contaminados*. Universidad de Alcalá, España.
- Oviedo, M; Domínguez, A. y Muñoz, M. P. 2015. *Estimate of influenza cases using generalized linear, additive and mixed models*. Human Vaccines & Immunotherapeutics, Vol. 11, No. 1: 298-301.

- Polansky, L. y Robbins, M. M. 2013. *Generalized additive mixed models for disentangling long-term trends, local anomalies, and seasonality in fruit tree phenology*. Ecology and Evolution, Vol. 3, No. 9: 3141-3151.
- Pedraza, O. F. 2000. *SIGMA. Sistema Generador de Modelos Altimétricos*. Colegio de Postgraduados. Campus Montecillo. México.
- R Core Team. 2016. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Radhakrishn, R. y Toutenburg, H. 1999. *Linear Models; Least Squares and Alternatives*. Editorial Springer. Segunda Edición. New York, USA.
- Reyes, T. E D; Cerino, C. F. J. y Suárez, H. M. A. 2006. *Remoción de metales pesados con carbón activado como soporte de biomasa*. Universidad Autónoma de Nuevo León. Facultad de Ciencias Químicas. Ingeniería Química. Ingenierías, Vol. 9, No. 31.
- Rodríguez, G. 2007. *Lecture Notes on Generalized Linear Models, Appendix B Generalized Linear Model Theory*. [<http://data.princeton.edu/wws509/notes/>: 30 de enero de 2016].
- Samper, F.J. y Carrera, J. 1990. *Geoestadística. Aplicaciones a la Hidrogeología Subterránea*. Segunda Edición. Centro Internacional de Métodos Numéricos en Ingeniería. Barcelona, España.
- SAS Institute Inc. 2008. *SAS/STAT® 9.2 User's Guide The GLIMMIX Procedure (Book Excerpt)*. Cary, NC: SAS Institute Inc.
- Seoane, P. J. 2005. *Modelos predictivos de distribución de aves esteparias en las Islas Orientales del Archipiélago Canario*. Universidad Autónoma de Madrid. Departamento Interuniversitario de Ecología. España.
- Servicio Geológico Mexicano (SGM). 2011. *Panorama minero del estado de Hidalgo*. Gobierno Federal. México.
- Soto, R. E; Lozano, R. T; Barbarín, C. J. M; Alcalá, R. M. 2004. *Remoción de metales pesados en aguas residuales mediante agentes químicos*. Universidad Autónoma de Nuevo León. Departamento de Ingeniería Química de la Facultad de Ciencias Químicas. Ingenierías, Vol. VII, No. 23: 46-51.

- Sun, B; Zhou, S. y Zhao, Q. 2003. *Evaluation of spatial and temporal changes of soil quality based on geostatistical analysis in the hill region of subtropical China*. Geoderma, Vol. 115: 85-99.
- Tiefelsdorf, M; Griffith, D. A. y Boots, B. 1999. *A variance-stabilizing coding scheme for spatial link matrices*. Environment and Planning A, Vol. 31: 165-180.
- Torabi, M. 2015. *Likelihood Inference for Spatial Generalized Linear Mixed Models*. Communications in Statistics - Simulation and Computation, Vol. 44, No. 7: 1692-1701.
- Vaca, R. A. y Golicher, D. J. 2016. *Integración del balance hídrico en la modelación de la distribución de especies de árboles mexicanos*. Botanical Sciences, Vol. 94, No. 1: 1-18.
- Venables, W. N. y Dichmont, C. M. 2004. *GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research*. Fisheries Research, Vol. 70: 319-337.
- Venables, W. N. y Ripley, B.D; 2002. *Modern Applied Statistics with S*. Editorial Springer-Verlag. Cuarta Edición. New York, USA.
- Viton, P. A. 2010. *Notes on Spatian Econometric Models*. The Ohio State University.
- Wolfinger, R. y O'Connell, M. 1993. *Generalized Linear Mixed Models: A Pseudo-likelihood Approach*. Journal of Statistical Computation and Simulation, Vol. 4: 233-243.
- Wood, S.N. 2006. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC Press.
- Wu, W; Hall, C. y Zhang, L. 2006. *Predicting the temporal and spatial probability of orographic cloud cover in the Luquillo Experimental Forest in Puerto Rico using generalized linear (mixed) models*. Ecological Modelling, Vol. 192: 473-498.
- Zwillinger, D. y Kokoska, S. 2000. *CRC standard probability and statistics tables and formulae*. Editorial Chapman & Hall.

## IX. ANEXOS

### Anexo 1. Tratamientos para generar las variables de respuesta.

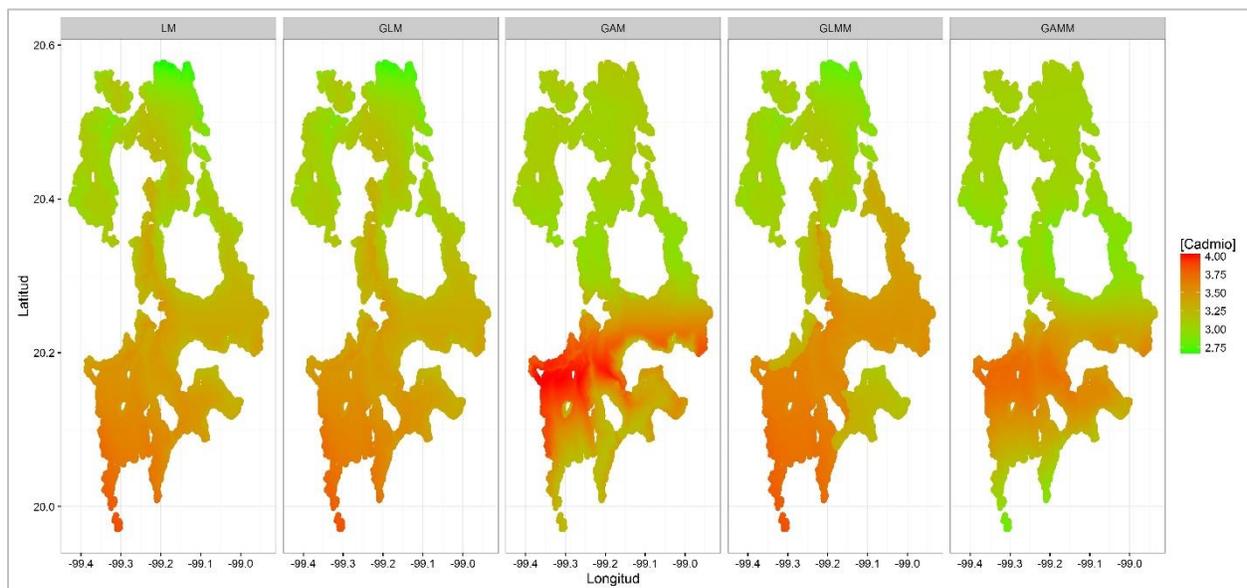
**Cuadro 17. Combinaciones para generar las variables de respuesta.**

