



# COLEGIO DE POSTGRADUADOS

---

---

INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN  
EN CIENCIAS AGRÍCOLAS

**CAMPUS MONTECILLO**

POSTGRADO DE SOCIOECONOMÍA, ESTADÍSTICA E  
INFORMÁTICA  
ESTADÍSTICA

**Análisis de valores extremos con datos censurados  
mediante regresión semiparamétrica**

Alejandro Ivan Aguirre Salado

T E S I S

PRESENTADA COMO REQUISITO PARCIAL PARA  
OBTENER EL GRADO DE:

DOCTOR EN CIENCIAS

MONTECILLO, TEXCOCO, EDO. DE MÉXICO  
2015

---

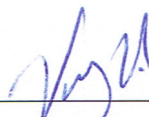
---

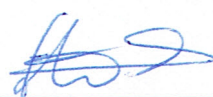
La presente tesis titulada: **Análisis de valores extremos con datos censurados mediante regresión semiparamétrica**, realizada por el alumno: **Alejandro Ivan Aguirre Salado**, bajo la dirección del Consejo Particular indicado ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

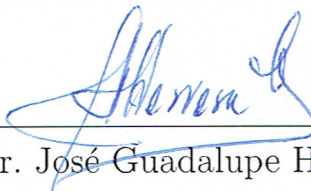
## DOCTOR EN CIENCIAS

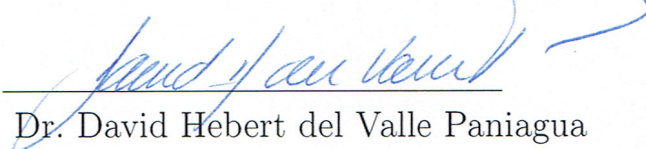
### SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA ESTADÍSTICA


#### CONSEJO PARTICULAR

CONSEJERO   
\_\_\_\_\_  
Dr. Humberto Vaquera Huerta

ASESOR   
\_\_\_\_\_  
Dr. José Luis Crossa Hiriart

ASESOR   
\_\_\_\_\_  
Dr. José Guadalupe Herrera Haro

ASESOR   
\_\_\_\_\_  
Dr. David Hebert del Valle Paniagua

ASESOR   
\_\_\_\_\_  
Dr. Barry C. Arnold

# Análisis de valores extremos con datos censurados mediante regresión semiparamétrica

Alejandro Ivan Aguirre Salado

Colegio de Postgraduados, 2015

En este trabajo se presenta un nuevo enfoque para analizar los valores extremos no estacionarios basados en una regresión semiparamétrica a los parámetros de localidad y escala de la distribución de valores extremos generalizada (GEV, por sus siglas en inglés). Para ello se utiliza una función suavizadora spline penalizada de base radial como predictor lineal del parámetro de localidad y escala; también, se utilizan como nodos los centroides del agrupamiento jerárquico de los datos estandarizados. Posteriormente extendemos nuestro modelo al caso bayesiano con datos censurados, de tal forma que es posible incluir información a priori para modelar en un entorno más flexible un conjunto de datos mas general como son los valores extremos censurados. Adicionalmente se incluye un análisis visual de la interacción de las variables sobre el efecto en los parámetros de la distribución GEV.

**Palabras clave:** Teoría de valores extremos, Spline Multivariados Penalizados, MCMC, Datos censurados, Cambio climático, Tendencia.

# Analysis of nonstationary extreme values using a semiparametric regression approach

Alejandro Ivan Aguirre Salado

Colegio de Postgraduados, 2015

In this work, a new approach is presented for analyzing non-stationary extreme values based on semi-parametric regression functions for the location and scale parameters of the generalized extreme value distribution (GEV). For this, penalized multivariate smoothing spline functions with radial basis are used as linear predictors of the location and scale parameters; while the centroids of hierarchical clustering of the standardized data are used as nodes. Results show a better fit to simulated and real data, compared with those obtained using vector generalized additive models. Subsequently we extend our model to Bayesian case with censored data, so it is possible to include a priori information modeling in a more flexible framework a set of more general dataset as are the censored extreme values. Finally, a plot of the variables against the estimated parameters of the GEV distribution is included.

**Key words:** Extreme value theory, multivariate penalized spline, MCMC, censored data, climate change, trend

## AGRADECIMIENTOS

A Dios por que me permitió concluir mis estudios.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico brindado, muy necesario para mi manutención durante la realización de mis estudios.

Al Colegio de Postgraduados, por haberme brindado la oportunidad de seguir mi formación académica en sus aulas.

A los integrantes de mi Consejo Particular, por compartir y transmitirme de manera desinteresada sus conocimientos adquiridos, por colaboración desinteresada en el presente trabajo y por su tiempo en la revisión de este trabajo de tesis.

A quienes fueron mis profesores y pusieron en mis manos herramientas valiosas que utilizaré en mi quehacer profesional.

A mis compañeros de clase y además amigos, quienes siempre brindaron apoyo.

A todas aquellas personas que no mencioné, pero de alguna manera influyeron en mi formación, a todos gracias.

## DEDICATORIA

A mis primeros maestros y además... padres: Margarito Aguirre Bravo y Bertha Salado Morales.

A mis hermanos, los cuales han sido fuente de motivación y además... admiró: Carlos Arturo Aguirre Salado, Olimpia Talya Aguirre Salado.

A todos mis familiares, a los que todavía están y a los desafortunadamente ya se han ido.

# Índice

<b>1. Introducción</b>	<b>1</b>
<b>2. Objetivos</b>	<b>3</b>
2.1. Objetivos Generales . . . . .	3
2.2. Objetivos particulares . . . . .	3
<b>3. Marco Teórico</b>	<b>4</b>
3.1. Máximos por bloques . . . . .	5
3.2. Distribución GEV . . . . .	5
3.3. Picos sobre umbral . . . . .	6
3.4. Censura por la derecha . . . . .	6
3.5. Estimación de los parámetros de la GEV bajo censura aleatoria . . . . .	7
3.5.1. Series no estacionarias . . . . .	8
3.6. Cadenas de Markov Monte Carlo . . . . .	9
3.6.1. El algoritmo de Metropolis-Hastings . . . . .	10
3.6.2. Implementación Bayesiana . . . . .	11
<b>4. Análisis de valores extremos a datos no censurados</b>	<b>12</b>

# Índice

---

4.1. Introducción . . . . .	12
4.2. Metodología . . . . .	13
4.2.1. Distribución GEV no estacionaria . . . . .	13
4.2.2. Modelo propuesto . . . . .	13
4.2.3. Estimación por máxima verosimilitud . . . . .	14
4.2.4. Elección del Kernel . . . . .	16
4.3. Estudio con datos sintéticos . . . . .	17
4.4. Valores extremos de máximos de lluvia . . . . .	19
4.5. Discusión . . . . .	21
<b>5. Análisis bayesiano de valores extremos con datos censurados</b>	<b>22</b>
5.1. Introducción . . . . .	22
5.2. Metodología . . . . .	22
5.3. Datos Sintéticos . . . . .	25
5.4. Niveles Máximos de contaminante PM10 . . . . .	29
5.5. Discusión . . . . .	34
<b>6. Conclusiones</b>	<b>35</b>
<b>Referencias</b>	<b>35</b>
<b>Anexos</b>	<b>39</b>
Anexo A: Rutinas en R-2.12.2 para implementar los modelos Frecuentistas y Bayesianos para datos censurados . . . . .	39



# Índice de tablas

4.1. Resultados de la simulación del sesgo y la desviación estándar de las estimaciones a partir de un tamaño de muestra $n$ de observaciones GEV. . .	17
5.1. Estimadores y 95 % intervalo de credibilidad para $\kappa$ (con valor verdadero de -0.5), por nivel de censura. . . . .	25
5.2. Estimadores y el intervalo de credibilidad al 95 % para $\kappa$ (verdadero valor de -0.5), el intercepto de $\mu$ y $\log \sigma$ . . . . .	29
5.3. Estimadores y 95 % intervalos de credibilidad para los parámetros en el modelo 4.2 usando los datos de máximos de PM10. . . . .	32

# Índice de figuras

3.1. Ilustración del esquema de censura en máximos de bloques en estación la Merced . . . . .	7
4.1. Función real (a,c) y ajustada (b,d) a los parámetros de datos simulados con $n = 3000$ de un modelo GEV no estacionario. . . . .	18
4.2. Mapa de Suiza y la distribución espacial de las estaciones estudiadas. . .	19
4.3. Funciones estimadas para el parámetro de localidad ( $a, b, c$ ) y escala ( $d, e, f$ ) para la región de estudio en 3 periodos de tiempo (1962, 1984, 2008). Los ejes $x$ y $y$ en cada gráfica representan la región geografica y los valores de los parametros están representados por la superficie. . . . .	20
5.1. Función real para $\mu$ (a) y $\sigma$ (b) usados en el estudio de simulación. . . .	26
5.2. Funciones estimadas para el parametro de localidad en cada uno de los niveles de censura de 0, 5, 10 y 15 % mostradas en las figuras (a),(b),(c),(d) respectivamente. . . . .	27
5.3. Funciones estimadas para el parámetro de escala en cada uno de los niveles de censura de 0, 5, 10 y 15 % mostradas en las figuras (a),(b),(c),(d) respectivamente. . . . .	28
5.4. Funciones estimadas para el parámetro de localidad (a) y escala (b) para niveles de censura de 15 %. . . . .	29
5.5. Valores extremos con datos censurados (puntos rojos) en la estación pedregal, Mex. . . . .	30
5.6. Funciones estimadas para el parámetro de localidad en el año 2000(a) y año 2015 (b) y escala en el año 2000(c) y año 2015 (d). . . . .	31

## Índice de figuras

---

5.7. Niveles de retorno con un periodo de retorno de 25 años para la región de estudio. . . . .	32
5.8. Funciones estimadas para el parámetro de localidad (a) y escala (b) en el año 2015. . . . .	33
5.9. Niveles de retorno con periodo de 25 años en la región de estudio. . . . .	34

# Capítulo 1

## Introducción

El análisis de valores extremos permite utilizar información anterior para calcular probabilidades de eventos extremos futuros (El Adlouni *et al.*, 2007). Los valores máximos pueden ser modelados mediante la distribución de valores extremos que emplea tres parámetros correspondientes a la localización, la escala y la forma (Jenkinson, 1955). Inicialmente, la estimación de dichos parámetros se realizaba empleando el método de máxima verosimilitud (Smith, 1985). No obstante, al aplicar este método se podrían presentar problemas de singularidades en las derivadas. Para superar esta dificultad, algunos investigadores propusieron el método de momentos, el de L momentos e inclusive el método de momentos con probabilidades ponderadas (Hosking, 1990, Hosking *et al.*, 1985, Madsen *et al.*, 1997).

Por otra parte, la distribución de los valores extremos se basa en una teoría asintótica y es válida para muestras grandes, por lo que el método de máxima verosimilitud puede ser inestable cuando la muestra es pequeña. Para resolver este detalle, existen soluciones donde se propone agregar una penalización en la verosimilitud y asignar una distribución a priori al parámetro de forma (Coles y Dixon, 1999, Martins y Stedinger, 2000).

Diversos autores consideraron diversos modelos jerárquicos bayesianos, Gaetan y Grigoletto (2007) propusieron usar campos aleatorios de Markov aproximados por suavizado kernel para modelación de los parámetros de la distribución GEV, Reich et al. estudio las ondas de calor mediante un modelo jerárquico bayesiano con distribución generalizada de Pareto y asignó a los parámetros un modelo de Markov dependiente del tiempo. Cooley y Sain (2010) estudiaron eventos de precipitaciones máximas asignando un modelo normal con covariables temporales a los parámetros de la distribución GPD. Sang y Gelfand (2010) estudiaron procesos estocásticos espaciales para valores extremos y modelaron la tendencia como función de covariables.

Por otro lado, cuando se utilizan datos reales, el supuesto de estacionariedad puede no cumplirse ya que se ha detectado la presencia de tendencias en valores extremos ([Leadbetter \*et al.\*, 1983](#)), esto ha sido encontrado principalmente en datos hidroclimatológicos, donde las condiciones futuras cambian ([Jain y Lall, 2001](#), [Kharin y Zwiers, 2005](#), [Wang \*et al.\*, 2004](#)). El encontrar tendencias en dichos valores justifica la utilización de covariables dentro de la distribución de valores extremos. Varios investigadores han introducido funciones para modelar ya sea el parámetro de localización o el parámetro de escala. Por ejemplo, para el parámetro de localización, [Weissman \(1978\)](#) empleó una función senoidal, mientras que [Tawn \(1988\)](#) y [Scarf \(1992\)](#) propusieron una función lineal, y otros investigadores como [Rosen y Cohen \(1996\)](#) y [Pauli y Coles \(2001\)](#) emplearon suavizadores de tipo spline para dar mayor flexibilidad al ajuste. Recientemente, los investigadores se han interesado en modelar el parámetro de la escala. Por ejemplo, [Yee y Stephenson \(2007\)](#) utilizaron el modelo spline para modelar el logaritmo del parámetro de la escala, [Rodríguez \*et al.\* \(2012\)](#) y [El Adlouni \*et al.\* \(2007\)](#) emplearon funciones lineales. En estos modelos no se consideró la interacción entre variables. Fue hasta que [Cannon \(2010\)](#) sugirió emplear funciones de redes neuronales así como la interacción entre las covariables del modelo.

Este trabajo introduce un nuevo tipo de función spline multivariado, basada en la norma, que permite incluir la interacción y modelar funciones de la localidad y escala, las cuales pueden ser ambas no lineales, en función de covariables y mediante una forma aproximadamente lineal. La estructura de este trabajo es como sigue. En el capítulo 3 se presenta la teoría de distribución de valores extremos no estacionaria. En el capítulo 4 se prueba el modelo propuesto para datos no censurados utilizando datos sintéticos y datos extremos de lluvia. En el capítulo 5 proponemos un modelo para datos censurados y finalmente en el capítulo 6 se presentan las principales conclusiones de este trabajo.

# Capítulo 2

## Objetivos

### 2.1. Objetivos Generales

- Desarrollar una nueva metodología en el análisis de valores extremos que permita capturar más información de un conjunto de datos.
- Demostrar con datos reales y mediante simulación que el modelo de valores extremos con un enfoque semiparamétrico es un método eficiente para el análisis de situaciones acordes a la realidad.

### 2.2. Objetivos particulares

- Incluir de forma implícita, los efectos de interacción entre covariables para modelar los parámetros de la distribución de valores extremos.
- Ajustar un modelo semiparamétrico bayesiano que permita incluir información a priori.
- Modelar valores extremos con datos censurados

# Capítulo 3

## Marco Teórico

Las dos metodologías comúnmente usadas para obtener los valores máximos de una serie de datos en el análisis de valores extremos son:

- Máximos por bloque.
- Picos sobre umbral.

En el primer caso la información se clasifica en bloques temporales o espaciales, y se escoge el valor máximo de cada bloque para realizar el análisis sucesivo de valores extremos, entonces se asume que cada extremo en un bloque siguen todos la misma distribución conocida como distribución generalizada de valores extremos (GEV por sus siglas en inglés).

En el caso de picos sobre umbral, como su nombre lo indica, los valores máximos son aquellos que sobrepasan un umbral, y en el subsecuente análisis se asume que estos extremos siguen todos la misma distribución generalizada de Pareto (GPD). Estos métodos presentan el inconveniente de que están basados en supuestos que se tienen que cumplir para que las inferencias sean correctas, i.e. estacionariedad, independencia, etc. para solventar esto, los modelos más recientes permiten relajar los modelos de tal forma que cada valor extremo sigue una distribución GEV o GPD con sus propios parámetros de localidad, escala y forma. La ventaja de esto es que permite flexibilizar el modelo, puesto que se modela a cada valor extremo de acuerdo a su información espacial, temporal, climática, etc.

