UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA

DIVISIÓN DE ESTUDIOS DE POSGRADO

ESTIMACIÓN DE PARÁMETROS EN MODELOS DE VALORES EXTREMOS USANDO ÁRBOLES DE DECISIÓN

TESIS

PARA OBTENER EL TÍTULO DE:

DOCTORA EN MODELACIÓN MATEMÁTICA

PRESENTA:

M.C. SONIA VENANCIO GUZMÁN

DIRECTOR DE TESIS:
DR. ALEJANDRO IVÁN AGUIRRE SALADO
CO-DIRECTOR:
DR. GUILLERMO ARTURO LANCHO ROMERO

H. Cd. de Huajuapan de León, Oaxaca, México, Octubre de 2023

Estimación de parámetros en modelos de valores extremos usando árboles de decisión

Venancio Guzmán Sonia

Dedicatoria

A mis padres, hermanos y amigos.

Agradecimientos

A cada una de las personas que están en mi vidas siempre apoyándome.

A mi asesor Dr. Alejandro Iván Aguirre Salado por guiarme en todo el proceso para adquirir conocimientos sobre la especialidad en estadística, por su paciencia y tiempo dedicado a este trabajo de tesis.

Al Dr. Guillermo Arturo Lancho Romero, al Dr. Sergio Palafox Delgado, al Dr. José del Carmen Jiménez Hernández, al Dr. Emmanuel Abdías Romano Castillo, al Dr. David Israel Celis Euan y a la Dra. María Guzmán Martínez por el apoyo brindado en la revisión de esta tesis.

A la División de Estudios de Posgrado de la universidad por haberme brindado la oportunidad de realizar el doctorado en Modelación Matemática en esta institución.

Al Consejo Nacional de Humanidades, Ciencia y Tecnología (CONAHCYT) por el apoyo económico que me brindo durante mi estancia en el doctorado.

Introducción

En la actualidad, el análisis de los datos es de gran importancia, ya que permite obtener información a través de datos observados. En la práctica es usual que los datos tengan un comportamiento muy variado, por lo cual, se requieren de técnicas estadísticas que nos permitan obtener la información deseada. De forma general, si el problema es realizar regresión o clasificación, los métodos paramétricos son una de las mejores alternativas para determinar una solución, esto a su vez implica, el estudio sobre la estimación de los parámetros del modelo propuesto. Más aún, la calidad o bondad de ajuste de dicho modelo depende primordialmente de la estimación de los parámetros correspondientes.

Por otra parte, al analizar un conjunto de datos, pueden encontrarse observaciones con valores "muy grandes" o "muy pequeños", formalmente estos son denominados Valores Extremos (VE), los cuales también proporcionan información relevante sobre los datos. Por tal motivo, es necesario preguntarse sobre cómo podemos predecir estos eventos que son poco usuales. Teóricamente, la Teoría de Valores Extremos (TVE), es una rama de la estadística que centra su interés en la modelación del comportamiento de estos valores máximos o mínimos de una serie de datos, por lo cual, formalmente se busca determinar la forma de la distribución límite a la cual estos valores extremos se pueden aproximar.

Aunque la historia de la teoría de los Valores Extremos (VE) es difícil de situar con exactitud, uno de los resultados más importantes en el análisis de dicha teoría debido a la influencia que ha tenido en diversas aplicaciones, es el primer teorema fundamental de valores extremos descrito en 1928 por R. A. Fisher y L. H. C. Tippett [10] y, posteriormente retomado por B. V. Gnedenko [14]. El resultado de V. Gnedenko hace referencia a la distribución asintótica o límite del máximo bajo una transformación, lo cual a su vez se extiende a la distribución límite del mínimo. Este teorema establece que la sucesión del máximo de una sucesión de variables aleatorias independientes e idénticamente distribuidas (denotado comúnmente por v.a.i.i.d.) adecuadamente normalizada con parámetros de localización y escala, converge en distribución a sólo una de las tres familias paramétricas: Gumbel, Weibull o Fréchet, y estas tres distribuciones límite, a la cual tal sucesión pueden converger se puede agrupar en una sola función de distribución propuesta por Von Mises en [27], la cual a su vez, es denominada como distribución de Valores Extremos Generalizados (DVEG). Posteriormente, una vez establecido la distribución límite a la cual la sucesión de variables converge, se puede llevar a cabo un proceso completo

de inferencia estadística para encontrar los estimadores de los parámetros de dicha distribución. Así, en este trabajo de investigación se realizó un estudio sobre la teoría de valores extremos y modelos basados en árboles de decisión con la finalidad primordial de estimar los parámetros del modelo propuesto. Por lo cual, los capítulos se dividen y describen de la siguiente forma:

- **Capítulo 1:** En el primer capítulo se realiza una breve descripción de algunos conceptos básicos necesarios para posteriormente sumergirse en la teoría de valores extremos, algunos de estos son: convergencia de variables aleatorias independientes e idénticamente distribuidos (v.a.i.i.d.) y el teorema del límite central (TLC). Posteriormente, se presenta la definición de las variables aleatorias: máximos y mínimos, y se enuncia el primer teorema fundamental de la TVE relacionado con la convergencia en distribución de dichas variables aleatorias. Además, se dan algunos ejemplos claves para mejor comprensión de la misma. Finalmente, se establece la relación existente entre la distribución de valores extremos con la distribución de valores extremos generalizados.

- **Capítulo 2:** En este capítulo se resume brevemente la teoría de árboles de regresión y se describe una metodología sobre cómo entrenar un modelo basado en árboles. Luego, se presenta uno de los temas principales: estimación de funciones, en el cual la función es expresada de forma aditiva: $F(\boldsymbol{x}; \{\beta_m, \boldsymbol{a_m}\}) = \sum_{m=1}^{M} \beta_m h(\boldsymbol{x}; \boldsymbol{a_m})$, donde las funciones $h(x; \boldsymbol{a_m})$ representan a los árboles de regresión, y son conocidas como funciones bases. Por su parte, β_m y $\boldsymbol{a_m}$ son los parámetros a estimar. Así pues, para la estimación de los parámetros óptimos, se recurre al estudio del método de descenso por gradiente.

  Por último, se describe paso a paso la metodología del algoritmo XGBoost, el cual tiene como objetivo estimar una función objetivo mediante árboles de decisión.

- **Capítulo 3:** Se expone un caso de estudio en donde se puede ver la aplicación de la teoría de valores extremos. Se presenta un modelo espacial basado en la distribución de VEG y modelos de conjuntos de árboles para analizar los niveles máximos de concentración de material particulado en el área metropolitana de la Ciudad de México. Las tendencias espaciales se modelaron a través de un árbol de decisión en el contexto de un modelo VEG no estacionario. Se utilizó un modelo de conjunto de árboles como predictor de los parámetros de VEG. El árbol de decisión se construyó utilizando un enfoque voraz por etapas, donde la función objetivo es el logaritmo de la función verosimilitud. Además, se verificó la validez del modelo mediante la verosimilitud y el criterio de información de Akaike.

Finalmente, con el análisis y estudio realizado se determinó que, una de las principales novedades de este trabajo está en la forma de asociar las covariables o variables independientes con los parámetros de la distribución VEG, este hallazgo condujo a la publicación de un artículo de investigación, [2]. Asimismo, en este estudio se propuso un árbol de decisión basado en los resultados

satisfactorios obtenidos en varias aplicaciones de aprendizaje automático. Además, es una de las primeras implementaciones donde los árboles se ajustaron simultáneamente a más de un parámetro. Se asumió que las observaciones en la misma localidad espacial tienen los mismos parámetros de forma y escala en la distribución de VEG. Sin embargo, también se consideró que el parámetro de ubicación varía espacialmente según una tendencia que es modelada por su respectivo árbol de decisión.

Con el modelo propuesto se garantiza que,

1. Localmente, la muestra proviene de la misma población, por lo cual se puede estimar la tendencia de forma conjunta en toda la región.

2. El enfoque propuesto permite obtener un modelo regularizado sin necesidad de incluir un término adicional para regularizar el modelo.

Índice general

Capítulo 1

Valores extremos

Existen diversas aplicaciones en donde el estudio de observaciones que son muy grandes o muy pequeñas son de gran importancia (comúnmente denominados valores extremos). Formalmente la Teoría de Valores Extremos (TVE por sus siglas en español o EVT por sus siglas en inglés), es una rama de la estadística cuyo interés es el estudio de este tipo de observaciones: muy grandes o muy pequeñas. Por esta razón, la teoría de valores extremos tiene diversas aplicaciones, algunas de estas son: en hidrología, con en el análisis de las inundaciones y precipitación máxima esperada en los próximos años, en finanzas, con el estudio de las series de precios de un activo financiero, en meteorología, en el estudio de cambios extremos de la temperatura, en la ingeniería en el estudio de resistencia de los materiales, entre otros.