Com	Distribución de y	Tipo de error	Tipo de W	Nivel de $\lambda, \rho$	Com	Distribución de y	Tipo de error	Tipo de W	Nivel de $\lambda, \rho$
1	Normal	SAR	W	0.00	48	Normal	SMA	S	0.50
2	Normal	SAR	W	0.20	49	Normal	SMA	S	0.70
3	Normal	SAR	W	0.50	50	Normal	SMA	S	0.90
4	Normal	SAR	W	0.70	51	Gamma	SAR	W	0.00
5	Normal	SAR	W	0.90	52	Gamma	SAR	W	0.20
6	Normal	SAR	B	0.00	53	Gamma	SAR	W	0.50
7	Normal	SAR	B	0.20	54	Gamma	SAR	W	0.70
8	Normal	SAR	B	0.50	55	Gamma	SAR	W	0.90
9	Normal	SAR	B	0.70	56	Gamma	SAR	B	0.00
10	Normal	SAR	B	0.90	57	Gamma	SAR	B	0.20
11	Normal	SAR	C	0.00	58	Gamma	SAR	B	0.50
12	Normal	SAR	C	0.20	59	Gamma	SAR	B	0.70
13	Normal	SAR	C	0.50	60	Gamma	SAR	B	0.90
14	Normal	SAR	C	0.70	61	Gamma	SAR	C	0.00
15	Normal	SAR	C	0.90	62	Gamma	SAR	C	0.20
16	Normal	SAR	U	0.00	63	Gamma	SAR	C	0.50
17	Normal	SAR	U	0.20	64	Gamma	SAR	C	0.70
18	Normal	SAR	U	0.50	65	Gamma	SAR	C	0.90
19	Normal	SAR	U	0.70	66	Gamma	SAR	U	0.00
20	Normal	SAR	U	0.90	67	Gamma	SAR	U	0.20
21	Normal	SAR	S	0.00	68	Gamma	SAR	U	0.50
22	Normal	SAR	S	0.20	69	Gamma	SAR	U	0.70
23	Normal	SAR	S	0.50	70	Gamma	SAR	U	0.90
24	Normal	SAR	S	0.70	71	Gamma	SAR	S	0.00
25	Normal	SAR	S	0.90	72	Gamma	SAR	S	0.20
26	Normal	SMA	W	0.00	73	Gamma	SAR	S	0.50
27	Normal	SMA	W	0.20	74	Gamma	SAR	S	0.70
28	Normal	SMA	W	0.50	75	Gamma	SAR	S	0.90
29	Normal	SMA	W	0.70	76	Gamma	SMA	W	0.00
30	Normal	SMA	W	0.90	77	Gamma	SMA	W	0.20
31	Normal	SMA	B	0.00	78	Gamma	SMA	W	0.50
32	Normal	SMA	B	0.20	79	Gamma	SMA	W	0.70
33	Normal	SMA	B	0.50	80	Gamma	SMA	W	0.90
34	Normal	SMA	B	0.70	81	Gamma	SMA	B	0.00
35	Normal	SMA	B	0.90	82	Gamma	SMA	B	0.20
36	Normal	SMA	C	0.00	83	Gamma	SMA	B	0.50
37	Normal	SMA	C	0.20	84	Gamma	SMA	B	0.70
38	Normal	SMA	C	0.50	85	Gamma	SMA	B	0.90
39	Normal	SMA	C	0.70	86	Gamma	SMA	C	0.00
40	Normal	SMA	C	0.90	87	Gamma	SMA	C	0.20
41	Normal	SMA	U	0.00	88	Gamma	SMA	C	0.50
42	Normal	SMA	U	0.20	89	Gamma	SMA	C	0.70
43	Normal	SMA	U	0.50	90	Gamma	SMA	C	0.90
44	Normal	SMA	U	0.70	91	Gamma	SMA	U	0.00
45	Normal	SMA	U	0.90	92	Gamma	SMA	U	0.20
46	Normal	SMA	S	0.00	93	Gamma	SMA	U	0.50
47	Normal	SMA	S	0.20	94	Gamma	SMA	U	0.70