## 3.1. Máximos por bloques

Uno de los métodos para obtener los valores extremos de una serie, es el llamado máximo por bloques, en donde se dividen los datos en secciones de igual tamaño y se escoge el valor más grande dentro de cada bloque. La ventaja de este método es que se escogen valores sobre todo el conjunto de datos, sin embargo, se pueden omitir los siguientes valores extremos dentro del mismo bloque que posiblemente sean mayores que el máximo dentro de otro bloque. Entonces se asume que la distribución que siguen los máximos por bloques es la distribución de valores extremos generalizada (GEV del inglés Generalized Extreme Value distribution).

## 3.2. Distribución GEV

Dada una una muestra de variables aleatorias independientes e idénticamente distribuidas,  $X_1, \dots, X_n \sim F_X(x)$ . El valor extremo es definido como:

$$M_n = \max \{X_i | i = 1, \dots, n\}$$

Puesto que las X's son independientes,

$$F_{M_n}(x) = P(M_n \leq x) = P(X_1 \leq x, \dots, X_n \leq x) = (F_X(x))^n$$

Desde que  $F_X(x) < 1$ , entonces  $(F_X(x))^n \rightarrow 0$  Para resolver esto  $F_X(x)$  es reescalada por constantes  $\mu$  y  $\sigma$  tal que:

$$M_n^* = \frac{M_n - \mu}{\sigma}$$

La distribución de valores extremos, fue propuesta originalmente por [Fisher y Tippett \(1928\)](#) e incluía tres familias: Gumbel, Fréchet y Weibull. Posteriormente [Jenkinson \(1955\)](#) combinó las tres familias en la distribución de valores extremos (GEV):

$$F(x, \mu, \alpha, \kappa) = \begin{cases} \exp \left\{ - \left( 1 + \kappa \frac{(y-\mu)}{\sigma} \right)^{-\frac{1}{\kappa}} \right\}, & \text{if } \kappa \neq 0, 1 - \kappa \frac{(y-\mu)}{\sigma} > 0; \\ \exp \left\{ - \exp \left( - \frac{(y-\mu)}{\sigma} \right) \right\}, & \text{if } \kappa = 0. \end{cases}$$

Con función de densidad de probabilidades dada por:

$$f(y, \mu, \alpha, \kappa) = \begin{cases} \frac{1}{\sigma} \left\{ \left( 1 + \kappa \frac{(y-\mu)}{\sigma} \right)^{-\frac{(1+\kappa)}{\kappa}} \right\} \exp \left\{ - \left( 1 + \kappa \frac{(y-\mu)}{\sigma} \right)^{-\frac{1}{\kappa}} \right\}, & \text{if } \kappa \neq 0, 1 - \kappa \frac{(y-\mu)}{\sigma} > 0; \\ \exp \left\{ - \frac{(y-\mu)}{\sigma} \right\} \exp \left\{ - \exp \left( - \frac{(y-\mu)}{\sigma} \right) \right\}, & \text{if } \kappa = 0. \end{cases}$$



### 3.3. Picos sobre umbral

---

Donde  $\mu + \sigma/\kappa \leq y \leq +\infty$  when  $\kappa < 0$  (Fréchet),  $-\infty \leq y \leq +\infty$  cuando  $\kappa = 0$  (Gumbel) y  $-\infty \leq y \leq \mu + \sigma/\kappa$  cuando  $\kappa > 0$  (Weibull). Aquí  $\mu \in \mathbb{R}, \sigma > 0$  y  $\kappa \in \mathbb{R}$  son los parámetros de localización, escala y forma respectivamente.

### 3.3. Picos sobre umbral

El método de picos sobre umbral POT, por sus siglas en inglés, selecciona los valores más grandes que sobrepasan un umbral, por lo que la mayor parte de los resultados de este método se basan en la distribución de los excesos sobre dicho umbral. Supóngase que se tiene una variable  $R$  con función de distribución  $F_R$ , la función condicional de  $R$  dado que es mayor que un umbral  $u$  se conoce como la distribución de los excesos de  $R$ ,  $F_{R,v}$  esta dada por:

$$F_{R,v} = P \{R - u \leq y | R > u\}$$

Para encontrar la distribución de los excesos, Balkema y de Hann ([Balkema y De Haan, 1974](#)), Pickands ([Pickands III, 1975](#)) desarrollaron el siguiente teorema: Para una gran clase de funciones de distribución, la distribución de los excesos de  $R$ ,  $F_{R,v}$ , para valores grandes de  $u$ , es aproximadamente igual a:

$$F_{R,v}(y) \approx G_{\xi,\beta} = \begin{cases} 1 - (1 + \xi y/\beta)^{1/\xi} & \text{si } \xi \neq 0 \\ 1 - \exp(-y/\beta) & \text{si } \xi = 0 \end{cases}$$

Donde  $\xi \in \mathbb{R}$ . Los parámetros  $\xi$  y  $\beta$  son conocidos como los parámetros de forma y escala, y  $G_{\xi,\beta}$  es conocida como la distribución generalizada de Pareto, GDP por sus siglas en inglés. Dependiendo del valor del parámetro  $\xi$  de la GPD se obtienen tres tipos de funciones de distribución: Si  $\xi > 0$ , la GPD es una distribución de Pareto con parámetros  $\alpha = 1/\xi$  y  $k = \beta/\xi$  para valores  $y \geq 0$ . Para  $\xi = 0$  la GPD corresponde a una distribución exponencial con parámetro  $1/\beta$  y  $y \geq 0$ . Finalmente si  $\xi < 0$ , las GPD toman la forma de una distribución tipo Pareto II, la cual está definida en el rango  $0 \leq y \leq \beta/\xi$ .

### 3.4. Censura por la derecha

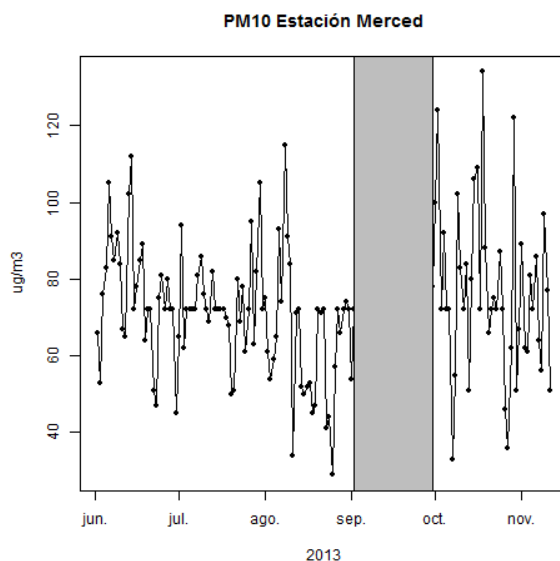
Decimos que una observación presenta censura por la derecha cuando el evento que se espera, sucede a la derecha de un valor utilizado como censura. En este caso la variable aleatoria se representa como  $Y_i = \min(X_i, C_i)$ , donde  $X_i$  representa al dato o evento y  $C_i$  al valor de la censura. Cuando  $C_i$  se considera fija y la misma para todos los datos, se le conoce como censura tipo I; si  $C_i$  se considera fija, pero diferente para cada dato, se le llama censura tipo I generalizada; si  $C_i$  se considera fija, la misma para todos los

### 3.5. Estimación de los parámetros de la GEV bajo censura aleatoria

---

datos y además es igual al  $r$ -ésimo estadístico de orden de los datos, se le llama censura tipo II, finalmente si  $C_i$  se considera una variable aleatoria y es independiente de  $X_i$  se le conoce como censura aleatoria.

En este trabajo, consideramos que un valor extremo dentro de un bloque es censurado si dentro del bloque existen observaciones faltantes, entonces escogemos el segundo valor extremo en dicho bloque para garantizar que los supuestos necesarios en la censura aleatoria se cumplan.



**Figura 3.1:** Ilustración del esquema de censura en máximos de bloques en estación la Merced

### 3.5. Estimación de los parámetros de la GEV bajo censura aleatoria

En la medición de fenómenos ambientales, es frecuente que produzcan problemas en equipos, lo que nos genera de datos censurados. Así, nosotros obtenemos los valores máximos  $M$ , en bloques temporales de información y asumimos que en tal bloque puede existir un valor censurado  $C$ . sumimos que las variables  $M$  y  $C$  son independientes. Sea  $Y = \min(M, C)$  y sea  $\delta$  la variable que indica si el valor de  $Y$  es censurado  $\delta = 0$  o no  $\delta = 1$ . Entonces la distribución de  $M$ , esta dada por:

$$M \sim GEV(\mu, \alpha, \kappa)$$

### 3.5. Estimación de los parámetros de la GEV bajo censura aleatoria

Sea  $G$  la función de distribución de  $M$ , y  $g$  su función de densidad de probabilidades, asuma que  $\theta = (\mu, \alpha, \kappa)$ . Similarmente sea  $F$  y  $f$  la función de distribución y la función de densidad de  $C$ , respectivamente.

Sea  $S_i = M_i \wedge C_i$  y  $\delta_i = 1 (Y_i = M_i)$ . Así,  $Y_i$  es el valor observado y posiblemente censurado valor extremo. Para datos con censura aleatoria tenemos:

$$\begin{aligned} P [Y_i = y, \delta_i = 1; s_i, \theta] &= P [M_i = y, C_i > y; s_i, \theta] \\ &= F(y) g(y; \theta, s_i) \end{aligned} \quad (3.1)$$

y

$$\begin{aligned} P [Y_i = y, \delta_i = 0; s_i, \theta] &= P [C_i = y, M_i > y; s_i, \theta] \\ &= f(y) G^*(y; \theta, s_i) \end{aligned} \quad (3.2)$$

Asuma que las  $s_i, \dots, s_n$ , las parejas  $(Y_i, \delta_i)$ , son independientes. La función de verosimilitud sobre los datos  $(Y_i = y_i, \delta_i, s_i), i = 1, \dots, n$ , condicional sobre las  $s_i, \dots, s_n$  es

$$L(\mu, \alpha, \kappa | y) = \prod_{i=1}^n [G(y_i; \mu, \alpha, \kappa) f(y_i)]^{1-\delta_i} [F(y_i) g(y_i; \mu, \alpha, \kappa)]^{\delta_i}$$

Arreglando los terminos obtenemos finalmente:

$$L(\mu, \alpha, \kappa | y) = \left\{ \prod_{i=1}^n [F(y_i)]^{\delta_i} [f(y_i)]^{1-\delta_i} \right\} \left\{ \prod_{i=1}^n [G(y_i; \mu, \alpha, \kappa)]^{1-\delta_i} [g(y_i; \mu, \alpha, \kappa)]^{\delta_i} \right\}$$

Así, si la censura es no informativa, i.e.,  $F_i(y)$  no contiene a los parametros en  $\theta$ , obtenemos

$$L(\mu_i, \sigma_i, \kappa | y_i) \propto \prod_{i=1}^n [G^*(y_i; \mu_i, \sigma_i, \kappa)]^{1-\delta_i} [g(y_i; \mu_i, \sigma_i, \kappa)]^{\delta_i}$$

#### 3.5.1. Series no estacionarias

Una serie  $X_1, \dots, X_n$  se dice que es estacionaria en el sentido estricto si:

$$F_{X_{t_1}, \dots, X_{t_n}}(X_{t_1}, \dots, X_{t_n}) = F_{X_{t_1+k}, \dots, X_{t_n+k}}(X_{t_1+k}, \dots, X_{t_n+k})$$

si la condición anterior no se cumple entonces la serie es no estacionaria. La estacionaridad en el sentido estricto es una condición muy fuerte difícil de probar en la práctica. Por ello existe otro concepto conocido como, estacionaridad en el sentido débil, en la cual:

### 3.6. Cadenas de Markov Monte Carlo

---

- La media de  $X_t$  es constante.
- La autocovarianza entre  $X_t$  y  $X_{t+h}$  solo depende de  $h$ .

## 3.6. Cadenas de Markov Monte Carlo

El término Cadenas de Markov Monte Carlo es usado para designar diferentes algoritmos cuyo objetivo es producir muestras de una distribución  $\pi(x)$ .

Una cadena de Markov es un proceso estocástico en donde la distribución en cualquier instante, solo depende de su valor en el instante anterior. Es lo mismo que decir  $P(x_{n+1})$  dado todos los valores anteriores es la misma que  $P(x_{n+1}|x_n)$ .

Un resultado importante acerca de una cadena de Markov, es que esta tiene una única distribución límite  $\pi(x)$ , si cumple con las siguientes propiedades:

- Es irreducible, es decir que la cadena puede ir del estado  $x$  a cualquier otro estado con probabilidad mayor que 0 en un numero finito de pasos.
- Es aperiódica, es decir que puede regresar a cualquier estado  $x$  en cualquier tiempo.
- Es positiva recurrente, es decir que el tiempo entre dos visitas a un mismo estado es finito.
- Satisface la con la condición de balance detallado.

La distribución límite es definida como:

$$\pi(x) = \lim_{i \rightarrow \infty} P(X_i = x)$$

también se le conoce como distribución estacionaria y esta dada por

$$\pi(y) = \sum_{x \in \chi} \pi(x) \cdot K(x, y) \quad \forall x \in \chi \quad (3.3)$$

$$\sum_{x \in \chi} \pi(x) = 1$$

la condición de balance detallado nos dice que una cadena de markov con matriz de transición  $P$  satisface la condición de balance detallado si existe una función  $\pi$  que

## 3.6. Cadenas de Markov Monte Carlo

---

satisface:

$$P(y, x) \pi(y) = P(x, y) \pi(x) \quad (3.4)$$

esta condición nos dice que la probabilidad de ir de  $x$  a  $y$  es la misma que de ir de  $y$  a  $x$ , y que la cadena de markov es reversible.

Note que la ecuación 3.4 es una condición suficiente para 3.3:

$$\begin{aligned} \sum_{x \in \mathcal{X}} P(y, x) \pi(y) &= \sum_{x \in \mathcal{X}} P(x, y) \pi(x) \\ &= \pi(y) \sum_{x \in \mathcal{X}} P(y, x) \\ &= \pi(y) \end{aligned}$$

### 3.6.1. El algoritmo de Metropolis-Hastings

La simulación de una distribución mediante MCMC se puede hacer con el algoritmo llamado Metropolis-Hastings el cual consiste en los siguientes pasos:

- Simula una muestra  $x_t$  de una distribución conocida  $Q(x_t|x_{t-1})$ ,  $x_0$  es cualquier valor inicial.
- La nueva muestra es aceptada o rechazada con probabilidad:  
$$A(x_t|x_{t-1}) = \min\left(1, \frac{P(x_t)Q(x_{t-1}|x_t)}{P(x_{t-1})Q(x_t|x_{t-1})}\right)$$
Esto se hace simulando un valor  $u \sim Uniform(0, 1)$  y se acepta la muestra si  $u < A(x_t|x_{t-1})$

El algoritmo de Metropolis-Hasting puede ser visto como un algoritmo de muestreo por importancia adaptativo.