Para comprender los principales resultados de la teoría de los valores extremos, a continuación se presentan algunos conceptos básicos y fundamentales de la teoría probabilística que permitirán profundizar en dicho tema.

1.1. Convergencia de variables aleatorias

En esta sección se estudian los distintos tipos de convergencia de las variables aleatorias, tales como: convergencia puntual, convergencia casi segura, convergencia en distribución, entre otros, véase [20].

Definición 1.1.1. (**Convergencia puntual**) Sean $\{X_n\}_{n\in\mathbb{N}}$ una sucesión de variables aleatorias definidas en el espacio de probabilidad $(\Omega, \mathcal{F}, P)$ y X una variable aleatoria. definida sobre el mismo espacio de probabilidad. La sucesión $\{X_n\}_{n\in\mathbb{N}}$ converge (puntualmente) a X si para cada $\omega \in \Omega$ se satisface que,

$$\lim_{n\to\infty} X_n(\omega) = X(\omega). \tag{1.1}$$

Esta convergencia comúnmente se denota por $X_n \overset{c.p.}{\to} X$.

Ejemplo 1.2. Sea $\{X_n\}_{n\in\mathbb{N}}$ una sucesión de variables aleatorias definidas en el espacio de Lebesgue $(\Omega, \mathcal{F}, P)$ donde $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B} := \sigma([0, 1])$ y $P = \lambda$. La sucesión X_n está definida como $X_n(\omega) = \omega^n$, vea Figura (1.1). La sucesión converge puntualmente a la variable aleatoria,

$$X(\omega) = \begin{cases} 0, & si \quad \omega \in [0,1), \\ \\ 1, & si \quad \omega = 1. \end{cases}$$

En efecto, dado $\varepsilon > 0$, nótese que si $\omega = 1$ entonces $X_n(\omega) = 1$ y se cumple:

$$|X_n(\omega) - X(\omega)| = |0 - 0| = 0 < \varepsilon.$$

Por otra parte, si $\omega \in [0,1)$ para $n > ln(\varepsilon)/ln(\omega)$ se tiene,

$$|X_n(\omega) - X(\omega)| = |\omega^n - 0| = \omega^n < \varepsilon.$$

Por lo tanto, $X_n(\omega) \to X(\omega)$ puntualmente.

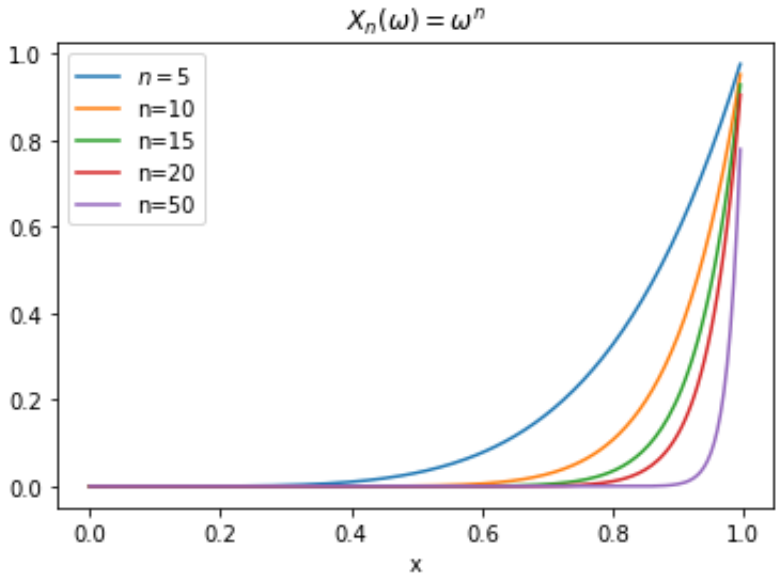


Figura 1.1: Gráfica de la variable aleatoria $X_n(\omega) = \omega^n$.

Como puede verse, la convergencia puntual es una condición fuerte en el sentido de que es necesario verificar la convergencia en cada punto del espacio muestral. Un tipo de convergencia más débil, es la convergencia casi segura, en el cual la convergencia se verifica salvo en un conjunto de medida cero.