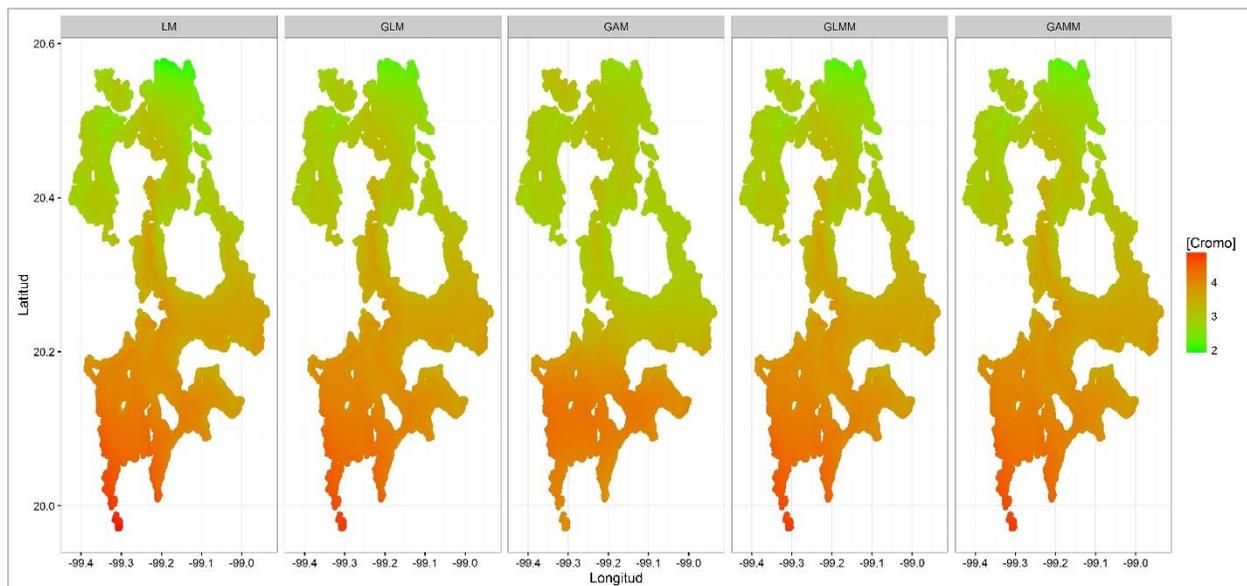
Com	Distribución de $y$	Tipo de error	Tipo de $W$	Nivel de $\lambda, \rho$
95	Gamma	SMA	U	0.90
96	Gamma	SMA	S	0.00
97	Gamma	SMA	S	0.20
98	Gamma	SMA	S	0.50
99	Gamma	SMA	S	0.70
100	Gamma	SMA	S	0.90
101	Inv. Gau.	SAR	W	0.00
102	Inv. Gau.	SAR	W	0.20
103	Inv. Gau.	SAR	W	0.50
104	Inv. Gau.	SAR	W	0.70
105	Inv. Gau.	SAR	W	0.90
106	Inv. Gau.	SAR	B	0.00
107	Inv. Gau.	SAR	B	0.20
108	Inv. Gau.	SAR	B	0.50
109	Inv. Gau.	SAR	B	0.70
110	Inv. Gau.	SAR	B	0.90
111	Inv. Gau.	SAR	C	0.00
112	Inv. Gau.	SAR	C	0.20
113	Inv. Gau.	SAR	C	0.50
114	Inv. Gau.	SAR	C	0.70
115	Inv. Gau.	SAR	C	0.90
116	Inv. Gau.	SAR	U	0.00
117	Inv. Gau.	SAR	U	0.20
118	Inv. Gau.	SAR	U	0.50
119	Inv. Gau.	SAR	U	0.70
120	Inv. Gau.	SAR	U	0.90
121	Inv. Gau.	SAR	S	0.00
122	Inv. Gau.	SAR	S	0.20

Com	Distribución de $y$	Tipo de error	Tipo de $W$	Nivel de $\lambda, \rho$
123	Inv. Gau.	SAR	S	0.50
124	Inv. Gau.	SAR	S	0.70
125	Inv. Gau.	SAR	S	0.90
126	Inv. Gau.	SMA	W	0.00
127	Inv. Gau.	SMA	W	0.20
128	Inv. Gau.	SMA	W	0.50
129	Inv. Gau.	SMA	W	0.70
130	Inv. Gau.	SMA	W	0.90
131	Inv. Gau.	SMA	B	0.00
132	Inv. Gau.	SMA	B	0.20
133	Inv. Gau.	SMA	B	0.50
134	Inv. Gau.	SMA	B	0.70
135	Inv. Gau.	SMA	B	0.90
136	Inv. Gau.	SMA	C	0.00
137	Inv. Gau.	SMA	C	0.20
138	Inv. Gau.	SMA	C	0.50
139	Inv. Gau.	SMA	C	0.70
140	Inv. Gau.	SMA	C	0.90
141	Inv. Gau.	SMA	U	0.00
142	Inv. Gau.	SMA	U	0.20
143	Inv. Gau.	SMA	U	0.50
144	Inv. Gau.	SMA	U	0.70
145	Inv. Gau.	SMA	U	0.90
146	Inv. Gau.	SMA	S	0.00
147	Inv. Gau.	SMA	S	0.20
148	Inv. Gau.	SMA	S	0.50
149	Inv. Gau.	SMA	S	0.70
150	Inv. Gau.	SMA	S	0.90

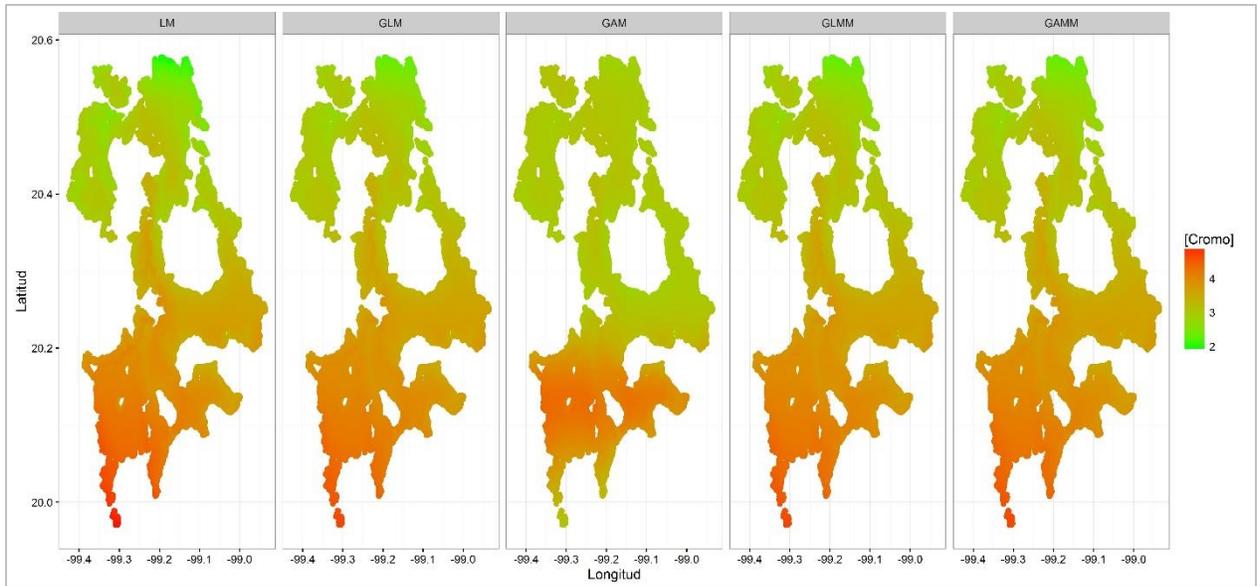
## Anexo 2. Mapas de predicción de metales pesados