Para ver por que el el algoritmo de Metropolis-Hasting funciona, vea que si  $A(x_t|x_{t-1}) < 1$  entonces  $\frac{P(x_{t-1})Q(x_{t-1}|x_t)}{P(x_t)Q(x_t|x_{t-1})} > 1$  y  $A(x_{t-1}|x_t) = 1$

Por otra parte suponga que  $A(x_t|x_{t-1}) < 1$  y  $A(x_{t-1}|x_t) = 1$ , entonces

$$A(x_t|x_{t-1}) = \frac{P(x_t) Q(x_{t-1}|x_t)}{P(x_{t-1}) Q(x_t|x_{t-1})}$$

$$A(x_t|x_{t-1}) P(x_{t-1}) Q(x_t|x_{t-1}) = P(x_t) Q(x_{t-1}|x_t)$$

$$A(x_t|x_{t-1}) P(x_{t-1}) Q(x_t|x_{t-1}) = P(x_t) Q(x_{t-1}|x_t) A(x_{t-1}|x_t)$$

$$P(x_{t-1}) T(x_t|x_{t-1}) = P(x_t) T(x_{t-1}|x_t)$$

### 3.6. Cadenas de Markov Monte Carlo

---

Donde  $T(x_{t-1}|x_t) = P(x_{t-1})Q(x_t|x_{t-1})$  es el kernel de transición, note que esta ultima ecuación es la condición de balance detallado, por lo tanto el algoritmo de metropolis-Hastings nos lleva a una distribución estacionaria  $P(x)$

#### 3.6.2. Implementación Bayesiana

Asumiendo que  $\pi(y_t|\theta_t)$  es la distribución GEV con parámetros con parámetros  $\theta_t = (\mu_t, \sigma_t, \kappa_t)$ , y que la relación que liga a los parámetros con las covariables es

$$\begin{aligned}\mu_t &= X\beta_{(1)} + Zu_{(1)}, \\ \log \sigma_t &= X\beta_{(2)} + Zu_{(2)}, \\ \kappa_t &= \kappa\end{aligned}\tag{3.5}$$

una formulación bayesiana para el modelo de valores extremos es la siguiente:

$$\pi(\theta_t, \omega|y_t) \propto \pi(y_t|\theta_t) \pi(\theta_t|\omega) \pi(\omega)\tag{3.6}$$

Donde  $\omega = (\beta_1, \beta_2, u_1, u_2)$  son los parámetros que ligan a las covariables con las observaciones de valores extremos. Este modelo es un modelo jerárquico con tres niveles: el nivel de datos  $\pi(y_t|\theta_t)$ ; el nivel del proceso  $\pi(\theta_t|\omega)$  y el nivel de la a priori  $\pi(\omega)$ .

Un inconveniente de este modelo es que durante la simulación por métodos de cadenas de markov Montecarlo, tiene un alto costo computacional, el número de parámetros a estimar se incrementa en proporción igual al tamaño de la muestra. Una modelo alternativo es presentado por [Bocci et al. \(2013\)](#), donde el modelo jerárquico bayesiano 3.6 puede ser visto directamente como función del espacio de parámetros  $\omega$ , de tal forma que podemos escribir la densidad aposteri como:

$$\pi(\omega|y_t) \propto \pi(y_t|\omega) \pi(\omega)$$

Donde  $\pi(y_t|\omega)$  es la densidad GEV con las condiciones sobre los parámetros 3.2 Para muestrear de la distribución aposteri, utilizamos el método de cadenas de markov monte carlo (MCMC) con probabilidad de aceptación está dado por:

$$\alpha(\theta^*|\theta) = \min\left(1, \frac{\pi(x|\theta^*) \pi(\theta^*) Q(\theta^*, \theta)}{\pi(x|\theta) \pi(\theta) Q(\theta, \theta^*)}\right)$$

Donde  $\pi(\theta)$  es la distribución a priori para los parámetros y  $\pi(x|\theta)$  es la verosimilitud.  $Q$  es la función generadora de candidatos.

# Capítulo 4

## Análisis de valores extremos a datos no censurados

### 4.1. Introducción

Recientemente se ha venido incrementado la atención para entender y modelar los eventos extremos. Inundaciones, vientos, huracanes, avalanchas, ondas de calor y sequías son algunos ejemplos en donde es utilizado el análisis de valores extremos. El conocimiento y entendimiento estos eventos es utilizado para predecir y minimizar el efecto de riesgos que estos fenómenos provocan, es por esto que se ha prestado especial interés en mejorar las metodologías para analizar dichos eventos.

La mayor parte de las metodologías para analizar los eventos extremos están basadas en los trabajos iniciales de [Fisher y Tippett \(1928\)](#) y posteriormente [Jenkinson \(1955\)](#), los cuales demostraron que las distribuciones límites de los valores máximos, solo pueden ser las distribuciones Gumbel, Fréchet o Weibull. Así, la literatura relativa al estudio de los valores extremos es vasta, y va desde el estudio de los valores extremos mediante copulas y procesos max estables ([Sang y Gelfand, 2010](#)) hasta las recientes metodologías semiparamétricas y bayesianas ([Bocci \*et al.\*, 2013](#)).

En este capítulo introducimos la idea principal sobre la cual empezamos a construir nuestro modelo, esto es, la introducción de splines multivariados que nos ayudan a estimar los parámetros de localidad y escala a través de covariables. Esta metodología consiste en generar bases gaussianas que incrementan el espacio de parámetros del modelo y consecuentemente mejoran la verosimilitud de los datos.

## 4.2. Metodología

La idea de aumentar el espacio paramétrico dentro de un modelo lineal generalizado, para mejorar las predicciones fue introducida formalmente en los modelos aditivos generalizados (GAM por sus siglas en ingles) (Hastie y Tibshirani, 1986), posteriormente se utilizaron nuevos modelos tales como las redes neuronales (Cannon, 2010) y los modelos de kernels reproducibles en espacios de Hilbert (Nosedal-Sanchez *et al.*, 2012). Todos estos modelos explotan el hecho de que al aumentar la dimensión de las variables independientes, los problemas no lineales se vuelve aproximadamente lineales y se disminuye la devianza.

En un principio, los modelos GAM fueron desarrollados para modelar el efecto no lineal de cada covariable en la variable respuesta por medio de un modelo aditivo. Estos modelos podían incluir explícitamente los términos correspondientes a las interacciones entre covariables. Posteriormente los modelos de kernels reproducibles en espacios de Hilbert, permitieron manejar la información de las covariables de manera multivariada, por medio de funciones kernel, las cuales aumentan la dimensión de las variables independientes. Algunas de las funciones kernel mas utilizadas son el kernel Gaussiano y el kernel ThinPlates (Walder y Chapelle, 2007).

### 4.2.1. Distribución GEV no estacionaria

En el caso no estacionario, asumimos que cada observación  $y_t$  se distribuye GEV, con parámetros  $\mu_t$ ,  $\alpha_t$  y  $\kappa_t$ :

$$f(y_t, \mu_t, \alpha_t, \kappa_t) = \begin{cases} \frac{1}{\sigma_t} \left\{ \left( 1 + \kappa_t \frac{(y_t - \mu_t)}{\sigma_t} \right)^{-\frac{(1+\kappa_t)}{\kappa_t}} \right\} \exp \left\{ - \left( 1 + \kappa_t \frac{(y_t - \mu_t)}{\sigma_t} \right)^{-\frac{1}{\kappa_t}} \right\}, & \text{if } \kappa_t \neq 0, \quad 1 - \kappa_t \frac{(y_t - \mu)}{\sigma_t} > 0 \\ \exp \left\{ -\frac{(y_t - \mu_t)}{\sigma_t} \right\} \exp \left\{ - \exp \left( -\frac{(y_t - \mu_t)}{\sigma_t} \right) \right\}, & \text{if } \kappa_t = 0. \end{cases}$$

Donde  $\mu_t + \sigma_t/\kappa_t \leq y_t \leq +\infty$  when  $\kappa_t < 0$  (Fréchet),  $-\infty \leq y_t \leq +\infty$  cuando  $\kappa_t = 0$  (Gumbel) y  $-\infty \leq y \leq \mu_t + \sigma_t/\kappa_t$  cuando  $\kappa_t > 0$  (Weibull). Aquí  $\mu_t \in \mathbb{R}$ ,  $\sigma_t > 0$  y  $\kappa_t \in \mathbb{R}$  son los parámetros de localid, escala y forma respectivamente.

### 4.2.2. Modelo propuesto

Para manejar el caso no estacionario, en este trabajo consideramos asignar un predictor lineal de los parámetros de localid y escala, y se mantiene constante el parámetro de forma, ver Yee y Stephenson (2007), y proponemos usar el siguiente modelo:



## 4.2. Metodología

---

$$\begin{aligned}\mu_t &= X\beta_{(1)} + Zu_{(1)}, \\ \log \sigma_t &= X\beta_{(2)} + Zu_{(2)}, \\ \kappa_t &= \kappa\end{aligned}\tag{4.1}$$

Donde  $X$  es una matriz  $n \times p_1$ ,  $\beta_{(i)}$   $i=1,2$  es un vector de tamaño  $p_1$ ,  $Z$  es una matriz  $n \times p_2$  de relaciones dada por la ecuación 4.2,  $u_{(i)}$   $i=1,2$  es un vector de tamaño  $p_2$ ,  $\underline{x}_i$  es el vector de covariables para la  $i$ -th observación, escalado and centrada, y  $\underline{k}_j$  corresponde a el  $j$ -th centroe obtenido por el método de clustering jerarquico. Note que incluimos la forma lineal  $Zu_{(i)}$  para intentar capturar las interacciones entre las covariables.

$$Z_{ij} = \exp \left( \|\underline{x}_i - \underline{k}_j\| \right)^2\tag{4.2}$$

### 4.2.3. Estimación por máxima verosimilitud

Para una muestra de  $n$  observaciones:  $\underline{x} = (x_1, \dots, x_n)$ , el estimador de máxima verosimilitud para los parámetros de la distribución GEV no estacionaria, puede ser determinado por maximizar la función de verosimilitud, dada por la forma general:

$$\begin{aligned}L(\mu_t, \sigma_t, \kappa_t | \underline{x}) &= \prod_{t=1}^{n_1} \frac{1}{\sigma_t} \exp \left\{ - \left[ 1 + \kappa_t \left( \frac{x_t - \mu_t}{\sigma_t} \right) \right]^{-\frac{1}{\kappa_t}} \right\} \times \left[ 1 + \kappa_t \left( \frac{x_t - \mu_t}{\sigma_t} \right) \right]^{-\left(1 + \frac{1}{\kappa_t}\right)} \\ &\times \prod_{t=n_1+1}^n \frac{1}{\sigma_t} \exp \left\{ - \left( \frac{x_t - \mu_t}{\sigma_t} \right) \right\} \exp \left\{ - \exp \left[ - \left( \frac{x_t - \mu_t}{\sigma_t} \right) \right] \right\}.\end{aligned}$$

Donde  $n_1$  es el número de observaciones tales que  $\kappa_t \neq 0$ . La función sobre  $\kappa_t$  se asume generalmente constante,  $\kappa_t = \kappa$  como se recomienda en [Yee y Stephenson \(2007\)](#), así  $n_1 = n$  y la log-verosimilitud es:

$$\ell(\mu_t, \sigma_t, \kappa | \underline{x}) = -n \log \sigma_t - \sum_{t=1}^n \left[ 1 + \kappa \left( \frac{x_t - \mu_t}{\sigma_t} \right) \right]^{-\frac{1}{\kappa}} - \sum_{t=1}^n \left( 1 + \frac{1}{\kappa} \right) \log \left[ 1 + \kappa \left( \frac{x_t - \mu_t}{\sigma_t} \right) \right]$$

Haciendo  $C = [X \ Z]$  y  $\underline{b}'_{(i)} = \left[ \underline{\beta}'_{(i)} \ \underline{u}'_{(i)} \right]$ , donde  $C$  es una matriz  $n \times p$ , con  $p = p_1 + p_2$ ;  $\underline{b}'_{(i)}$  es un vector de  $p \times 1$  parámetros, entonces los predictores lineales pueden ser escritos:

$$\mu_t = Cb_1 ; \log \alpha_t = Cb_2 ; \kappa_t = \kappa$$

## 4.2. Metodología

Para el presente trabajo, asignamos una penalización ( $P$ ) al vector de parámetros, tal que:

$$P = \begin{bmatrix} \frac{1}{\sigma_\beta^2} \otimes I_{p_1} & 0 \\ 0 & \frac{1}{\sigma_u^2} \otimes I_{p_2} \end{bmatrix}$$

Donde  $I_{p_1}$  y  $I_{p_2}$  son matrices identidades de orden  $p_1$  y  $p_1$  respectivamente,  $\sigma_\beta^2$  y  $\sigma_u^2$  son valores que controlan el grado de regularización del modelo.

De tal forma que la log-verosimilitud del modelo queda de la siguiente manera:

$$\ell_n^p(\mu_t, \sigma_t, \kappa \mid \underline{x}) = \ell_n(\underline{x} \mid \mu_t, \sigma_t, \kappa) - \sum_{i=1}^2 \underline{b}'_{(i)} P \underline{b}_{(i)} - \frac{1}{\sigma_\kappa^2} \kappa^2$$

Nótese que también podemos reescribir la verosimilitud como:

$$\ell_n(\mu_t, \sigma_t, \kappa \mid \underline{x}) = \sum_{t=1}^n \ell_t(\mu_t, \sigma_t, \kappa \mid \underline{x}) - \sum_{i=1}^2 \underline{b}'_{(i)} P \underline{b}_{(i)} - \frac{1}{\sigma_\kappa^2} \kappa^2$$

Donde:

$$\ell_t(\mu_t, \sigma_t, \kappa \mid \underline{x}) = -\log \sigma_t - \left[ 1 - \kappa \left( \frac{x_t - \mu_t}{\sigma_t} \right) \right]^{\frac{1}{\kappa}} \log \left[ 1 - \kappa \left( \frac{x_t - \mu_t}{\sigma_t} \right) \right]$$

Así el gradiente de la verosimilitud está dado por:

$$\begin{aligned} \frac{\partial \ell_n^p(\mu_t, \sigma_t, \kappa \mid \underline{x})}{\partial b_1} &= \sum_{t=1}^n \frac{\ell_t(\mu_t, \sigma_t, \kappa \mid \underline{x})}{\partial \mu_t} \otimes \underline{c}_t - 2P \underline{b}_{(1)} \\ \frac{\partial \ell_n^p(\mu_t, \sigma_t, \kappa \mid \underline{x})}{\partial b_2} &= \sum_{t=1}^n \frac{\ell_t(\mu_t, \sigma_t, \kappa \mid \underline{x})}{\partial \sigma_t} \otimes \underline{c}_t - 2P \underline{b}_{(2)} \\ \frac{\partial \ell_n^p(\mu_t, \sigma_t, \kappa \mid \underline{x})}{\partial \kappa} &= \sum_{t=1}^n \frac{\ell_t(\mu_t, \sigma_t, \kappa \mid \underline{x})}{\partial \kappa} - \frac{2\kappa}{\sigma_\kappa^2} \end{aligned}$$

Aquí, el operador  $\otimes$  es el producto Kronecker, y la matriz Hessiana esta dada por:

$H = \sum_{t=1}^n H_{(t)}$ , donde:

$$H_{(t)} = \begin{bmatrix} H_{1(t)} & H'_{2(t)} \\ H'_{2(t)} & H_{3(t)} \end{bmatrix}$$

y

$$\begin{aligned} H_{1(t)} &= \begin{bmatrix} \frac{\partial \ell_t(\mu_t, \sigma_t, \kappa \mid \underline{x})}{\partial \mu_t^2} & \frac{\partial \ell_t(\mu_t, \sigma_t, \kappa \mid \underline{x})}{\partial \mu_t \partial \sigma_t} \\ \frac{\partial \ell_t(\mu_t, \sigma_t, \kappa \mid \underline{x})}{\partial \mu_t \partial \sigma_t} & \frac{\partial \ell_t(\mu_t, \sigma_t, \kappa \mid \underline{x})}{\partial \sigma_t^2} \end{bmatrix} \otimes \underline{c}_t \underline{c}'_t - (I_4 \otimes 2P) \\ H_{2(t)} &= \begin{bmatrix} \frac{\partial \ell_t(\mu_t, \sigma_t, \kappa \mid \underline{x})}{\partial \mu_t \partial \kappa} & \frac{\partial \ell_t(\mu_t, \sigma_t, \kappa \mid \underline{x})}{\partial \sigma_t \partial \kappa} \end{bmatrix} \otimes \underline{c}'_t \end{aligned}$$

## 4.2. Metodología

---

$$H_{3(t)} = \frac{\partial \ell_t(\mu_t, \sigma_t, \kappa \mid \underline{x})}{\partial \kappa^2} - \frac{2}{\sigma_\kappa^2}$$

Como valores iniciales para nuestro algoritmo se toman los estimadores de máxima verosimilitud del modelo de valores extremos estacionario.