Definición 1.2.1. (Convergencia casi segura) Sea $\{X_n\}_{n\in\mathbb{N}}$ una sucesión de variables aleatorias definidas en el espacio de probabilidad $(\Omega, \mathcal{F}, P)$ se dice que la sucesión converge de forma casi segura (o con probabilidad 1) a la variable aleatoria X si,

$$P\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1,$$

o equivalentemente,

$$P\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\right\}\right) = 0.$$

La convergencia casi segura se denota como $X_n \overset{c.s.}{\to} X$, o bien, $\lim_{n\to\infty} X_n = X$.

Ejemplo 1.3. Considere el espacio de probabilidad $([0,1], \mathcal{B}, \lambda)$, con λ la medida de Lebesgue sobre $\mathcal{B} := \sigma([0,1])$ y las variables aleatorias X_n y X definidas por $X_n(\omega) = \omega^n$ y $X(\omega) = 0$ para cada $\omega \in [0,1]$ respectivamente. Entonces,

$$X_n(\omega) \to 0 \text{ para } \omega \in [0,1).$$

Nótese que la sucesión efectivamente converge a X, salvo en el punto 1, más aún, el conjunto en donde no se da la convergencia a la variable X es de medida cero. Por lo tanto, la convergencia es casi segura.

Definición 1.3.1. (Convergencia en probabilidad) Sea $\{X_n\}_{n\in\mathbb{N}}$ una sucesión de variables aleatorias definidas en el espacio de probabilidad $(\Omega, \mathcal{F}, P)$, se dice que la sucesión converge en probabilidad a la variable aleatoria X, si para cada $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon\}\right) = 0,$$

o equivalentemente,

$$\lim_{n\to\infty} P\left(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| \leq \epsilon\}\right) = 1.$$

La convergencia en probabilidad suele denotarse como $X_n \xrightarrow{P} X$.

Ejemplo 1.4. Considere el espacio de probabilidad $((0,1], \mathcal{B}, \lambda)$, así, dado que $P = \lambda$, P asigna a cada intervalo su longitud. Para cada $n \in \mathbb{N}$, definimos $X_n = 1_{(0,1/n]}$, esto es,

$$X_n(\omega) = \begin{cases} 1, & \text{si } \omega \in (0, 1/n], \\ \\ 0, & \text{en otro caso.} \end{cases}$$

Se satisface entonces que, $X_n \xrightarrow{P} 0$.

En efecto, dado $\varepsilon > 0$ se tiene:

$$\begin{aligned} P(|X_n| > \varepsilon) &= P(X_n > \varepsilon) \\ &= \begin{cases} 1/n, & \text{si } \varepsilon < 1, \\ \\ 0, & \text{en otro caso,} \end{cases} \end{aligned}$$

de donde, $\lim_{n\to\infty} P(|X_n| > \varepsilon) = 0$ y por lo tanto $X_n \xrightarrow{P} 0$. Vea Figura (1.2).

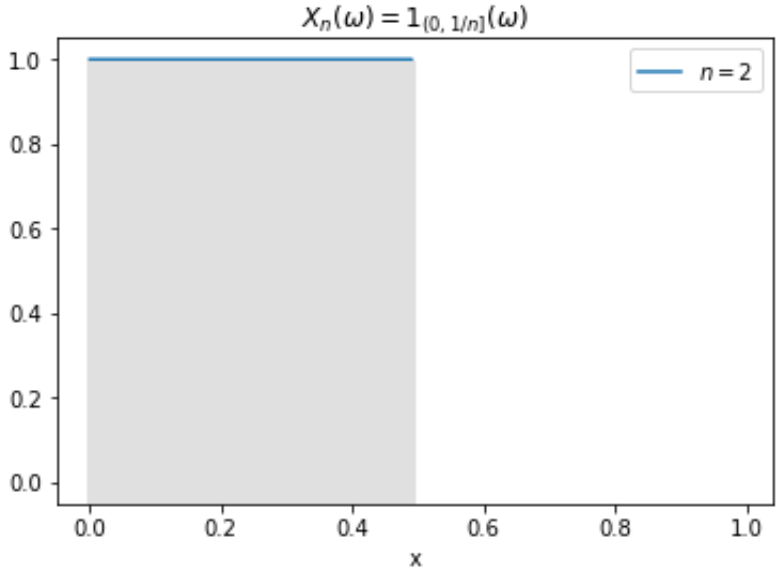


Figura 1.2: Gráfica de la variable aleatoria X_n.