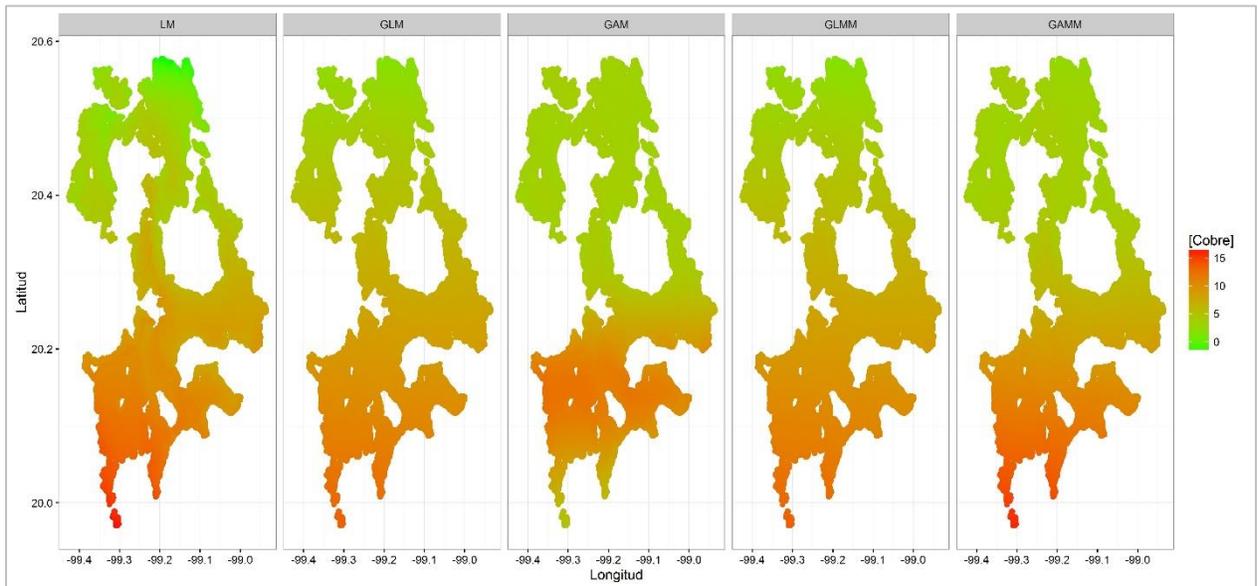
**Figura 14. Mapas de predicción para [Cd] con distribución Normal para LM y Gamma para el resto de los modelos.**



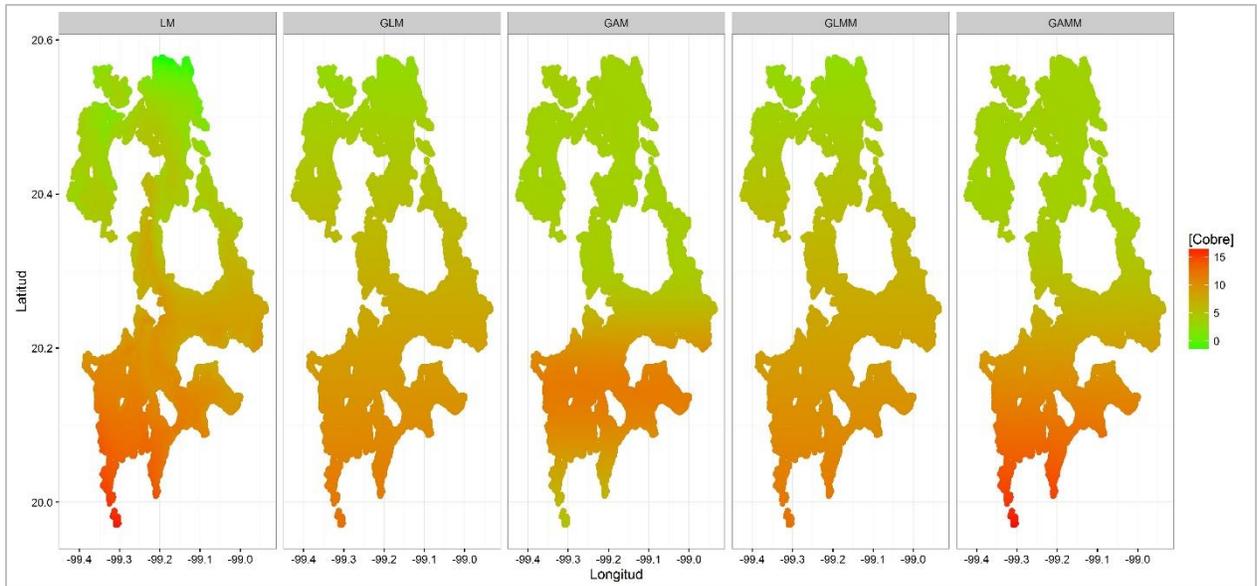
**Figura 15. Mapas de predicción para [Cr] con distribución Normal para LM y Gamma para el resto de los modelos.**