Una forma simple de comparar la validez de un modelo comparado con otro, uno anidado dentro otro es aproximadamente  $\chi^2$ , con los grados de libertad igual a la diferencia de parámetros entre ambos modelos (Coles, 2001). Como medida de comparación entre modelos no anidados, una medida útil es el criterio de información de (Akaike, 1974) (AIC), dado por:

$$AIC = 2k - 2\log(L)$$

Donde  $k$  es el número de parámetros estimados en el modelo, y  $L$  es el valor de la función de log-verosimilitud evaluada en los parámetros estimados.

### 4.2.4. Elección del Kernel

Un kernel es una función  $T$ , que mapea un espacio vectorial  $X$  en otro espacio vectorial  $Y$ , generalmente de mayor dimensión. En este trabajo proponemos usar el kernel gaussiano, el cual transforma el vector de entrada  $x_i \in R^n$  en un vector de salida  $z_i \in R^m$  por medio de la siguiente expresión

$$z_{ij} = \exp\left(-\frac{\|x_i - \underline{k}_j\|^2}{2\sigma^2}\right)$$

Donde  $\underline{k}_j$  es un vector arbitrario de la dimensión de  $x_i$  y  $j = 1, \dots, m$ . Una posible forma de elegir a  $\underline{k}_j$  es por medio de los centroides encontrados en un algoritmo de clustering sobre la observaciones  $x_i$ .

Bocci *et al.* (2013) llevo a cabo un análisis de valores extremos a máximos de lluvias empleando el kernel ThinPlates. En este kernel, la transformación de un conjunto de observaciones  $X$ , con  $x_i \in R^n$  a otro conjunto  $Z$ , con  $z_i \in R^m$  es por medio de la siguiente igualdad

$$Z = [C(x_i - k_j)]_{1 \leq i \leq n; 1 \leq j \leq m} \cdot [C(k_h - k_j)]_{1 \leq h, j \leq m}^{-1/2}$$

$$C(v) = \|v\|^2 \log \|v\|$$

### 4.3. Estudio con datos sintéticos

Este kernel tiene la desventaja de que para su calculo necesitamos evaluar una inversa y en algunos casos, esta puede no existir.

### 4.3. Estudio con datos sintéticos

Para ilustrar el desempeño del modelo de valores extremos (GEV) propuesto se utilizan datos sintéticos. Realizamos 500 generamos datos del modelo GEV con diferentes tamaños de muestra ( $n = 25, 50, 100, 500$ ) con dos covariables,  $x_1$  y  $x_2$ . Los valores de las covariables  $x_1$  son generados de datos igualmente espaciados en el intervalo de 0 to  $4\pi$ , los valores de la covariable son generados de la  $x_2 \sim Uniforme(0, 4\pi)$ , con  $\mu_t = 0$ , gráfícados en a y b en la figura 4.1.

$$\log(\sigma_t) = \frac{1}{2\pi |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (x_t - \phi)^t \Sigma^{-1} (x_t - \phi)\right) \quad (4.3)$$

Donde  $x_i = (x_{1(t)}, x_{2(t)})$ ,  $\phi = (6, 6)$  y  $\Sigma = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$  y  $\kappa = -0.1$ , denotan el modelo verdadero.

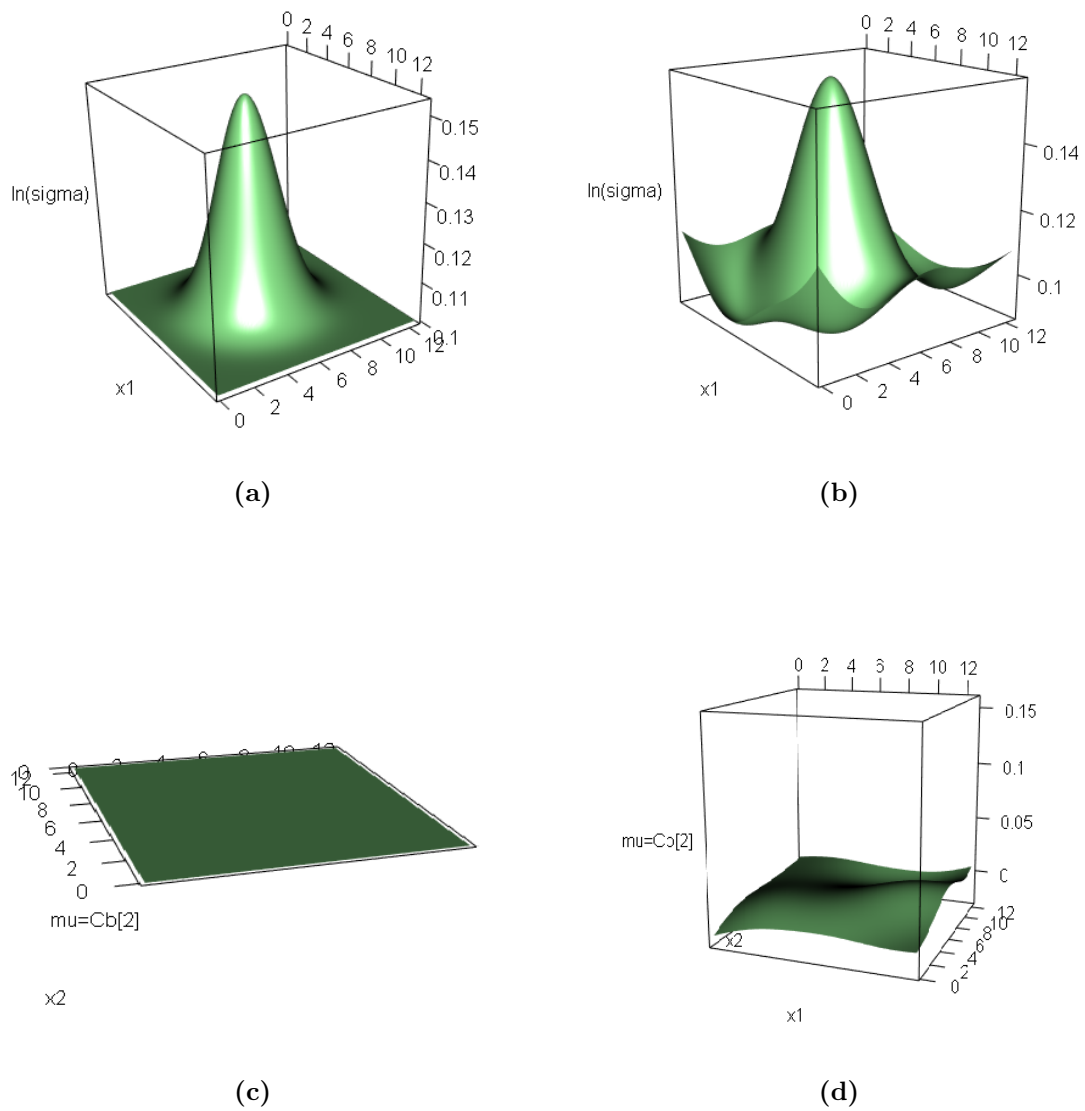
Los resultados de la simulación son mostrados en la Tabla 4.1, donde se pueden observar los comportamientos de nuestros estimadores al cambiar el tamaño de muestra, para ( $n = 25, 50, 100, 500$ ). Para cada muestra simulamos 500 ejecuciones y ajustamos el modelo 4.3. El sesgo y el error estándar pueden ser calculados con la media muestral y la desviación estándar de las estimaciones. Los resultados de la Tabla 4.1 muestran la tendencia esperada, es decir, conforme aumenta el tamaño de la muestra, el sesgo y la desviación estándar disminuyen, evidenciando que, nuestros estimadores son consistentes.

**Tabla 4.1:** Resultados de la simulación del sesgo y la desviación estándar de las estimaciones a partir de un tamaño de muestra n de observaciones GEV.

	Sample size			
	$n = 25$	$n = 50$	$n = 100$	$n = 500$
Bias ( $\hat{\kappa}$ )	1.7205357	1.5960004	0.09572314	0.06465314
Bias( $b$ )	-0.132707	-0.02537	0.03449517	0.00609907
SD ( $\hat{\kappa}$ )	2.1917617	2.1948049	0.00552024	0.01146129
SD ( $b$ )	0.3450831	0.8789689	0.55762510	0.19040480

Para demostrar la ventaja de nuestro modelo frente al modelo actual VGAM Yee y Stephenson (2007), en situaciones donde el modelo depende de la interacción entre variables, se ajustaron 500 nuevas realizaciones del modelo 4.3, para cada uno de los 2

### 4.3. Estudio con datos sintéticos



**Figura 4.1:** Función real (a,c) y ajustada (b,d) a los parámetros de datos simulados con  $n = 3000$  de un modelo GEV no estacionario.

#### 4.4. Valores extremos de máximos de lluvia

modelos, con 10 nodos totales en ambos modelos. Se calculó la media del AIC de los 500 ajustes a las realizaciones, resultando un AIC promedio de  $-4332$  para el modelo VGAM y un AIC promedio de  $-4336$  para nuestro modelo, demostrando que nuestro modelo es ligeramente mejor.

#### 4.4. Valores extremos de máximos de lluvia

En este ejemplo se ilustra el modelo GEV no estacionario para el conjunto de datos de máximos de lluvia obtenidos del paquete de *R* package SpatialExtremes Ribatet *et al.* (2008). Estos datos corresponden a los máximos de lluvia entre los años de 1962 al 2008 durante los meses de junio a agosto en 79 sitios de Suiza. Las covariables fueron la coordenada  $x_1$ , coordenada  $x''$  (longitud);  $x_2$ , coordenada  $y''$  (latitud), y el  $x_3$  tiempo.

Los resultados son mostrados para el modelo:



**Figura 4.2:** Mapa de Suiza y la distribución espacial de las estaciones estudiadas.

Resultados mostrados para el modelo:

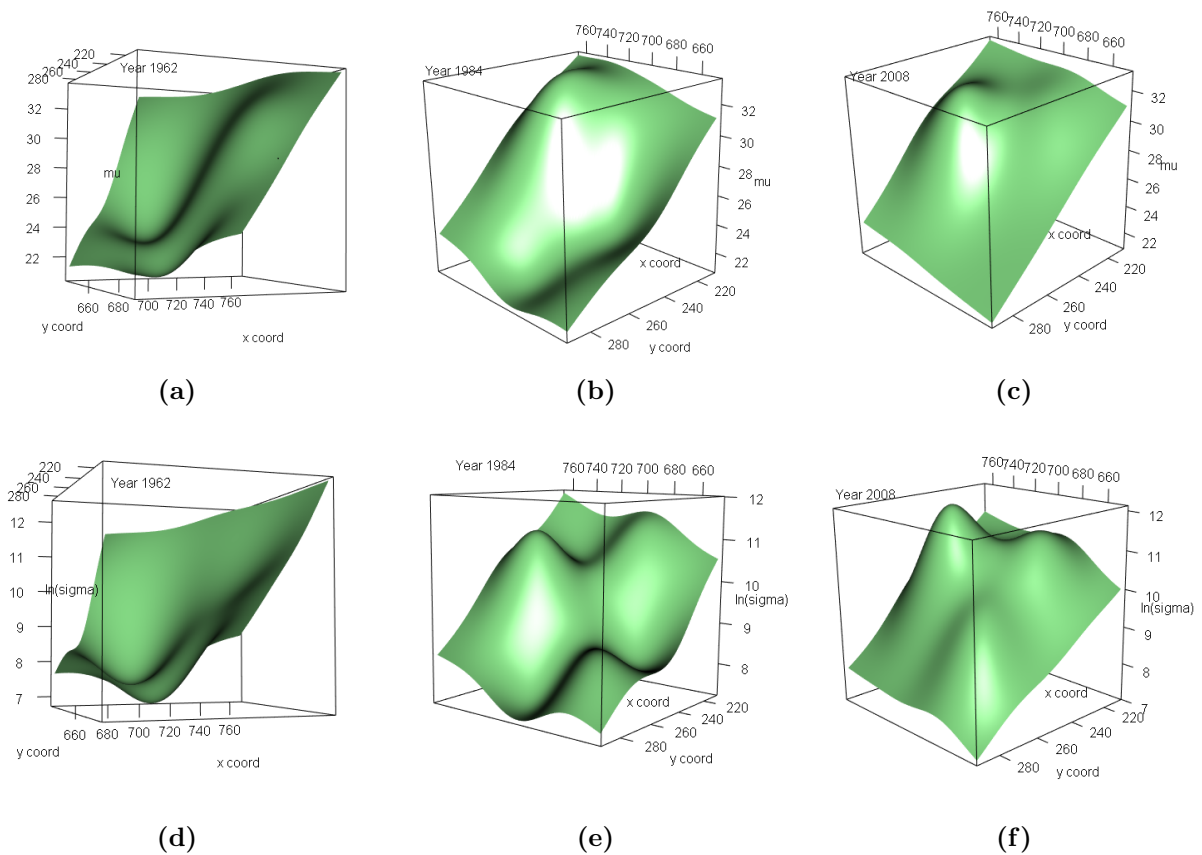
$$\begin{aligned}\mu_t &= \beta_{(1)0} + \beta_{(1)1}x_{t1} + \beta_{(1)2}x_{t2} + \beta_{(1)3}x_{t3} + \sum_{j=1}^6 u_{(1)j}z_{tj} \\ \log \sigma_t &= \beta_{(2)0} + \beta_{(2)1}x_{t1} + \beta_{(2)2}x_{t2} + \beta_{(2)3}x_{t3} + \sum_{j=1}^6 u_{(2)j}z_{tj} \\ \kappa &= \beta_{(3)0}\end{aligned}$$

Para el cálculo  $z_i$ , elegimos seis nodos obtenidos como los centroides de las variables es-

#### 4.4. Valores extremos de máximos de lluvia

tandarizadas. Para comparar nuestro modelo elegimos un modelo VGAM con funciones suavizadas spline con seis nodos para cada una de las tres variables, sumando en total 12 nodos.

Los resultados son mostrados en la figura 4.3. El estimador del parámetro de forma fue de  $-0.12$ , con una desviación estándar de  $0.013$ , como medida de comparación entre modelos, el AIC del modelo propuesto fue de  $29,132$  y el AIC del modelo VGAM es de  $29,134$ , lo cual indica que nuestro modelo es ligeramente mejor. Por otra parte el AIC del modelo estacionario es  $29,837$ , lo que indica que existe una fuerte evidencia de que el modelo no es estacionario. Analizando la figura 4.3 podemos observar que a partir del año 1984, hubo una inversión de la tendencia de los máximos en la mayor parte de la región espacial estudiada. También se puede observar que la variabilidad no ha permanecido constante en el periodo de tiempo estudiado.



**Figura 4.3:** Funciones estimadas para el parámetro de localidad (*a, b, c*) y escala (*d, e, f*) para la región de estudio en 3 periodos de tiempo (1962, 1984, 2008). Los ejes *x* y *y* en cada gráfica representan la región geografica y los valores de los parámetros están representados por la superficie.

### 4.5. Discusión

En este capítulo se presentó una nueva manera de analizar los eventos máximos no estacionarios, implementando una regresión semiparamétrica sobre el parámetro de localización y escala, similar al trabajo hecho por (Yee y Stephenson, 2007). Sin embargo, debido a la forma particular de la regresión, nuestro modelo es capaz de incluir en el modelo de forma implícita el efecto de interacción entre variables, a través de un modelo lineal, que facilita el cálculo de las operaciones de maximización del modelo.

Los resultados de nuestro método son comparables con los modelos aditivos VGAM, quienes muestran resultados muy similares; sin embargo en nuestro modelo incluimos implícitamente el efecto de interacción entre covariables. Un enfoque similar fue el empleado por (Cannon, 2010) quien modeló los parámetros de la distribución con redes neuronales. No obstante, ese método depende de los valores iniciales y generalmente no siempre encuentra la solución óptima global.

Por otra parte, nuestro método, nos permite una estimación más flexible, y no se limita a determinadas formas funcionales en el predictor de cada parámetro, como en la mayoría de los modelos actuales El Adlouni *et al.* (2007).