Ejemplo 1.5. Considere el espacio de probabilidad $((0,1], \mathcal{B}, \lambda)$ y considere la sucesión definida por,

$$X_n(\omega) = \begin{cases} 1_{(0,1/n]}(\omega), & \text{si } n \text{ es impar,} \\ \\ 1_{(1/n,1]}(\omega), & \text{si } n \text{ es par.} \end{cases}$$

Se puede verificar que la sucesión no converge en probabilidad.

En efecto, dado $\varepsilon > 0$ y n impar se tiene,

$$P(|X_n| > \varepsilon) = \begin{cases} 1/n, & \text{si } \varepsilon < 1, \\ \\ 0, & \text{en otro caso,} \end{cases}$$

de este modo, $X_{2n+1} \xrightarrow{P} 0$.

Ahora, nótese que para n par, se obtiene que,

$$P(|X_n - 1| > \varepsilon) = \begin{cases} 1/n, & \text{si } \varepsilon < 1, \\ \\ 0, & \text{en otro caso,} \end{cases}$$

así, $X_{2n} \xrightarrow{P} 1$ y por lo tanto X_n no converge en probabilidad.

La convergencia menos restrictiva de todas las descritas anteriormente, es la convergencia en distribución, también conocido como convergencia débil o convergencia en Ley. Este tipo de convergencia involucra únicamente a las funciones de distribución, con lo cual, se podría tener el caso en el que las variables aleatorias estén definidas en distintos espacios de probabilidad.

Definición 1.5.1. (**Convergencia en ley o en distribución**) Una sucesión de variables aleatorias $X_1, X_2, \ldots$, con función de distribución $F_{X_1}, F_{X_2}, \ldots$,

respectivamente, se dice que converge en distribución a la variable aleatoria X teniendo función de distribución F_X si,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \tag{1.2}$$

para cada punto en donde F es continua. La convergencia en distribución se denota por $F_n(x) \xrightarrow{d} F(X)$.

Ejemplo 1.6. Si X es una variable aleatoria con distribución normal de media μ y varianza σ^2, esto es $X \sim N(\mu, \sigma)$, entonces la variable aleatoria $Z = \frac{x-\mu}{\sigma}$ se distribuye de forma normal con media 0 y varianza 1.
Más aún, se sabe del teorema del límite central que, si $X_1, X_2, \ldots, X_n$ es una sucesión de variables aleatorias con distribución F, de media μ y varianza σ^2 entonces:

$$\frac{\overline{X} - \mu}{\sigma\sqrt{n}} \xrightarrow{d} Z, \tag{1.3}$$

donde $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ es la media muestral y $Z \sim N(0, 1)$.

Ejemplo 1.7. Considere la sucesión $X_1, X_2, \ldots$, con función de distribución $N(0, \sigma^2/n)$ para cada X_n respectivamente. La sucesión, converge a la variable aleatoria constante $X = 0$, con función de distribución:

$$F_X(x) = \begin{cases} 0, & \text{si } x < 0, \\ \\ 1, & \text{si } x \geq 0. \end{cases}$$

En efecto, dado que $X_n \sim N(0, \sigma^2/n)$ se tiene,

$$F_{X_n}(x) = \frac{1}{\sqrt{2\pi\sigma^2/n}} \int_{-\infty}^{x} e^{-u^2/2(\sigma^2/n)} du. \tag{1.4}$$

La gráfica de la función de distribución se puede ver en la Figura (1.3).
Nótese que en el límite la función de distribución es,

$$\lim_{n\to\infty} F_{X_n}(x) = \begin{cases} 0, & \text{si } x < 0, \\ \\ 1/2, & \text{si } x = 0, \\ \\ 1, & \text{si } x > 0. \end{cases}$$

Con lo cual puede verse fácilmente que,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \tag{1.5}$$

para cada punto x en donde $F_X(x)$ es continua, esto es, excepto en el punto $x = 0$. Vea Figura (1.3).