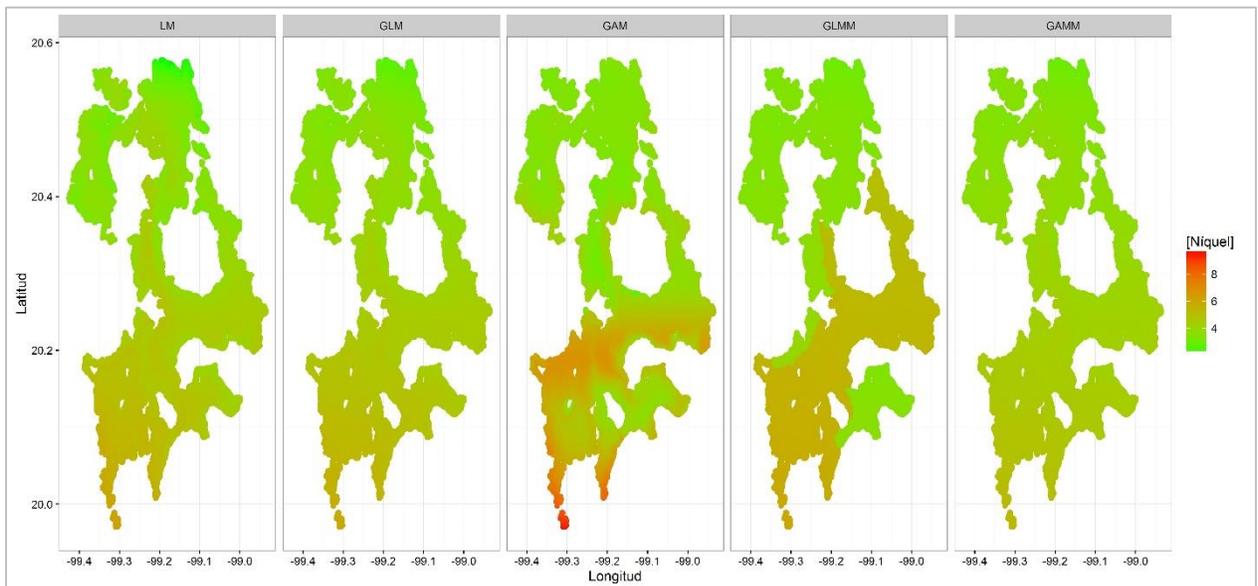
**Figura 16. Mapas de predicción para [Cr] con distribución Normal para LM e Inversa Gaussiana para el resto de los modelos.**



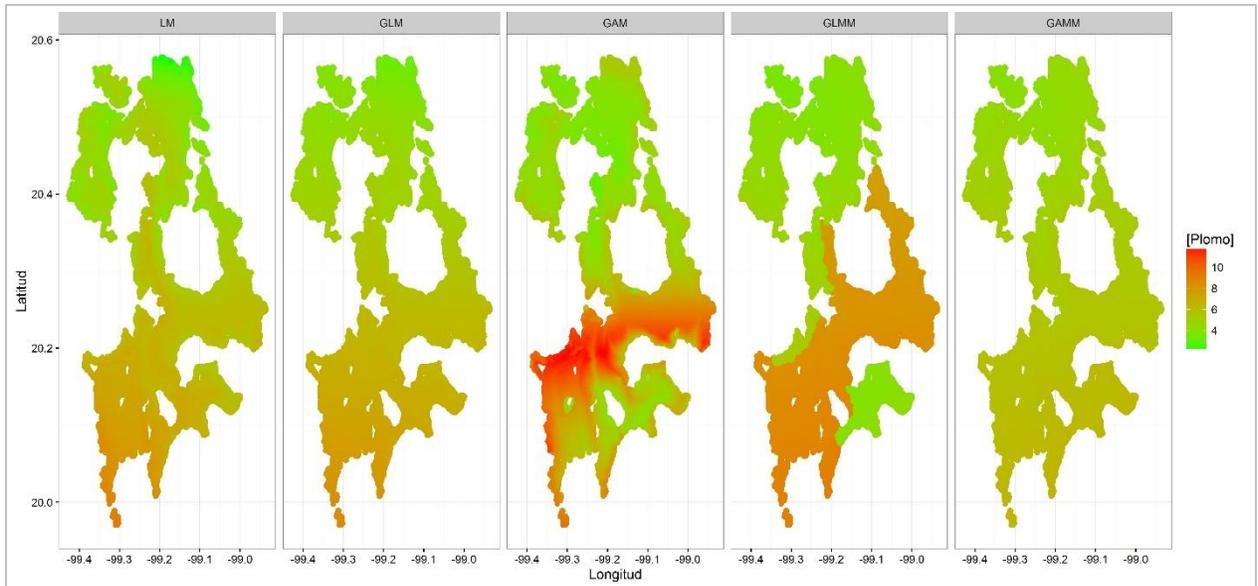
**Figura 17. Mapas de predicción para [Cu] con distribución Normal para LM y Gamma para el resto de los modelos.**



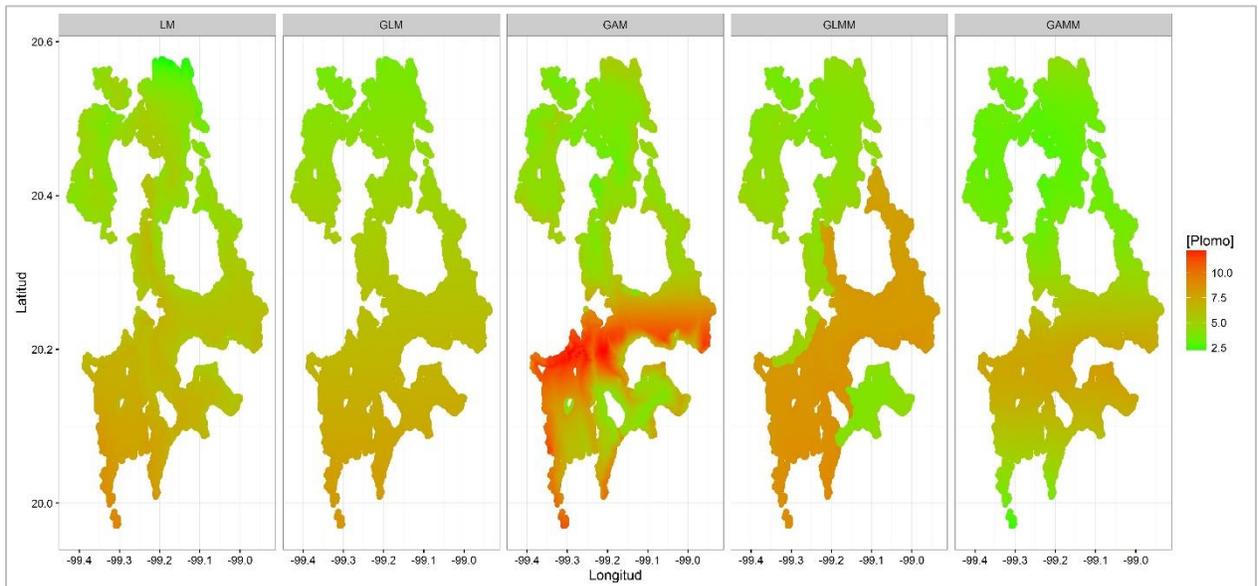
**Figura 18. Mapas de predicción para [Cu] con distribución Normal para LM e Inversa Gaussiana para el resto de los modelos.**