En el estudio de simulación, se observó que los estimadores obtenidos con nuestro método son consistentes. Además se obtuvo, de forma aproximada, la forma original de la función usada para simular los parámetros. Sin embargo, debido a que usamos un enfoque de máxima verosimilitud, el ajuste no es muy preciso para muestras pequeñas.

El modelo presentado en este trabajo constituye una herramienta eficaz en el análisis de valores extremos para el caso no estacionario, en situaciones donde los parámetros de la distribución son funciones de covariables, ya sea temporales o espaciales.



# Capítulo 5

## Análisis bayesiano de valores extremos con datos censurados

### 5.1. Introducción

El análisis de extremos con censura aleatoria es un campo relativamente nuevo ([Einmahl \*et al.\*, 2008](#)) y se ha enfocado principalmente en la estimación de algunos estadísticos de la distribución de valores extremos (GEV) ([Gomes y Neves, 2011](#)). Nosotros tratamos de vincular estas metodologías para mejorar la estimación en la distribución GEV en los casos en los que se presenta censura y así obtener conclusiones más acordes a la realidad.

Por otra parte, recientemente se han venido trabajando novedosas formas de trabajar con los valores extremos, principalmente en el área de los modelos predicción por medio de suavizamientos como es el caso de los modelos VGLM ([Yee y Stephenson, 2007](#)), modelos mixtos ([Laurini y Pauli, 2009](#), [Padoan y Wand, 2008](#)), modelos geoaditivos ([Bocci \*et al.\*, 2013](#)), entre otros.

En este capítulo, proponemos usar kernels gaussianos para generar las bases de nuestro algoritmo de suavizado, similar al trabajo presentado en [Bocci \*et al.\* \(2013\)](#) en donde utilizaron ThinPlates para obtener las bases para suavizar el parámetro de localidad de la distribución GEV.

### 5.2. Metodología

En orden para determinar el impacto de los covariables en los parámetros de la distribución de valores extremos no estacionarios con datos censurados, hemos visto que esto

## 5.2. Metodología

---

puede hacerse usando un kernel gaussiano aplicado a la norma de la diferencia entre las observaciones y los nodos encontrados con el método cluster. En este sentido, realizamos un primer ajuste usando el conjunto completo de covariables, y lo implementamos en el modelo 1; seguidamente realizamos un segundo ajuste dividiendo el conjunto completo de covariables en dos conjuntos, separando a las covariables espaciales del resto y este enfoque lo trabajamos en el modelo 2.

Así el modelo 1 consiste en un modelo jerárquico bayesiano, donde la función que liga a los parámetros con las covariables es:

$$\begin{aligned}\mu_i &= X\beta_{(1)} + Zu_{(1)}, \\ \log \sigma_i &= X\beta_{(2)} + Zu_{(2)}, \\ \kappa_i &= \kappa\end{aligned}\tag{5.1}$$

Donde  $X$  es una matrix  $n \times p_1$  de covariables escaladas y centradas incluyendo el intercepto,  $\beta_{(i)}$   $i=1,2$  es un vector de tamaño  $p_1$ ,  $u_{(i)}$   $i=1,2$  es un vector de tamaño  $p_2$ ,  $Z$  es una matrix de tamaño  $n \times p_2$  tal que  $Z_{ij} = \exp(\|\underline{x}_i - \underline{k}_j\|)^2$ ,  $\underline{x}_i$  es el vector de covariables para la  $i$ -th observación, escalada y centrada, y  $\underline{k}_j$  es el  $j$ -th centroide obtenido por el método de hierarchical clustering.

La densidad a posteriori del modelo es:

$$\pi(\omega|y_t) \propto \pi(y_t|\omega) \pi(\omega)$$

Donde  $\pi(y_t|\omega)$  es la densidad GEV bajo el condiciones sobre los parámetros 5.1 y  $\omega = (\beta_1, \beta_2, u_1, u_2)$ . Este modelo es equivalente a un modelo jerárquico donde el nivel de los datos esta dado por  $\pi(y_t|\theta_t)$  y el nivel de la apriori es  $\pi(\omega)$ . La distribución apriori de  $\omega$  es tal que la distribución apriori para  $\beta$ 's es no informativa y

$$u's \sim N(0, \sigma_u)$$

$$\kappa \sim N(0, \sigma_\kappa)$$

Para muestrear de la distribución posteriori, usamos el método de random walks metropolis.

Por otro lado, en el modelo 2, separamos la componente espacial por medio del término  $Z_s$ , así, la función que liga a los parámetros con las covariables es:

## 5.2. Metodología

---

$$\begin{aligned}
 \mu_i &= X\beta_{(1)} + Z_x u_{(1)x} + Z_s u_{(1)s}, \\
 \log \sigma_i &= X\beta_{(2)} + Z_x u_{(2)x} + Z_s u_{(2)s}, \\
 \kappa_i &= \kappa
 \end{aligned}
 \tag{5.2}$$

Donde  $X$  es una matriz  $n \times p$  de covariables escaladas y centradas incluyendo el intercepto,  $\beta_{(i)}$   $i=1,2$  es un vector de tamaño  $p$ ,  $u_{(i)x}$   $i=1,2$  es un vector de tamaño  $K_x$ ,  $Z_x$  es una matriz de  $n \times K_x$  tal que  $\{Z_x\}_{ij} = \exp\left(\frac{\|\underline{x}_i - \underline{k}_j\|^2}{2}\right)$ ,  $i = 1, \dots, n$ ;  $j = 1, \dots, K_x$ ,  $\underline{x}_i$  el vector de covariables para el  $i$ -th observación, escaladas y centradas, y  $\underline{k}_j$  el  $j$ -th centroide obtenido por el método de hierarchical clustering;  $Z_s$  es una matriz  $n \times K_s$  de bases para el vector de coordenadas para la  $i$ -th observación.

La función de densidad a posteriori es:

$$\pi(\omega^*, \omega^{**} | y_t) \propto \pi(y_t | \omega^*) \pi(\omega^* | \omega^{**}) \pi(\omega^{**})$$

Donde  $\pi(y_t | \omega^*)$  es la densidad GEV. La distribución a priori  $\omega^*$  es tal que

$$\begin{aligned}
 \beta_1 &\sim N(0, 10^4 I) \\
 \beta_2 &\sim N(0, 10^4 I) \\
 u_{(x1)} | \sigma_{x1} &\sim N(0, \sigma_{x1} I_{K_{x1}}) \\
 u_{(x2)} | \sigma_{x2} &\sim N(0, \sigma_{x2} I_{K_{x2}}) \\
 u_{(s1)} | \sigma_{s1} &\sim N(0, \sigma_{s1} I_{K_{s1}}) \\
 u_{(s2)} | \sigma_{s2} &\sim N(0, \sigma_{s2} I_{K_{s2}}) \\
 \kappa &\sim Uniform(-5, 5)
 \end{aligned}$$

Finalmente la distribución a priori para los hiperparametros  $\omega^{**}$  es:

$$\begin{aligned}
 \sigma_{x1} &\sim Half - Cauchy(25) \\
 \sigma_{x2} &\sim Half - Cauchy(25) \\
 \sigma_{s1} &\sim Half - Cauchy(25) \\
 \sigma_{s2} &\sim Half - Cauchy(25)
 \end{aligned}$$

## 5.3. Datos Sintéticos

En este ejemplo presentamos el modelo GEV con datos censurados usando datos simulados. Simulamos 500 valores extremos de un modelo GEV con dos covariables  $x_1$  y  $x_2$ , usamos diferentes proporciones de censura. Los valores de la covariable  $x_1$  son generados de datos igualmente espaciados en el intervalo de 0 a  $4\pi$ , los valores de la covariable son muestras independientes tal que  $x_2 \sim Uniform(0, 4)$ , con

$$\begin{aligned} \mu_t &= \sin \left( \frac{\left( \sqrt{(x_1 - 2\pi)^2 + (x_2 - 2\pi)^2} \right)^{\frac{3}{4}}}{5} \right) \\ \sigma_t &= \sigma = 0.3 \\ \kappa_t &= \kappa = -0.5 \end{aligned} \tag{5.3}$$

Primeramente presentamos los resultados del Modelo 1, donde ajustamos un modelo para los niveles de censura de 0, 5, 10 y 15 por ciento. Para cada modelo corrimos 200,000 iteraciones para obtener muestras de la distribución posteriori, después de un quemado de 50,000, y adelgazamos la cadena, manteniendo 1 de cada 5 iteraciones consecutivas. Los resultados para el valor estimado del parámetro  $\kappa$ , así como sus correspondientes intervalos de credibilidad al 95%, son mostrados en la tabla 5.1. De la tabla observamos que todos los intervalos con diferentes grados de censura contienen al verdadero valor, además observamos que el sesgo y la longitud del intervalo de credibilidad aumenta cuando se incrementa el porcentaje de datos censurados.

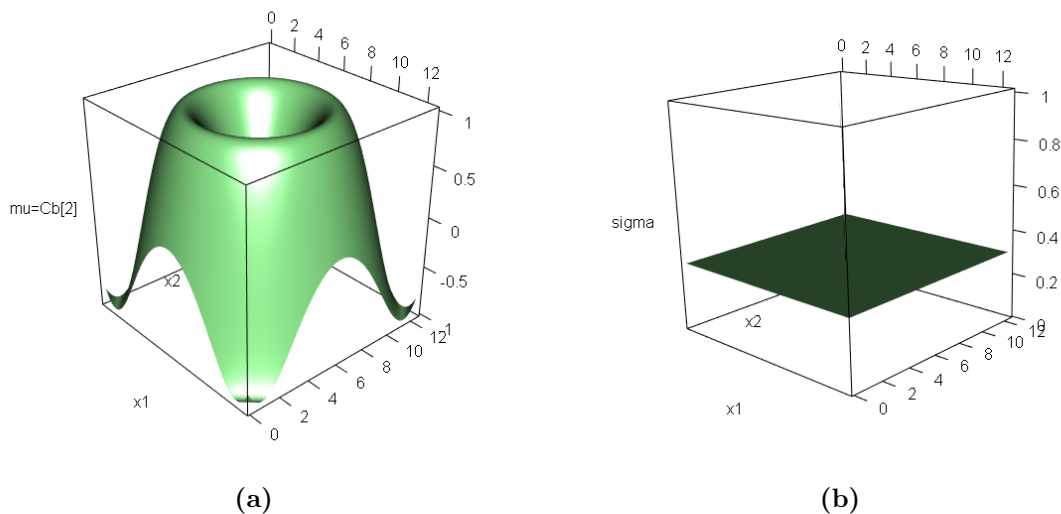
**Tabla 5.1:** Estimadores y 95% intervalo de credibilidad para  $\kappa$  (con valor verdadero de -0.5), por nivel de censura.

% Censura	Mean	95 % CI
0	-0.5013227	(-0.5992741, -0.4083045)
5	-0.5283462	(-0.6050321, -0.4647396)
10	-0.5163329	(-0.6000799, -0.3636123)
15	-0.5473067	(-0.5923786, -0.4960348)

La figura 5.1 muestra los valores de  $\mu$  y  $\sigma$  en función de las covariables  $x_1$  y  $x_2$ . Note que la función para  $\mu$  corresponde a una función que no puede separarse únicamente en efectos principales de las covariables, por lo que no puede ajustarse con la mayoría de los modelos tradicionales para valores extremos, la función para  $\sigma$  corresponde a un plano en el espacio de las covariables.

La figura 5.2 muestra la función estimada para el parámetro de localidad, en el plano de las covariables  $x_1$  y  $x_2$ , para cada uno de los niveles de censura de 0, 5, 10 y 15%. La forma original del parámetro de localidad, es recuperada por el modelo, inclusive para niveles del 15% de censura.

### 5.3. Datos Sintéticos



**Figura 5.1:** Función real para  $\mu$  (a) y  $\sigma$  (b) usados en el estudio de simulación.

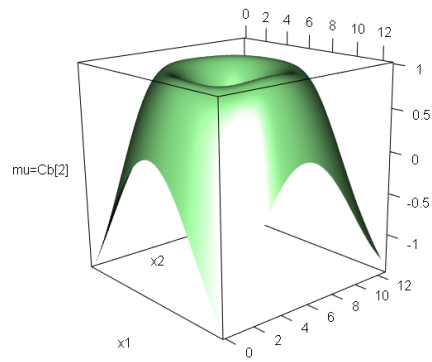
A fin de evaluar el rendimiento de nuestro modelo, dejamos fuera del ajuste a un grupo de 1000 datos considerados como datos de prueba, y después utilizamos estos datos para evaluar la correlación con sus valores predichos. Específicamente realizamos esta validación para el parámetro de localidad, donde obtenemos una correlación de 0.99 entre los valores predichos por nuestro modelo y el grupo considerado como datos de prueba.

Las estimaciones para la función del parámetro de escala es mostrada en la figura 5.3, en este caso, tenemos que la estimación, presenta pequeñas irregularidades con una diferencia máxima en valor absoluto de 0.1 con respecto al valor real del parámetro las cuales se hacen más notorias en los bordes de las gráficas con mayor nivel de censura.

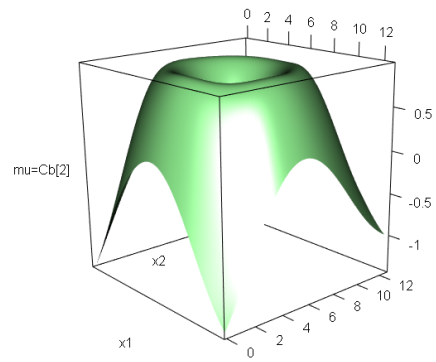
Con respecto al modelo 2, ajustamos el modelo para niveles de censura de 15 por ciento. Para nuestro modelo corrimos 250,000 iteraciones para obtener muestras de la densidad aposteriori, después de quemar 50,000 iteraciones y adelgazar cada 5 iteraciones. Los resultado para el parámetro  $\kappa$ , el intercepto para  $\mu$  y  $\sigma$  y su intervalo de credibilidad correspondiente al 95 %, son mostrados en la tabla 5.2. De la tabla 5.2 observamos que el intervalo para  $\kappa$  contiene el verdadero valor, note que el sesgo y la longitud de el intervalo de credibilidad es simétrico y no contiene al cero.

### 5.3. Datos Sintéticos

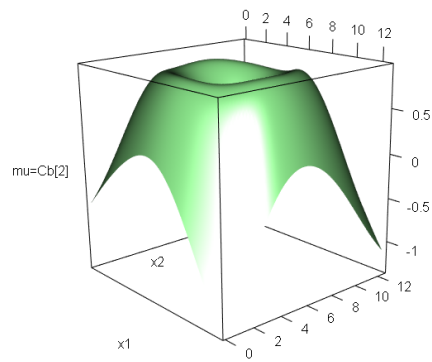
---



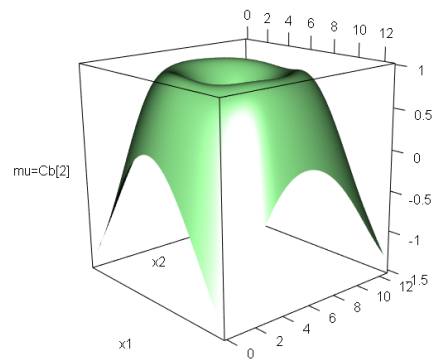
(a)



(b)



(c)

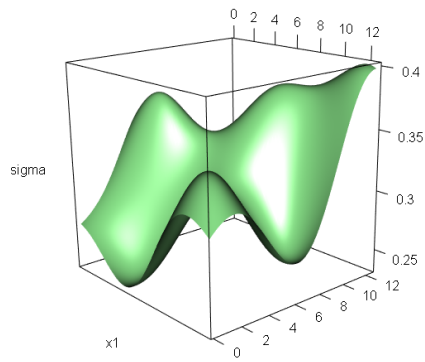


(d)

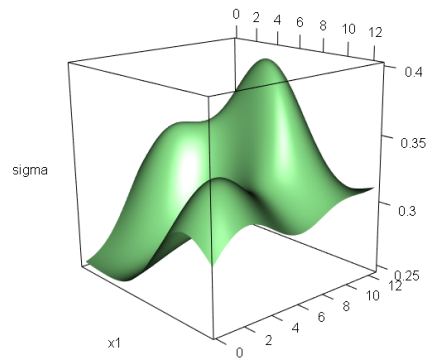
**Figura 5.2:** Funciones estimadas para el parametro de localidad en cada uno de los niveles de censura de 0, 5, 10 y 15 % mostradas en las figuras (a),(b),(c),(d) respectivamente.