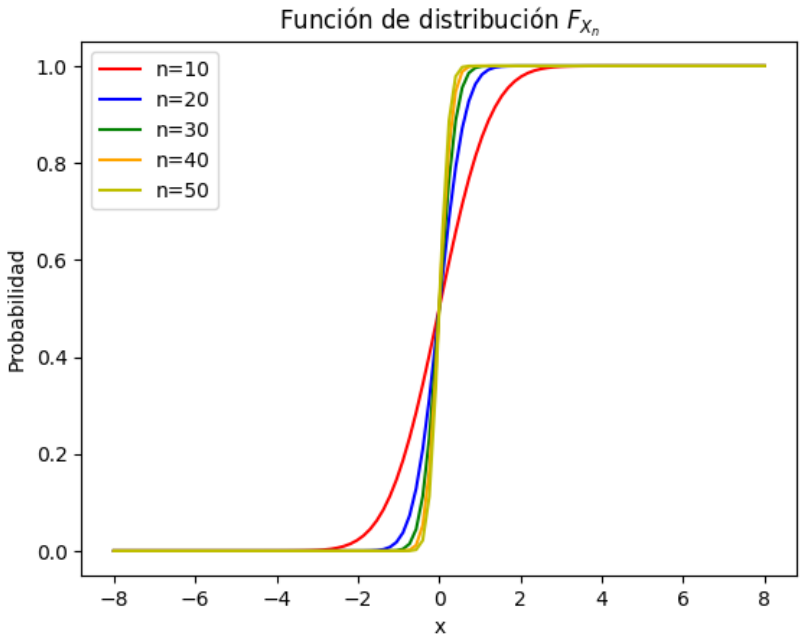


Figura 1.3: Función de distribución de X_n, donde $X_n \sim N(0, 10/n)$.

Los tipos de convergencias descritos hasta ahora no son equivalentes, pero se dan algunas implicaciones, estas se enuncian en el siguiente teorema.

Teorema 1.7.1. *Sean $X_1, X_2, \ldots,$ y X variables aleatorias definidas sobre el mismo espacio de probabilidad $(\omega, \mathcal{F}, P)$, con función de distribución $F_{X_1}, F_{X_2}, \ldots,$ y F_X, respectivamente. En general, se satisface que,*

1. *La convergencia casi segura, implica convergencia en probabilidad:*

$$X_n \xrightarrow{c.s.} X \Longrightarrow X_n \xrightarrow{P} X.$$

2. *La convergencia en probabilidad implica convergencia en distribución:*

$$X_n \xrightarrow{P} X \Longrightarrow X_n \xrightarrow{d} X.$$

De este modo podemos deducir que la convergencia casi segura, implica convergencia en distribución.

Para la demostración de este teorema se requiere conocer equivalencias de la convergencia en probabilidad, otras teorías estadísticas y teoría de la medida, lo cual esta fuera del alcance de esta tesis, por lo cual, para el propósito de este trabajo sólo se citará [24].

Por otra parte, aún cuando la convergencia en probabilidad no implica necesariamente la convergencia casi segura, si la sucesión de variables aleatorias es creciente, entonces la convergencia en probabilidad sí implica convergencia casi segura (con probabilidad 1).

Proposición 1.7.1. *Sea $X_1, X_2, \ldots$ una sucesión de variables aleatorias definidas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{F}, P)$ tales que, $X_n \leq X_{n+1}$ para cada $n \in \mathbb{N}$. Si $X_n \to X$ en probabilidad, entonces $X_n \to X$ con probabilidad 1.*

La convergencia en distribución no necesariamente implica la convergencia en probabilidad, no obstante, si el límite es una constante, entonces sí se satisface. Esto se enuncia a continuación.

Proposición 1.7.2. *Sean $X_1, X_2, \ldots,$ una sucesión de variables aleatorias definidas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{F}, P)$ y c una constante. Si $X_n \to c$ en distribución, entonces $X_n \to c$ en probabilidad.*