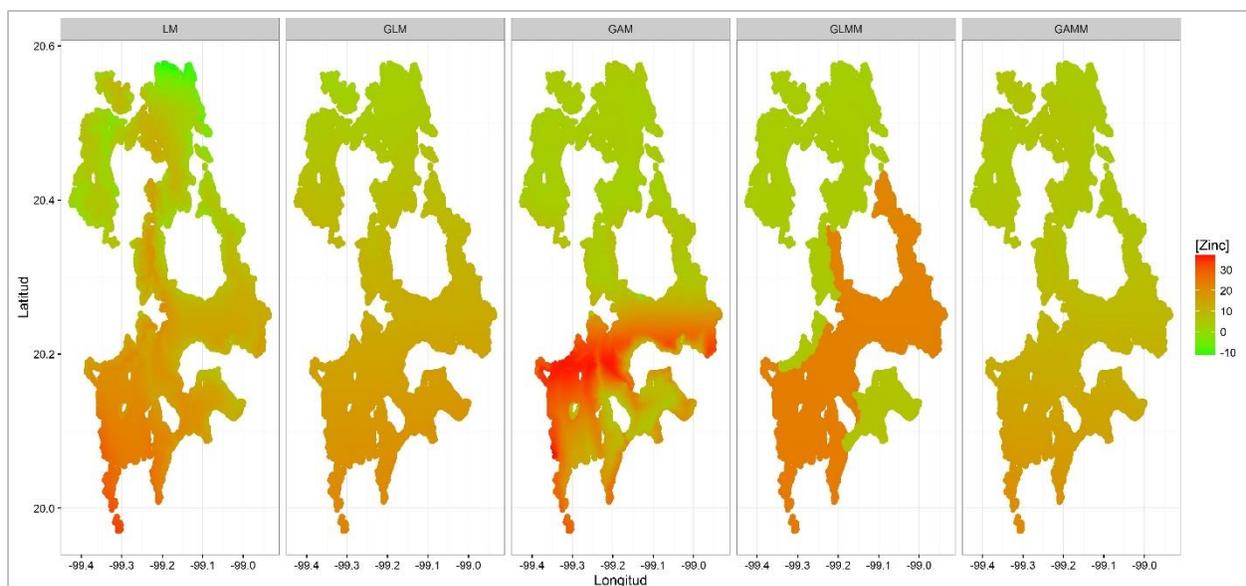
**Figura 19. Mapas de predicción para [Ni] con distribución Normal para LM e Inversa Gaussiana para el resto de los modelos.**



**Figura 20. Mapas de predicción para [Pb] con distribución Normal para LM y Gamma para el resto de los modelos.**



**Figura 21. Mapas de predicción para [Pb] con distribución Normal para LM e Inversa Gaussiana para el resto de los modelos.**



**Figura 22. Mapas de predicción para [Zn] con distribución Normal para LM y Gamma para el resto de los modelos.**

### **Anexo 3. Efectos de altas concentraciones de metales pesados**

Al hablar de contaminación por metales hay que tener en cuenta que más importante que el contenido total de un elemento en el suelo es la forma o especie química bajo la que se encuentra, es decir, su especiación. Así, la forma resultante de dicha especiación va a influir decisivamente en su distribución en el suelo, condicionando su solubilidad, su movilidad en el suelo y las aguas superficiales y subterráneas, su biodisponibilidad y toxicidad y, por tanto, su comportamiento como contaminante potencial (Mulligan *et al.*, 2001).

Los metales pesados presentes en los suelos no se comportan como elementos estáticamente inalterables, sino que siguen patrones de movilidad. La dinámica de los metales pesados en el suelo puede clasificarse en cuatro vías Navarro *et al.* (2007):

- Movilización a las aguas superficiales o subterráneas.
- Transferencia a la atmósfera por volatilización.
- Absorción por las plantas e incorporación a las cadenas tróficas.
- Retención de metales pesados en el suelo de distintas maneras: disueltos o fijados, retenidos por adsorción, complejación y precipitación.

La toxicidad de los metales pesados se puede definir como elevada, tanto para microorganismos como para animales y plantas. Los metales pesados pasan del suelo a las plantas, y de ahí a los

mamíferos. Los efectos conocidos de los metales pesados en las plantas y animales, así como en los seres humanos se detallan en el Cuadro 18 y

Cuadro 19 respectivamente.

**Cuadro 18. Efectos de algunos metales tóxicos en plantas y animales.**

<b>Metal</b>	<b>Efecto</b>
Arsénico	Reducción del crecimiento y alteración de la concentración de Ca, K, P y Mn en la planta.
Cadmio*	Plantas y algunos animales lo absorben y lo concentran dentro de sus tejidos. Bloquea el calcio en la membrana plasmática. Desplaza al calcio y al zinc en determinadas proteínas además de causar estrés oxidativo. Como resultado daño en el ADN y en los lípidos. Inhibición de la fotosíntesis y la transpiración. Inhibición de la síntesis de clorofila. Modifica las concentraciones de Mn y K. Se acumula en los riñones, el hígado y los órganos reproductores.
Cobre*	Desbalance iónico, alteración de la permeabilidad de la membrana celular, reducción del crecimiento e inhibición de la fotosíntesis.
Cromo*	Degradación de la estructura del cloroplasto, inhibición de la fotosíntesis, alteración de las concentraciones de Fe, Ca, K y Mg.
Mercurio	Alteración de la fotosíntesis, inhibición del crecimiento, alteración de la captación de K.
Plomo*	Inhibición de la fotosíntesis, el crecimiento de la acción enzimática. Los animales pueden absorber plomo por inhalación o ingestión, el plomo tiende a acumularse. La anemia es el primer síntoma de envenenamiento crónico producido por el plomo en los animales, dado que interfiere en la síntesis del grupo hemo, reflejándose en síntomas tales como náuseas, vómitos y dolores abdominales. También puede degradar el tejido en el sistema nervioso central. El Pb suele aprovechar la metabolización del Ca para sustituirlo y dañar la célula.
Zinc*	Alteración de la permeabilidad de la membrana celular, inhibición de la fotosíntesis, alteración de las concentraciones de Cu, Fe y Mg.