### 5.3. Datos Sintéticos

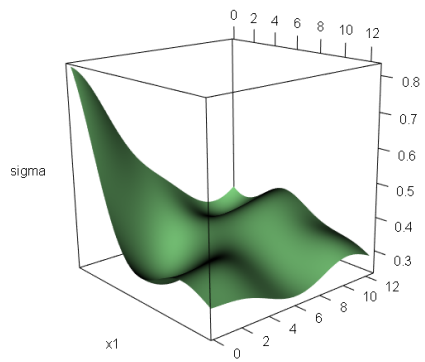
---



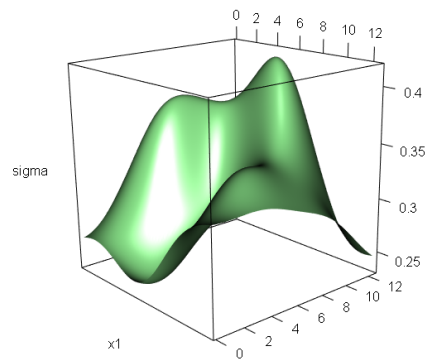
(a)



(b)



(c)



(d)

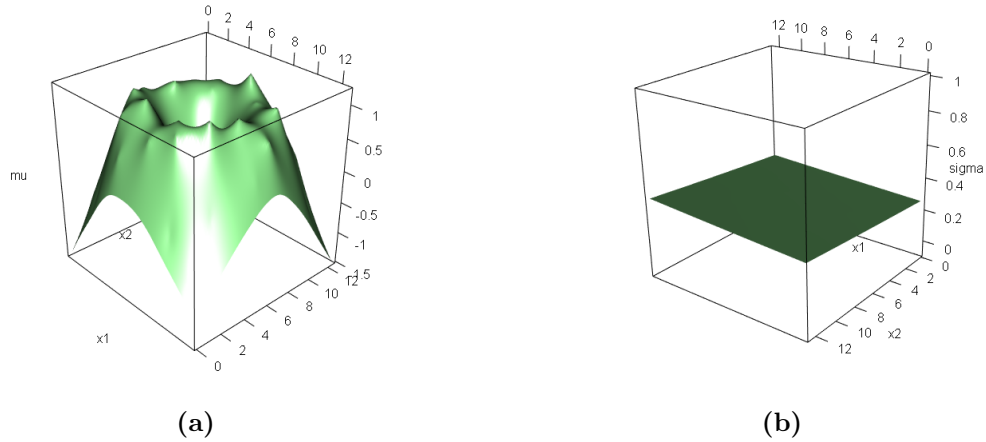
**Figura 5.3:** Funciones estimadas para el parámetro de escala en cada uno de los niveles de censura de 0, 5, 10 y 15% mostradas en las figuras (a),(b),(c),(d) respectivamente.

## 5.4. Niveles Máximos de contaminante PM10

**Tabla 5.2:** Estimadores y el intervalo de credibilidad al 95 % para  $\kappa$  (verdadero valor de -0.5), el intercepto de  $\mu$  y  $\log \sigma$ .

% Parameter	Mean	95 % CI
$\beta_{0,\mu}$	-3.280	(-3.876, -2.524)
$\beta_{0,\log(\sigma)}$	-1.159	(-1.254, -1.072)
$\kappa$	-0.498	(-0.580, -0.408)

La figura 5.4 muestra la función estimada para el parámetro de forma, en el plano de covariables  $x_1$  y  $x_2$ , para niveles de censura de 15 %. Observamos que la forma original del modelo es recuperada por el algoritmo propuesto, aunque se observan ligeros picos en los bordes de la figura, posiblemente debido a la inclusión de datos censurados. La estimación para la función estimada para el parámetro de escala parámetro son mostrados en la figura 5.4 (b), en este caso, observamos que los estimaciones corresponden a un plano ligeramente inclinado que recuperan la función original.



**Figura 5.4:** Funciones estimadas para el parámetro de localidad (a) y escala (b) para niveles de censura de 15 %.

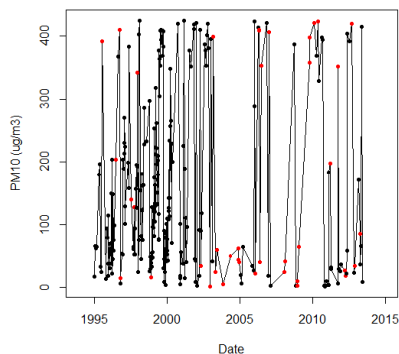
## 5.4. Niveles Máximos de contaminante PM10

En este ejemplo se ilustra el modelo GEV no estacionario para datos censurados para el conjunto de datos de máximos de contaminación atmosférica de partículas menores a 10 micrómetros (PM10). Estos datos corresponden a 1479 observaciones de niveles máximos de PM10, tomados de la base de datos de la Red Automática de Monitoreo Atmosférico



## 5.4. Niveles Máximos de contaminante PM10

(RAMA), entre los años de 1995 al 2014 en 11 estaciones ubicadas en la Zona Metropolitana del Valle de México, específicamente, Tlalnepanta (TLA), Xalostoc (XAL), Merced (MER), Pedregal (PED), Tultitlán (TLI), Villa de Flores (VIF), Tlahuác (TLA), Santa Úrsula (SUR), FES Acatlán (FAC), San Agustín (SAG) e Iztacalco (IZT). Estos datos contienen un 13.25 por ciento de observaciones censuradas. Las covariables fueron la longitud, latitud y tiempo.

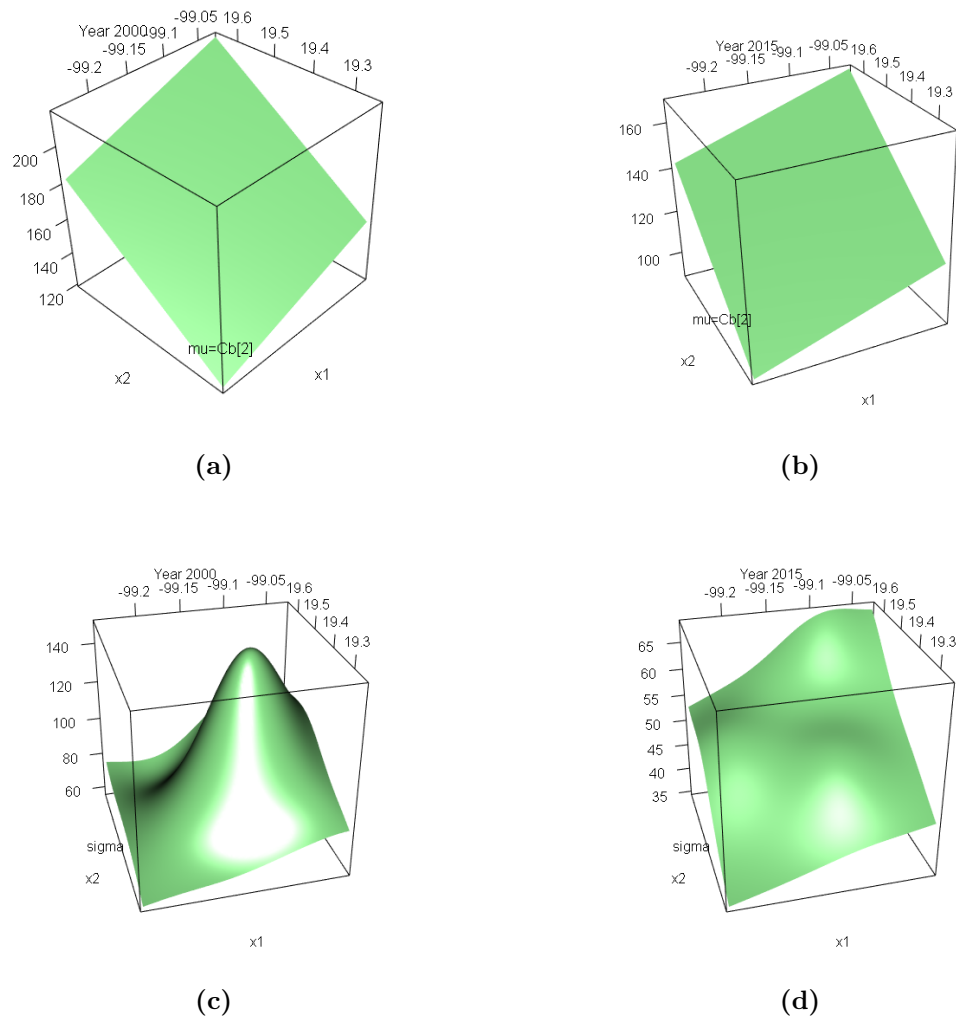


**Figura 5.5:** Valores extremos con datos censurados (puntos rojos) en la estación pedregal, Mex.

Nuevamente primero mostramos resultados del modelo 1. Los resultados para los parámetros de localidad y escala son mostrados en la 5.6. La función ajustada para  $\mu$  en el plano de coordenadas  $xy$ , muestran una tendencia espacial, de forma aproximadamente lineal para el parámetro de localidad, el cual tiende a incrementarse en la dirección noroeste, esta tendencia se ha mantenido invariante en los años 1995 y 2015. El parámetro de escala, también se incrementa en la misma dirección noroeste, pero a diferencia del parámetro de localidad, se observa un incremento en la zona ubicada entre las estaciones Merced y Xalostoc para el año 1995 pero cambia para el año 2015, en la cual el comportamiento se hace lineal, manteniendo la tendencia de aumentar en la dirección noroeste. el estimador para el parámetro de forma es 0.2427 y su intervalo de credibilidad al 95

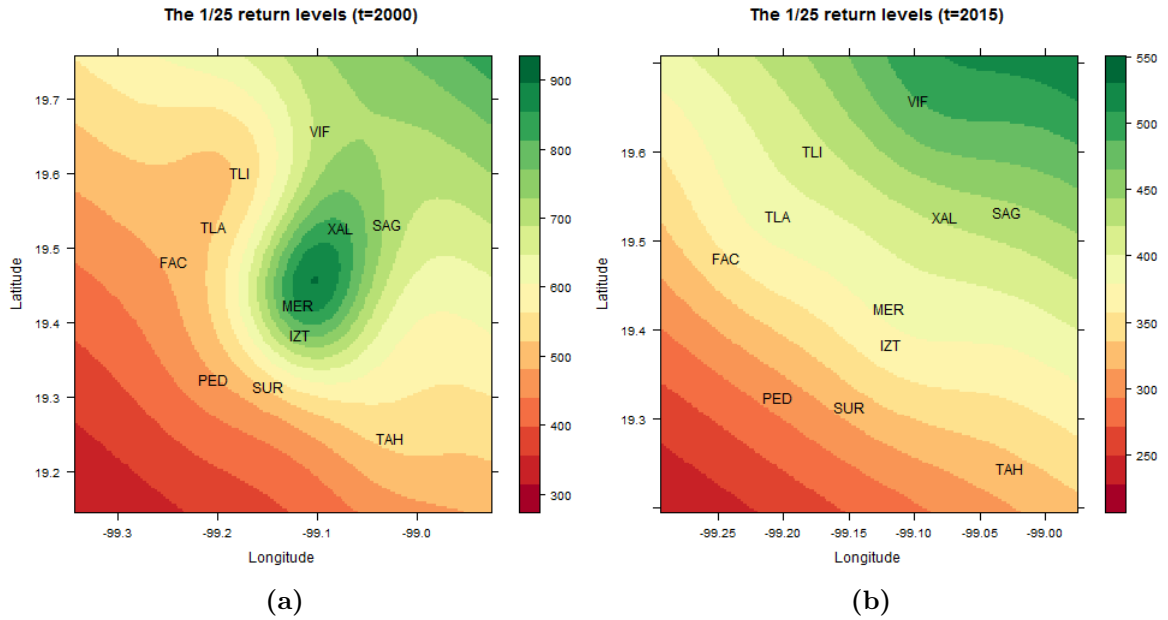
El umbral a partir del cual un valor extremo es excedido con probabilidad  $p$  es conocido como el nivel de retorno  $Z_p$ , el cual se espera que se exceda una vez cada  $1/p$  años Coles (2001). Aunque los niveles de retorno pierden su importancia cuando los procesos cambian en el tiempo Sang y Gelfand (2010), podemos ver en nuestro estudio que los niveles de retorno para el periodo de retorno de 25 años, cambian para cada uno de los dos años en la figura 5.7. De acuerdo a esto, el nivel de riesgo era mayor en el periodo posterior al año 1995 y este disminuyó para el año 2015, sin embargo permanece la tendencia de un riesgo mayor en la dirección noroeste de la zona de estudio.

## 5.4. Niveles Máximos de contaminante PM10



**Figura 5.6:** Funciones estimadas para el parámetro de localidad en el año 2000(a) y año 2015 (b) y escala en el año 2000(c) y año 2015 (d).

## 5.4. Niveles Máximos de contaminante PM10



**Figura 5.7:** Niveles de retorno con un periodo de retorno de 25 años para la región de estudio.

Con respecto a los resultados del Modelo 2. Los resultados para los parámetros de las funciones de escala y localización en el modelo 5.2, así como el parámetro de forma son mostradas en la tabla 5.3.

**Tabla 5.3:** Estimadores y 95 % intervalos de credibilidad para los parámetros en el modelo 4.2 usando los datos de máximos de PM10.

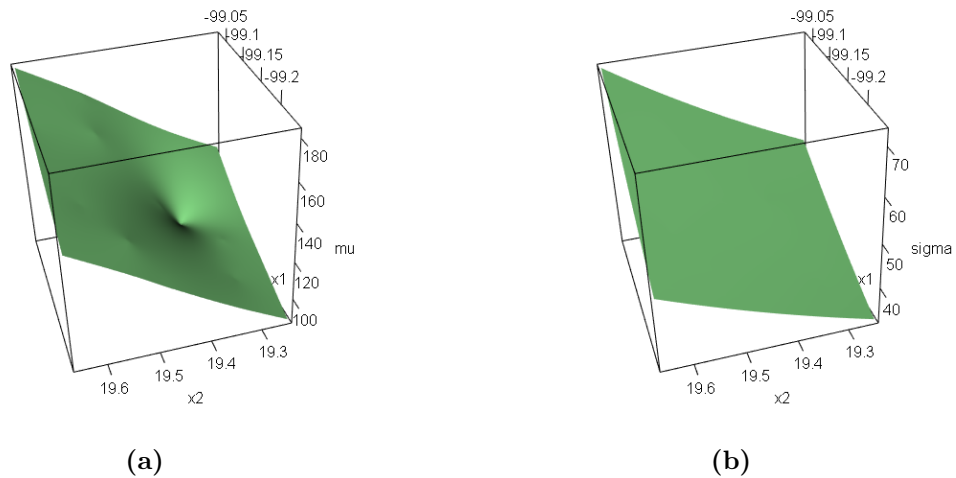
% Parametros	Media	95 % CI
$\beta_{(1)0}$	171.8	(150.6, 189.1)
$\beta_{(1)x}$	-19.51	(-28.56, -9.460)
$\beta_{(1)s1}$	11.17	(6.852, 15.90)
$\beta_{(1)s2}$	16.61	(12.12, 20.90)
$\beta_{(2)0}$	-0.160	(-0.340, 0.0512)
$\beta_{(2)x}$	-3.280	(-3.876, -2.524)
$\beta_{(2)s1}$	0.130	(0.083, 0.182)
$\beta_{(2)s2}$	0.113	(0.067, 0.152)
$\kappa$	0.272	(0.219, 0.337)
$\sigma_{(1)x}^2$	7.988	(1.601, 21.45)
$\sigma_{(2)x}^2$	$2.5e - 06$	$(3.0e - 07, 1.5e - 04)$
$\sigma_{(1)s}^2$	25.59	(8.556, 45.96)
$\sigma_{(2)s}^2$	0.020	(0.008, 0.042)

## 5.4. Niveles Máximos de contaminante PM10

Note que ninguno de los intervalos de credibilidad en la tabla 5.3 contienen al cero, excepto  $\beta_{(2)0}$  correspondientes al intercepto de  $\log \sigma$ . Recuerde que los parámetros contienen que  $\beta_{(1)}$  corresponden a  $\mu$ , y aquellos que contienen a  $\beta_{(2)}$  corresponden al  $\log \sigma$ .