Demostración: La función de distribución de la variable aleatoria constante c, esta dada por:

$$F(x) = \begin{cases} 0, & si \quad x < c, \\ \\ 1, & si \quad x \geq c, \end{cases}$$

la cual es una función discontinua en $x = c$.
Supóngase que $\lim_{n\to\infty} F_{X_n}(x) = F(x)$ para cada $x \neq c$. Sea $\varepsilon > 0$, se tiene:

$$\begin{aligned} P(|X - c| \geq \varepsilon) &= P(X_n \leq c - \varepsilon) + P(X_n \geq c + \varepsilon) \\ &\leq P(X_n \leq c - \varepsilon) + P(X_n \geq c + \varepsilon/2) \\ &= F_{X_n}(c - \varepsilon) + 1 - F_{X_n}(c + \varepsilon/2), \end{aligned}$$

así, considerando la definición de $F(x)$,

$$\begin{aligned} \lim_{n\to\infty} P(|X - c| \geq \varepsilon) &= F(c - \varepsilon) + 1 - F(c + \varepsilon/2) \\ &= 0 + 1 - 1 = 0. \quad \blacksquare \end{aligned}$$

1.8. Leyes de los grandes números

A continuación se enuncian otros resultados muy utilizados en estadística. En estos se puede ver algunos de los tipos de convergencia descritos en la sección anterior.

Las leyes de los grandes números son resultados representativos de la estadística que permiten estudiar el comportamiento de la media muestral de una sucesión de variables aleatorias que tiende a infinito. Considerando algunas condiciones, estas leyes establecen que la media muestral de una sucesión de variables aleatorias converge a la media poblacional, cuando el número de sumandos tiende a infinito. Existen dos leyes: la ley débil y la ley fuerte de los grandes números que se diferencian y caracterizan por el tipo de convergencia. La ley débil se caracteriza por presentar la convergencia en probabilidad y la ley fuerte establece convergencia de forma casi segura. De este modo la ley fuerte implica entonces la ley débil.

Teorema 1.8.1. ***(Ley débil de los grandes números)*** *Sea $X_1, X_2, \ldots,$ una sucesión de variables aleatorias independientes tales que,*

$$E(X_n) = \mu, \ \ y\ V(x_n) = \sigma^2, \quad n = 1, 2, \ldots \tag{1.6}$$

Para cualquier $\varepsilon > 0$ se satisface que,

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) = 0. \tag{1.7}$$

Esto es, la sucesión de la media muestral converge a la media común μ de las variables $X_1, X_2, \ldots,$ en probabilidad.

Teorema 1.8.2. ***(Ley fuerte de los grandes números o Ley fuerte de Kolmogorov)*** *Sea $X_1, X_2, \ldots,$ una sucesión de variables aleatorias independientes tales que,*

$$E(X_n) = \mu < \infty, \quad n = 1, 2, \ldots,$$

donde C es una constante. La variable,

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}, \tag{1.8}$$

converge a μ casi seguramente.

Ejemplo 1.9. (Lanzamiento de una moneda) En este ejemplo las variables aleatorias son de tipo Bernoulli con parámetro $p = 0.5$. Se puede verificar gráficamente que las proporciones de caras cuando se lanza una moneda n veces, es decir, las medias muestrales $\overline{X}_n$ tienden casi seguramente hacia p, vea Figura (1.4).

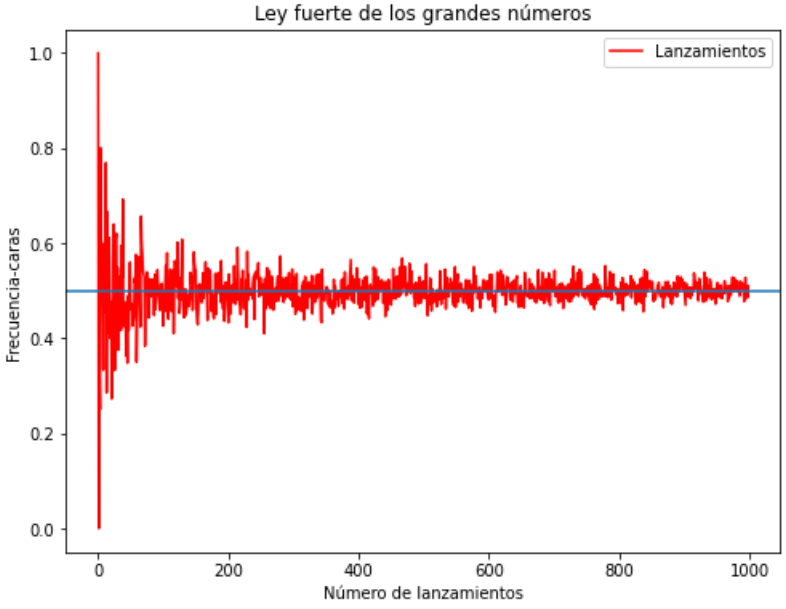


Figura 1.4: Lanzamiento de una moneda.