**Nota:** \* Elementos que se modelaron en este estudio. **Fuente:** Navarro *et al.* (2007).

**Cuadro 19. Efectos de algunos metales tóxicos en el ser humano.**

<b>Metal</b>	<b>Efecto</b>
Cadmio*	Cáncer de pulmón, mayor excreción urinaria de proteínas de bajo peso molecular, bronquitis, enfisema, nefrotoxicidad, infertilidad, cáncer de próstata, alteraciones neurológicas, hipertensión, enfermedades vasculares y óseas.
Mercurio	Temblor mercurial, proteinuria, descenso del CI, alteraciones neurológicas y sistema respiratorio. Enfermedad de Minamata (pérdida de visión, capacidades mentales o muerte), promueve también la pérdida de Ca y Mg
Plomo*	Enfisema, menor velocidad de conducción del impulso nervioso en las fibras, alteraciones neurológicas, nefrotoxicidad, anemia, cáncer de riñón.

**Nota:** \* Elementos que se modelaron en este estudio; CI: Coeficiente intelectual. **Fuente:** Morton (2006) y Navarro *et al.* (2007).

#### **Anexo 4. Métodos de tratamiento de aguas residuales y recuperación de suelos**

A través del tiempo se han propuesto métodos y tecnologías de tratamiento de aguas residuales, métodos de recuperación y/o eliminación de metales pesados en las aguas residuales y métodos de recuperación de suelos contaminados; en la actualidad se dispone de un amplio abanico de dichos métodos y tecnologías, algunas de aplicación habitual y otras todavía en fase experimental. Como es de esperarse, su aplicación depende de las características propias del problema, de la eficiencia esperada con cada tratamiento, la viabilidad económica así como del tiempo estimado para su desarrollo y ejecución. En las siguientes líneas se describen algunos métodos y tecnologías empleadas.

Basso *et al.* (2002) examinaron la capacidad de diferentes especies de macroalgas marinas (*Corallina officinalis* L., *Porphira columbina*, y *Codium fragile*) de la costa atlántica del sur argentino de remover metales pesados de aguas contaminadas, a partir de ensayos de biosorción utilizando principalmente soluciones diluidas de ión cadmio, como modelo. Se investigó el efecto de las dosis de muestras empleadas sobre la biosorción de cadmio, en condiciones pre-establecidas de equilibrio. Las *Corallina* y *Porphira* presentaron una alta efectividad (> 90%), superior a la determinada para *Codium* y para un carbón activado comercial empleado con fines comparativos, para dosis mayores a 0.3 g/100 mL. Se determinaron y representaron satisfactoriamente las isotermas de sorción de cadmio para las tres especies de algas mediante el modelo de Langmuir. Los valores estimados para la máxima capacidad de sorción indicaron que las *Porphira* presentan el mayor potencial para la biosorción de cadmio de aguas contaminadas, seguidas en secuencia decreciente por las *Corallina* y *Codium*. Además, a partir de ensayos de biosorción de níquel, se encontró que todas las algas son relativamente menos efectivas para capturar esta especie metálica.

Garces *et al.* (2004) propone la técnica de fotocátalisis como alternativa para el tratamiento de las aguas residuales. Es mencionado en su trabajo de investigación que en los últimos años se han implementado los procesos de oxidación avanzada (POA) como una alternativa tecnológicamente viable y novedosa para el tratamiento de los efluentes líquidos de industrias. Los POA consisten básicamente en la formación de radicales hidróxilo altamente oxidantes, los cuales contribuyen a la mineralización total de los compuestos contaminantes presentes en los efluentes industriales. Entre los procesos más utilizados se encuentra la fotocátalisis que consiste en la aceleración de una fotoreacción mediante un catalizador. Existe una aplicación de la energía solar para la eliminación

de contaminantes en aguas residuales mediante el uso de un fotocatalizador adecuado. Este sistema permite aprovechar directamente la energía solar que llega a la superficie de la tierra para provocar una serie de reacciones químicas (redox) que dan lugar a la eliminación de compuestos orgánicos en las aguas de vertidos urbanos, industriales y agrícolas y cambiar el estado de oxidación de los metales pesados de tal forma que pasan de estar disueltos a una forma insoluble.

Soto *et al.* (2004) presentaron una alternativa para remover metales pesados en aguas residuales. La experimentación se realizó utilizando cuatro coagulantes:  $Al_2(SO_4)_3$ ,  $FeSO_4$ ,  $Fe(SO_4)_3$  y  $FeCl_3$ , donde éste último mostró el mejor nivel de remoción de metales pesados así como una velocidad de sedimentación más rápida.

Reyes *et al.* (2006) estudiaron la remoción de metales pesados a través de la combinación de un método tradicional (adsorción) con uno nuevo (biosorción). En este nuevo método se implementó diferentes tipos de biomasa como microorganismos inmovilizados (*Aspergillus terreus*, *Arthrobacter viscosus*, *Escherichia coli*) sobre diversos soportes (carbón activado) y se observó la capacidad máxima de adsorción para diferentes metales, los resultados fueron favorables.