La función de densidad a posteriori fue implementada en R 3.0.1, debido a que tenemos un tamaño de muestra relativamente grande, los  $K_x = 40$  y  $K_s = 30$ , el tiempo para obtener las 250,000 muestras de el algoritmo MCMC fue cerca de 30 horas.

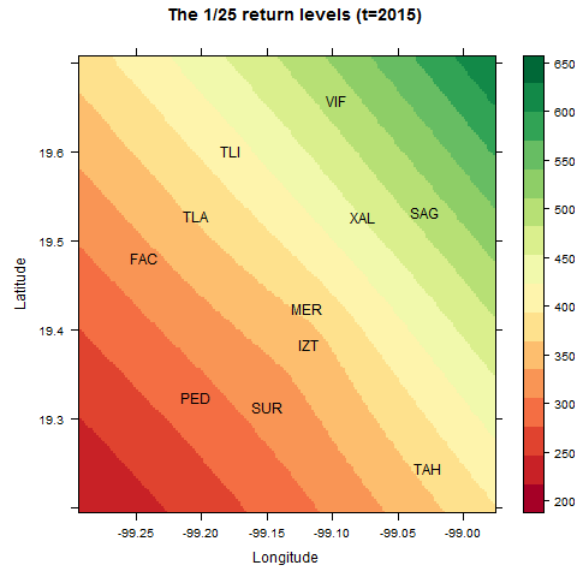
Los resultados para la funciones de localización y escala son mostrados en la figura 5.8. Las funciones estimadas para el parámetro de localización en el plano xy, muestran una tendencia espacial, aproximadamente lineal, la cual tiende a incrementarse en la dirección noroeste. El parámetro de escala se incrementa en la misma dirección, pero a diferencia de la función del parámetro de localidad, se observa un ligero decremento en el área cercana a Iztacalco, este parámetro presenta también un comportamiento lineal en el plano xy, manteniendo la tendencia de incrementarse hacia la región noroeste. El estimador del parámetro de forma es 0.2726 y su intervalo de credibilidad es 95 % (0.2198, 0.3378).



**Figura 5.8:** Funciones estimadas para el parámetro de localidad (a) y escala (b) en el año 2015.

De acuerdo a la figura 5.8, en los años 2015 permanece la tendencia de incrementarse en la región noroeste del área de estudio, alcanzando los más grandes niveles de riesgo en áreas cercanas a Villa Flores y San Agustín, y niveles más pequeños de riesgo en áreas alrededor de la estación Pedregal.

## 5.5. Discusión



**Figura 5.9:** Niveles de retorno con periodo de 25 años en la región de estudio.

## 5.5. Discusión

En este capítulo presentamos una metodología para estudiar valores extremos no estacionarios con datos censurados, realizamos un estudio de simulación y finalmente lo utilizamos para analizar datos de contaminación por partículas menores a 10 micrómetros (PM10), en el área metropolitana de la ciudad de México. El análisis lo realizamos usando un enfoque bayesiano semiparamétrico. Los resultados, muestran una clara tendencia espacial de incremento en los niveles máximos de PM10 en la dirección noroeste de la zona de estudio, así como también se observa una tendencia de cambio en los niveles máximos en el tiempo.

Concluimos que podemos usar esta metodología para generar mapas de riesgo de eventos extremos de lluvia, vientos, ondas de calor, etc. En particular cuando se tienen datos censurados que presentan información adicional y posiblemente multivariada, con el objetivo de medir sus efectos e implícitamente el de sus interacciones, para así poder encontrar relaciones entre estos y los valores extremos.

# Capítulo 6

## Conclusiones

El análisis de valores extremos es una herramienta para poder medir los riesgos que presentan los eventos desastrosos tales como las inundaciones, los vientos, las sequías y muchos otros fenómenos hidroclimatológicos e inclusive eventos económicos tales como pérdidas máximas, inflación, etc. que provocan las crisis y quiebras de grandes instituciones e inclusive, países. Es por ello, que se ha venido estudiando sistemáticamente a la par de otros grandes descubrimientos en el campo de la estadística.

En este trabajo, hemos estudiado y tratado de mejorar los modelos que se han venido trabajando, de tal forma que podamos solventar las limitaciones que presentan estos modelos de manera individual, y así intentamos proponer un análisis mas general en el estudio de los valores extremos. Particularmente introducimos, por un lado, una forma para tratar los datos censurado y por otro, mejorar la predicciones al usar los splines multivariados para explicar los parámetros de localización y de escala de la distribución GEV.

En los ejemplos de aplicación observamos que se pueden encontrar tendencias espacio temporales sobre los máximos y podemos generar mapas de riesgo, así como también algunas otras estadísticas útiles tales como mapas de probabilidades máximas de ocurrencia de los valores extremos, entre otras. Por otro lado, dada la forma del modelo, y más específicamente, en la parte correspondiente a los spline, notamos que estos hacen más sencillo el proceso de estimación y consecuentemente, hallar soluciones que nos permiten recuperar en detalle la forma en que se comportan los parámetros de la distribución, y de esta podemos hacer predicciones más precisas de acuerdo a las circunstancias y características del lugar que se esta estudiando.

Finalmente podemos concluir que el modelo presentado en este trabajo, constituye una herramienta eficaz en el análisis de valores extremos para el caso no estacionario, con posible información censurada y en situaciones en las cuales los parámetros de la distribución son funciones de covariables temporales, espaciales, etc.

# Referencias

- Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 19, 716-723.
- Balkema, A. A. y De Haan, L. (1974). Residual life time at great age. *The Annals of probability*, 792-804.
- Bocci, C., Caporali, E. y Petrucci, A. (2013). Geoadditive modeling for extreme rainfall data. *ASTA Advances in Statistical Analysis*, 97, 2, 181-193. ISSN 1863-8171.
- Cannon, A. J. (2010). A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology. *Hydrological Processes*, 24, 6, 673-685. ISSN 1099-1085.
- Coles, S. G. (2001). *An Introduction to Statistical Modelling of Extreme Values*. Springer, London, primera edición.
- Coles, S. G. y Dixon, M. (1999). Likelihood-Based Inference for Extreme Value Models. *Extremes*, 2, 1, 5-23. ISSN 1386-1999.
- Cooley, D. y Sain, S. (2010). Spatial Hierarchical Modeling of Precipitation Extremes From a Regional Climate Model. *Journal of Agricultural, Biological, and Environmental Statistics*, 15(3), 381-402.
- Einmahl, J. H., Fils-Villetard, A., Guillou, A. *et al.* (2008). Statistics of extremes under random censoring. *Bernoulli*, 14, 1, 207-227.
- El Adlouni, S., Ouarda, T. B. M. J., Zhang, X., Roy, R. y Bobée, B. (2007). Generalized maximum likelihood estimators for the nonstationary generalized extreme value model. *Water Resources Research*, 43, 3, W03410. ISSN 1944-7973.
- Fisher, R. A. y Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24, 180-190.
- Gaetan, C. y Grigoletto, M. (2007). A hierarchical model for the analysis of spatial rainfall extremes. *Journal of Agricultural, Biological, and Environmental Statistics*, 12(4), 434-449.
- Gomes, M. y Neves, M. (2011). Estimation of the extreme value index for randomly censored data. *Biometrical Letters*, 48, 1, 1-22.

## Referencias

---

- Hastie, T. y Tibshirani, R. (1986). Generalized additive models. *Statistical science*, 297–310.
- Hosking, J. R. M. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *J. R. Stat. Soc.*, 52, 105–124.
- Hosking, J. R. M., Wallis, J. R. y Wood, E. F. (1985). Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments. *Technometrics*, 27, 3, 251–261.
- Jain, S. y Lall, U. (2001). Floods in a changing climate: Does the past represent the future? *Water Resources Research*, 37, 12, 3193–3205. ISSN 1944-7973.
- Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, 81, 158–171.
- Kharin, V. V. y Zwiers, F. W. (2005). Estimating extremes in transient climate change simulations. *J. Clim.*, 18, 1156–1173.
- Laurini, F. y Pauli, F. (2009). Smoothing sample extremes: The mixed model approach. *Computational Statistics & Data Analysis*, 53, 11, 3842–3854.
- Leadbetter, M. R., Lindgren, G. y Rootzen, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. 336pp, Springer, New York.
- Madsen, H., Rasmussen, P. F. y Rosbjerg, D. (1997). Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events: 1. At-site modeling. *Water Resources Research*, 33, 4, 747–757. ISSN 1944-7973.
- Martins, E. y Stedinger, J. (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*, 36, 737–744.
- Nosedal-Sanchez, A., Storlie, C. B., Lee, T. C. y Christensen, R. (2012). Reproducing kernel hilbert spaces for penalized regression: A tutorial. *The American Statistician*, 66, 1, 50–60.
- Padoan, S. y Wand, M. (2008). Mixed model-based additive models for sample extremes. *Statistics & Probability Letters*, 78, 17, 2850–2858.
- Pauli, F. y Coles, S. (2001). Penalized likelihood inference in extreme value analyses. *J. Appl. Stat*, 28, 547–560.
- Pickands III, J. (1975). Statistical inference using extreme order statistics. *the Annals of Statistics*, 119–131.
- R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Ribatet, M., Singleton, R. y Team, R. C. (2008). SpatialExtremes: Modelling Spatial Extremes. R package version 2.0-1.
- Rodriguez, S., Reyes, H., Perez, P. y Vaquera, H. (2012). Selection of a Subset of Meteorological Variables for Ozone Analysis: Case Study of Pedregal Station in Mexico City. *Journal of Environmental Science and Engineering A*, 1, 11–20.



## Referencias

---

- Rosen, O. y Cohen, A. (1996). Extreme percentile regression. *In: Härdle, W. and M.G. Schimek, (eds.) Statistical Theory and Computational Aspects of Smoothing: Proceedings of the COMP-STAT 94 Satellite Meeting held in Semmering, Austria, August 1994, Physica-Verlag, Heidelberg, 2728.*
- Sang, H. y Gelfand, A. E. (2010). Continuous Spatial Process Models for Spatial Extreme Values. *Journal of Agricultural, Biological, and Environmental Statistics*, 15(1), 49–65.
- Scarf, P. A. (1992). Estimation for a four parameter generalized extreme value distribution. *Commun. Stat. Theory Methods.*, 21, 2185–2201.
- Smith, R. L. (1985). Maximum likelihood estimation in a class of non-regular cases. *Biometrika*, 72, 67–92.
- Tawn, J. (1988). Bivariate extreme value theory: models and estimation. *Biometrika*, 75, 397–415.
- Walder, C. y Chapelle, O. (2007). Learning with transformation invariant kernels. En *Advances in Neural Information Processing Systems*, 1561–1568.
- Wang, X. L., Zwiers, F. W. y Swail, V. (2004). North Atlantic Ocean wave climate scenarios for the 21st century. *J. Clim.*, 17, 2368–2383.
- Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *J. Am. Stat. Assoc.*, 73, 812–815.
- Yee, T. W. y Stephenson, A. G. (2007). Vector generalized linear and additive extreme value models. *Extremes*, 10, 119.

# Anexos

## Anexo A: Rutinas en R-3.0.1 para implementar los modelos bayesianos de valores extremos con datos censurados

```
pwrM=function(x,p)
{
x=as.matrix(x)
e=eigen(x)
#e$values[e$values<=0]=1
return(e$vectors%*%diag(e$values^p)%*%t(e$vectors))
}
```

```
distancia=function(v1,v2)
{
#print(v2)
#return(sqrt(sum((v1-v2)^2)))
return(sqrt(sum((v1-v2)^2)))
}
```

```
pwrM=function(x,p)
{
x=as.matrix(x)
e=eigen(x)
return(e$vectors%*%diag(e$values^p)%*%t(e$vectors))
}
```

```
thin_plates_fun=function(v){return(v*v*log(v))}
C_thin_plates=function(v)
{
ret=v
}
```

```
ret[ret!=0]=thin_plates_fun(ret[ret!=0])
return(ret)
}

bases_THIN_PLATES=function(x,centro)
{
x=as.matrix(x)
if(ncol(x)!=ncol(centro))
{
x=t(x)
}
n=nrow(centro)
basess=as.matrix(sapply(1:n,function(v2){return(apply(x,1,distancia,centro[v2,]))}))
if(ncol(basess)!=nrow(centro))
{
basess=t(basess)
}
thinpl=C_thin_plates(basess)
return(thinpl%%pwrM(var(thinpl),-0.5))

#return(thinpl%%pwrM(C_thin_plates(as.matrix(dist(centro))),-0.5))
#return(basess*basess*log(basess))#thin plates
}

bases_GAUSSIAN=function(x,centro,h=1)
{
x=as.matrix(x)
if(ncol(x)!=ncol(centro))
{
x=t(x)
}
n=nrow(centro)
basess=as.matrix(sapply(1:n,function(v2){return(apply(x,1,distancia,centro[v2,]))}))
if(ncol(basess)!=nrow(centro))
{
basess=t(basess)
}
return(exp(-h*basess))
#return(basess*basess*log(basess))#thin plates
}

bases=bases_THIN_PLATES
```

```
x_features=function(xx,k,lambda=0.1)
{
xx=as.matrix(xx)
#xx=xx%*%pwrM(var(xx),0.5)
n=nrow(xx)
clust=hclust(dist(xx))
g=cutree(clust,k)
centro=c()
#xx=cbind(xx,xx)
xx=as.matrix(xx)
for (i in 1:k)
{
if(sum(g==i)!=1)
centro=rbind(centro,t(as.matrix(apply(as.matrix(xx[g==i,]),2,mean))))
if(sum(g==i)==1)
centro=rbind(centro,t(as.matrix(apply(t(as.matrix(xx[g==i,])),2,mean))))
}
Ik=diag(rep(1,k))
D=bases(xx,centro)
ret=list()
ret[[1]]=D
ret[[2]]=centro
return(ret)
}
```