De acuerdo a su página electrónica oficial, la Universidad de Alicante (España) ha experimentado la recuperación y/o eliminación de metales pesados en aguas residuales mediante tecnología electroquímica. La recuperación del metal se lleva a cabo mediante su deposición en forma metálica sobre el cátodo en un reactor electroquímico. Esto supone la formación de una nueva fase sólida. El tipo de reactor electroquímico se determina en función del valor del metal recuperado y de las leyes medioambientales. En segundo lugar, la elección del diseño está determinada por la posibilidad de reciclar el metal en forma metálica o como una disolución concentrada. La recuperación de metales por electrodeposición normalmente se lleva a cabo a partir de disoluciones concentradas usando un reactor abierto. Esta geometría facilita la extracción y la recuperación masiva del metal. El uso de procesos electroquímicos permite obtener metales de gran pureza y de una forma mucho más ecológica que con los tratamientos convencionales, además es una tecnología respetuosa con el medioambiente ya que evita la emisión de gases, sulfuros y partículas metálicas.

De acuerdo a Ortiz *et al.* (2007), en función de los objetivos que se quieren alcanzar a la hora de recuperar un suelo contaminado se distinguen tres técnicas de recuperación:

- Técnicas de contención, que aíslan el contaminante en el suelo sin actuar sobre él, generalmente mediante la aplicación de barreras físicas en el suelo.
- Técnicas de confinamiento, que reducen la movilidad de los contaminantes en el suelo para evitar su migración actuando directamente sobre las condiciones fisicoquímicas bajo las que se encuentran los contaminantes.
- Técnicas de descontaminación, dirigidas a disminuir la concentración de los contaminantes en el suelo.

En el Cuadro 20 se resumen los principales tipos de tratamientos de recuperación y su aplicación. De acuerdo al objeto de estudio que son los metales pesados, los tratamientos importantes, son aquellos en los que el tipo de contaminante tratado hace referencia a metales y/o componentes inorgánicos.

**Cuadro 20. Principales técnicas de recuperación de suelos.**

<b>Técnica</b>	<b>Tratamiento</b>	<b>Tipo de contaminantes tratados</b>		
<b>Contención</b>	Físico- Químico	Extracción*	Compuestos orgánicos volátiles y semivolátiles, combustibles, metales, PCBs, hidrocarburos aromáticos policíclicos, disolventes halogenados y clorados, etc.	
		Lavado*	Compuestos orgánicos semivolátiles, hidrocarburos derivados del petróleo, cianuros y metales.	
		Flushing*	Todo tipo de contaminantes, especialmente inorgánicos.	
		Electrocínética*	Especialmente metales solubles o complejados.	
		Adición de enmiendas*	Fundamentalmente sales y metales.	
		Barreras permeables activas*	Contaminantes orgánicos biodegradables, metales, nitratos, sulfatos.	
		Inyección de aire comprimido	Solventes clorados, sustancias volátiles y semivolátiles ligeras como xileno, benceno, tolueno, tetracloruro de carbono, tricloroetano, cloruro de metilo, etc.	
		Pozos de recirculación	Tricloroetileno, derivados del petróleo, compuestos orgánicos no halogenados, semivolátiles, pesticidas y compuestos inorgánicos.	
		Oxidación ultravioleta	Amplio espectro de contaminantes orgánicos y explosivos (hidrocarburos del petróleo, hidrocarburos clorados, compuestos orgánicos volátiles y semivolátiles, alcoholes, cetonas, aldehidos, fenoles, éteres, pesticidas, dioxinas, PCBs, TNT, RDX y HMX)	
		Biológico	Biodegradación asistida	Amplio espectro de contaminantes orgánicos biodegradables.
			Biotransformación de metales*	Metales.
			Fitorrecuperación*	Metales, pesticidas, solventes, explosivos, hidrocarburos aromáticos policíclicos, crudo.
			Bioventing	Hidrocarburos del petróleo de peso mediano, explosivos (DDT, DNT).
			Landfarming	Fundamentalmente hidrocarburos del petróleo de peso mediano.
	Biopilas		Derivados del petróleo, compuestos orgánicos volátiles halogenados y no halogenados, compuestos orgánicos semivolátiles y pesticidas.	
	Térmico	Compostaje	Compostaje Explosivos (TNT, RDX y HMX), hidrocarburos aromáticos policíclicos, hidrocarburos del petróleo, clorofenoles y pesticidas.	
		Lodos biológicos	Compuestos orgánicos volátiles y semivolátiles no halogenados, explosivos, hidrocarburos del petróleo, petroquímicos, solventes y pesticidas.	
		Incineración	Explosivos, hidrocarburos clorados, PCBs y dioxinas.	
		Desorción térmica	Compuestos orgánicos volátiles no halogenados, combustibles, algunos compuestos orgánicos semivolátiles, hidrocarburos aromáticos policíclicos, PCBs, pesticidas y metales volátiles.	
		Mixto	Extracción multifase	Compuestos orgánicos volátiles, en disolución y compuestos orgánicos no acuosos en fase libre.
Atenuación natural			Atenuación natural Compuestos BTEX (benceno, tolueno, etil benceno y xileno), hidrocarburos clorados, algunos pesticidas y compuestos inorgánicos.	

<b>Técnica</b>	<b>Tratamiento</b>	<b>Tipo de contaminantes tratados</b>
<b>Confinamiento</b>	Barreras verticales*	Contaminantes orgánicos e inorgánicos.
	Barreras horizontales*	Contaminantes orgánicos e inorgánicos.
	Barreras de suelo seco	Contaminantes orgánicos e inorgánicos.
	Sellado profundo	Contaminantes orgánicos e inorgánicos.
	Barreras hidráulicas	Contaminantes orgánicos e inorgánicos.
<b>Descontaminación</b>	Estabilización físico-química*	Fundamentalmente compuestos inorgánicos como metales pesados.
	Inyección de solidificantes*	Fundamentalmente compuestos inorgánicos, eficacia mucho menor para compuestos orgánicos semivolátiles y pesticidas
	Vitrificación*	Contaminantes inorgánicos (principalmente Hg, Pb, Cd, As, Ba, Cr y cianuros) y algunos orgánicos.

**Nota:** \*Tratamiento que sugiere ser aplicado a metales. **Fuente:** Ortiz *et al.* (2007)