#####

```
gev_FEATURES=function(x1,x2,y,knotss1=10,knotss2=10)
{
if(!is.null(x1))
{
x1=as.matrix(x1)
}
if(!is.null(x2))
{
x2=as.matrix(x2)
}
#meannn=apply(as.matrix(x),2,mean)
#meann=meannn
#meann=rep(meann,nrow(x))
#dim(meann)=dim(t(x))
#meann=t(meann)
#x=x-meann
#sqrtM=pwrM(var(x),-0.5)
#x=x%*%sqrtM
```

```
y=as.matrix(y)
gev_F=list()
#h=10#numero de nodos spline
#p=ncol(x)
X=c()
Z1=c()
Z2=c()
gev_F[[1]]=c()
gev_F[[2]]=c()
gev_F[[3]]=c()
gev_F[[4]]=c()
longZ1=0
longZ2=0
if(!is.null(x1))
{
X=x1
Z_feat1=cbind(x_features(x1,k=knotss1))
Z1=Z_feat1[[1]]
gev_F[[2]]=Z_feat1[[2]]
longZ1=ncol(Z1)
}
if(!is.null(x2))
{
X=cbind(X,x2)
Z_feat2=cbind(x_features(x2,k=knotss2))
Z2=Z_feat2[[1]]
gev_F[[3]]=Z_feat2[[2]]
longZ2=ncol(Z2)
}
X=cbind(rep(1,length(y)),X)
T=cbind(X,Z1,Z2)

#betas=[b,ux,us,sigma,xi,var_ux,var_uy]
gev_F[[1]]=T
gev_F[[4]]=c(ncol(X),ncol(X)+longZ1,ncol(X)+longZ1+longZ2)
return(gev_F)
}

#####
derivb_val=function(bs,bm,xi,y,x,fun)
{
x=as.matrix(x)#x es un vector columna y "y" es un valor
dim(x)=c(1,prod(dim(x)))
bs=as.matrix(bs)
dim(bs)=c(prod(dim(bs)),1)
bm=as.matrix(bm)
```

```
dim(bm)=c(prod(dim(bm)),1)
mu=x%*%bm
#print(mu)
#solo intercepto para sigma
lsigma=x%*%bs
if((1+xi*(y-mu)/exp(lsigma))>0){ret=fun(lsigma, mu, xi, y)}else{ret=-999999}
if(is.infinite(ret))
{
ret=-999999
}
#print("lsigma")
#print(lsigma)

#print("mu")
#print(mu)

#print("xi")
#print(xi)

#print("ret")
#print(ret)

return(ret)
}
```

```
deriv_gev_val_cens=function(bs,bm,xi,vec_cens,y,x)#y es un vector y x es una matrix
{
gev=function(lsigma, mu, xi, y){(-lsigma-((1/xi)+1)*log(1+xi*(y-mu)/exp(lsigma))-
(1+xi*(y-mu)/exp(lsigma))^-1/xi)}
gev1=function(lsigma, mu, xi, y){log(1-exp(-((1+xi*(y-mu)/exp(lsigma))^-1/xi)))}
x=as.matrix(x)
y=as.matrix(y)
n=nrow(x)
p=length(bs)+length(bm)+1
value=0
for (i in 1:n)
{
if(ncol(x)>1)
{
if(vec_cens[i]==0)
{
tmp=derivb_val(bs,bm,xi,y[i],x[i,],fun=gev)
}
else
{
tmp=derivb_val(bs,bm,xi,y[i],x[i,],fun=gev1)
}
```

```

}
}
if(ncol(x)==1)
{
if(vec_cens[i]==0)
{
tmp=derivb_val(bs,bm,xi,y[i],x[i],fun=gev)
}
else
{
tmp=derivb_val(bs,bm,xi,y[i],x[i],fun=gev1)
}
}
if(is.infinite(tmp))
{
#print("bs")
#print(bs)
#print("x[i,]")
#print(x[i,])
#print("bs%*%x[i,]")
#print(sum(c(bs)*c(x[i])))
stop("aqui")
}
value=value+tmp[[1]]
}
return(value)
}

lv_gev_censura=function(betas,vec_cens,y,x)
{
np=length(betas)
p=(np-1)/2
bm=as.matrix(betas[1:p])
bs=as.matrix(betas[(p+1):(np-1)])
xi=as.matrix(betas[np])
der=deriv_gev_val_cens(bs,bm,xi,vec_cens=vec_cens,y,x)
res= der
return(res)
}

dmtnorm=function(x,mu,sigma)
{
return(((1/(((2*pi)^(length(mu)/2))*sqrt(det(sigma)))))*
exp(-0.5*t(x-mu)%*%solve(sigma)%*%(x-mu)))
}

```

```
dLmtnorm=function(x,mu,sigma)
{
return(log(1/(((2*pi)^(length(mu)/2))*sqrt(det(sigma))))+
(-0.5*t(x-mu)%*%solve(sigma)%*(x-mu)))
}

library("LaplacesDemon")

#betas=[b,ux,us,sigma,xi,var_ux,var_us]
lv_bayes=function(betas,vec_cens,y,xx,k,separador_variables,
solo_intercepto_SIGMA=TRUE)#original
{
b=c()
ux=c()
us=c()
sigma=c()
xi=c()
var_ux=c()
var_us=c()
b=betas[1:separador_variables[1]]
if((separador_variables[1]+1)<=separador_variables[2])
ux=betas[(separador_variables[1]+1):separador_variables[2]]
if((separador_variables[2]+1)<=separador_variables[3])
us=betas[(separador_variables[2]+1):separador_variables[3]]
if((separador_variables[3]+1)<=separador_variables[4])
sigma=betas[(separador_variables[3]+1):separador_variables[4]]
if((separador_variables[4]+1)<=separador_variables[5])
xi=betas[(separador_variables[4]+1):separador_variables[5]]
if((separador_variables[5]+1)<=separador_variables[6])
var_ux=exp(betas[(separador_variables[5]+1):separador_variables[6]])
if((separador_variables[6]+1)<=separador_variables[7])
var_us=exp(betas[(separador_variables[6]+1):separador_variables[7]])

aprioris=0

#formamos los betas requeridos en gev
if(solo_intercepto_SIGMA)
{
betas_PAR=c(b,ux,us,sigma,rep(0,separador_variables[3]-1),xi)
}else{
betas_PAR=c(b,ux,us,sigma,xi)
b_sigma=c()
ux_sigma=c()
us_sigma=c()
}
```



```

var_ux_sigma=c()
var_us_sigma=c()

b_sigma=betas[(separador_variables[3]+1):(separador_variables[3]
+separador_variables[1])]
if((separador_variables[1]+1)<=separador_variables[2])
ux_sigma=betas[(separador_variables[3]+
separador_variables[1]+1):(separador_variables[3]+separador_variables[2])]
if((separador_variables[2]+1)<=separador_variables[3])
us_sigma=betas[(separador_variables[3]+
separador_variables[2]+1):(separador_variables[3]+separador_variables[3])]
if((separador_variables[7]+1)<=separador_variables[8])
var_ux_sigma=exp(betas[(separador_variables[7]+1):separador_variables[8]])
if((separador_variables[8]+1)<=separador_variables[9])
var_us_sigma=exp(betas[(separador_variables[8]+1):separador_variables[9]])

apriori_ux_sigma=0
apriori_us_sigma=0
apriori_b_sigma=0
apriori_half_cauchi_sigma=0
if(!is.null(ux_sigma))
apriori_ux_sigma=dLmtnorm(ux_sigma,rep(0,length(ux_sigma)),
diag(rep(var_ux_sigma,length(ux_sigma))))
if(!is.null(us_sigma))
apriori_us_sigma=dLmtnorm(us_sigma,rep(0,length(us_sigma)),
diag(rep(var_us_sigma,length(us_sigma))))
if(!is.null(b_sigma))
apriori_b_sigma=dLmtnorm(b_sigma,rep(0,length(b_sigma)),
diag(rep(10000,length(b_sigma))))
if(!is.null(var_ux_sigma))
apriori_half_cauchi_sigma=dhalfcauchy(var_ux_sigma,scale=25,log=TRUE)
if(!is.null(var_us_sigma))
apriori_half_cauchi_sigma=apriori_half_cauchi_sigma+
dhalfcauchy(var_us_sigma,scale=25,log=TRUE)
aprioris=apriori_b_sigma+apriori_ux_sigma+apriori_us_sigma+
apriori_half_cauchi_sigma
}

apriori_ux=0
apriori_us=0
apriori_b=0
apriori_half_cauchi=0

if(!is.null(ux))
apriori_ux=dLmtnorm(ux,rep(0,length(ux)),diag(rep(var_ux,length(ux))))
#dmvnorm(ux,rep(0,length(ux)),diag(rep(.1,length(ux))),log=TRUE)
if(!is.null(us))

```

```
apriori_us=dLmtnorm(us,rep(0,length(us)),diag(rep(var_us,length(us))))
if(!is.null(b))
apriori_b=dLmtnorm(b,rep(0,length(b)),diag(rep(10000,length(b))))
if(!is.null(var_ux))
apriori_half_cauchi=dhalfcauchy(var_ux,scale=25,log=TRUE)
if(!is.null(var_us))
apriori_half_cauchi=apriori_half_cauchi+dhalfcauchy(var_us,scale=25,log=TRUE)

if(solo_intercepto_SIGMA)
{
if(!is.null(sigma))
apriori_half_cauchi=apriori_half_cauchi+dhalfcauchy(sigma,scale=25,log=TRUE)
}

#apriori de shape es log(1/10)
apriori_shape=ifelse(-10<xi&&xi<10,log(1/10),-100000000)#=log(dbeta(shape,6,9))
aprioris=aprioris+apriori_b+apriori_ux+apriori_us+
apriori_half_cauchi+apriori_shape
return(0-(lv_gev_censura(betas=betas_PAR,vec_cens=censura,y=y,x=xx)+aprioris))
}

recupera_parametros=function(betas,solo_intercepto_SIGMA=TRUE)
{
ret=list()
betas_PAR=c()

b_sigma=c()
ux_sigma=c()
us_sigma=c()
var_ux_sigma=c()
var_us_sigma=c()

b=c()
ux=c()
us=c()
sigma=c()
xi=c()
var_ux=c()
var_us=c()

b=betas[1:separador_variables[1]]

if((separador_variables[1]+1)<=separador_variables[2])
ux=betas[(separador_variables[1]+1):separador_variables[2]]
if((separador_variables[2]+1)<=separador_variables[3])
us=betas[(separador_variables[2]+1):separador_variables[3]]
```

```
if((separador_variables[3]+1)<=separador_variables[4])
sigma=betas[(separador_variables[3]+1):separador_variables[4]]
if((separador_variables[4]+1)<=separador_variables[5])
xi=betas[(separador_variables[4]+1):separador_variables[5]]
if((separador_variables[5]+1)<=separador_variables[6])
var_ux=exp(betas[(separador_variables[5]+1):separador_variables[6]])
if((separador_variables[6]+1)<=separador_variables[7])
var_us=exp(betas[(separador_variables[6]+1):separador_variables[7]])

#formamos los betas requeridos en gev
if(solo_intercepto_SIGMA)
{
betas_PAR=c(b,ux,us,sigma,rep(0,separador_variables[3]-1),xi)
b_sigma=sigma
}else{
betas_PAR=c(b,ux,us,sigma,xi)
b_sigma=betas[(separador_variables[3]+1):(separador_variables[3]+
separador_variables[1])]
if((separador_variables[1]+1)<=separador_variables[2])
ux_sigma=betas[(separador_variables[3]+
separador_variables[1]+1):(separador_variables[3]+separador_variables[2])]
if((separador_variables[2]+1)<=separador_variables[3])
us_sigma=betas[(separador_variables[3]+
separador_variables[2]+1):(separador_variables[3]+separador_variables[3])]
if((separador_variables[7]+1)<=separador_variables[8])
var_ux_sigma=exp(betas[(separador_variables[7]+1):separador_variables[8]])
if((separador_variables[8]+1)<=separador_variables[9])
var_us_sigma=exp(betas[(separador_variables[8]+1):separador_variables[9]])
}

ret[[1]]=c(b,ux,us)#parametros betas de mu
ret[[2]]=c(b_sigma,ux_sigma,us_sigma)#parametros betas de sigma
ret[[3]]=xi#shape
ret[[4]]=c(var_ux,var_us)#varianzas de mu
ret[[5]]=c(var_ux_sigma,var_us_sigma)#varianzas de sigma

return(ret)
}

#setwd("/home/gregorio/alex/0simula")
setwd("C:/Users/alejandro/Documents/0Doctorado/00Articulo/000simulacion")
library(SpatialExtremes)
data(rainfall)
```

```
op <- par(mfrow = c(1,1),pty = "s", mar=c(4,4,2,2))
swiss(city = TRUE,axes = TRUE,xlab="x coord",ylab="y coord")
idx.site <- c(1, 10, 20)
plot(1962:2008, rain[,1], type = "b", xlab = "Year", ylab =
"Precipitation (cm)", ylim = c(0, 120), col = 2)
points(coord[-idx.site,])
points(coord[idx.site,], pch = 15, col = 2:4)
plot(1962:2008, rain[,1], type = "b", xlab = "Year", ylab =
"Precipitation (cm)", ylim = c(0, 120), col = 2)
lines(1962:2008, rain[,10], col = 3, type = "b")
lines(1962:2008, rain[,20], col = 4, type = "b")

rain_data=matrix(nrow=0,ncol=5)
for (i in 1:nrow(coord))
{
rain_data=rbind(rain_data,cbind(rain[,i],coord[i,1],coord[i,2],
as.matrix(1:nrow(rain)),coord[i,3]))
}
colnames(rain_data)=c("maxi","x","y","t","e")
y=rain_data[,1]
x1=rain_data[,4:5]
x2=rain_data[,2:3]

censura=as.matrix(rep(0,length(y)))

info_spline=gev_FEATURES(x1,x2,y,knotss1=20,knotss2=30)
T=info_spline[[1]]
separador_variables=info_spline[[4]]

solo_intercepto_SIGMA=TRUE
if(solo_intercepto_SIGMA)
{
separador_variables=c(separador_variables,
separador_variables[length(separador_variables)]+1)
}else{
separador_variables=c(separador_variables,2*
separador_variables[length(separador_variables)])
}

# el for es para cada uno de: xi,var_ux,var_us
for (i in 1:3)
{
separador_variables=c(separador_variables,
separador_variables[length(separador_variables)]+1)
}

if(!solo_intercepto_SIGMA)
```

```
{
# el for es para cada uno de: var_ux_sigma,var_us_sigma
for (i in 1:2)
{
separador_variables=c(separador_variables,
separador_variables[length(separador_variables)]+1)
}
}

numero_parametros=separador_variables[length(separador_variables)]
set.seed(9181)
start_par=runif(numero_parametros)
betas=start_par
recupera_parametros(start_par,solo_intercepto_SIGMA)
x=T
vec_cens=censura
lv_bayes(start_par,vec_cens=censura,y=y,xx=T,k=k,
separador_variables=separador_variables,
solo_intercepto_SIGMA=solo_intercepto_SIGMA)

con=list()
con$fnscale=1
con$maxit=500
con$trace=1
con$REPORT=1
#esto tarda una hora
control=list()
control$trace=TRUE
#control$iter.max=5
#control$eval.max=3

#miopt=nlminb(lv_bayes,start=start_par,control=control,
vec_cens=censura,y=y,x=T,k=k,separador_variables=separador_variables,
solo_intercepto_SIGMA=solo_intercepto_SIGMA)
#start_par=miopt$par
miopt=optim(par=start_par,fn=lv_bayes,
method="BFGS",hessian=T,control=con,vec_cens=censura,y=y,xx=T,k=k,
separador_variables=separador_variables,
solo_intercepto_SIGMA=solo_intercepto_SIGMA)
#betas=miopt$par
#lv_bayes(betas,vec_cens=censura,y=y,xx=T,k=k,
separador_variables=separador_variables,
solo_intercepto_SIGMA=solo_intercepto_SIGMA)
#recupera_parametros(betas,solo_intercepto_SIGMA)
```

```
rwmetro <- function(target,N,paramss,VCOV,burnin=0,...)
{
require(MASS) #requires package MASS for normal sampling
samples <- paramss
for (i in 2:(burnin+N))
{
prop <- mvrnorm(n = 1, paramss, VCOV)
if (runif(1) < min(1, target(prop,...)/target(paramss,...)))
paramss <- prop
samples <- rbind(samples,paramss)
cat(i,"\n")
}
samples[(burnin+1):(N+burnin),]
}

VarCov=-(1/length(y))*solve(miopt$hessian)

post.samp=rwmetro(target=lv_bayes,N=100,paramss=miopt$par,
VCOV=VarCov,vec_cens=censura,y=y,x=T,k=k)

param_est=apply(as.matrix(post.samp),2,mean)

#plot(post.samp[,ncol(post.samp)])
cadenas=as.matrix(rw)

#setwd("C:/Users/alejandro/Documents/0Doctorado/00Articulo/000simulacion")

save(post.samp,fitt,cadenas,file=paste("modelo_bocci_geo.RData",sep=""))
```