Uno de los teoremas más importantes en estadística por su amplio uso en diversas aplicaciones, es el teorema del límite central (TLC). El teorema del límite central permite realizar estudios probabilísticos con distribuciones de todo tipo. El enunciado básicamente expresa que sin importar la distribución de las variables aleatorias, la media muestral tiende a una distribución normal, sin embargo, esto ocurre cuando el tamaño de la muestra es suficientemente grande (es común considerar $n \geq 30$). Una prueba de este teorema puede consultarse en [36].

Teorema 1.9.1. ***(Teorema del límite central)***. *Sea $X_1, X_2, \ldots,$ una sucesión de variables aleatorias independientes e idénticamente distribuidas, tales que para cada $n \in \mathbb{N}$, $E(X_n) = \mu$ y $Var(X_n) = \sigma^2$. Se cumple:*

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1).$$

Este teorema también puede verse como:

$$\overline{X}_n = \frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow{d} N(\mu, \sigma^2/n).$$

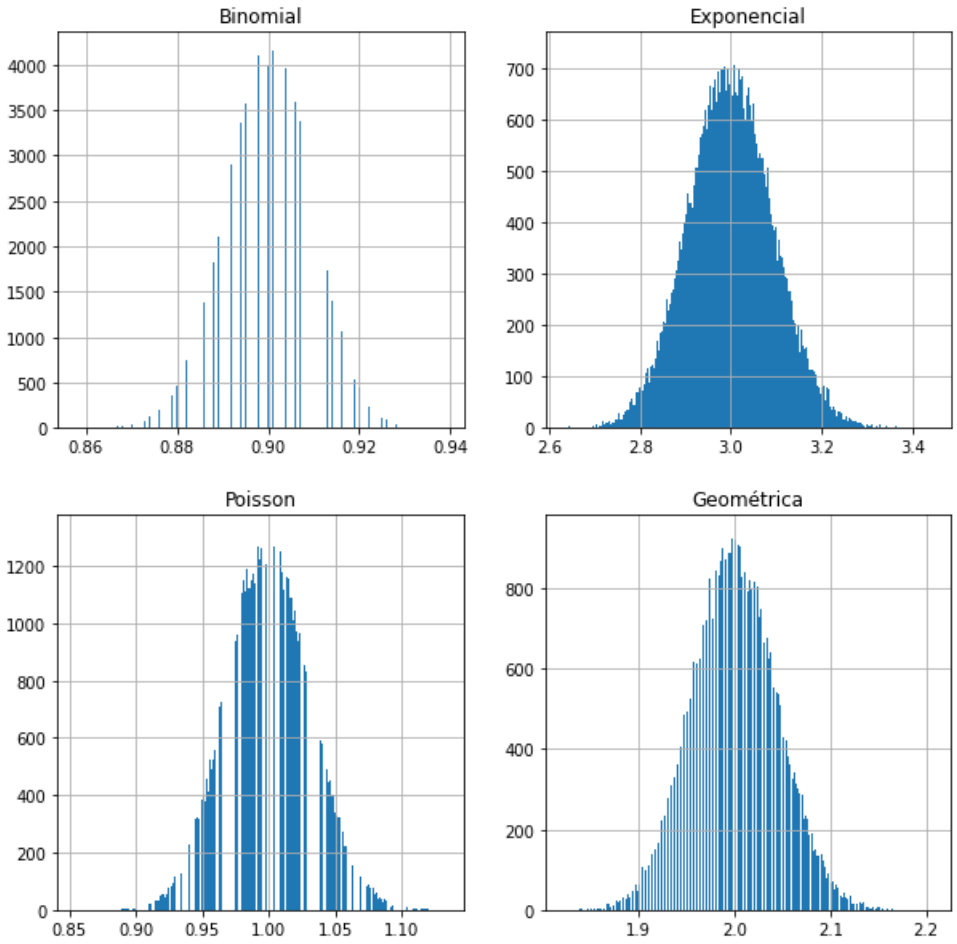


Figura 1.5: Histogramas de la media muestral para variables con función de distribución: Binomial, Exponencial, Poisson y Geométrica.

Con esto puede verse que, la distribución de la media muestral se puede obtener realizando una transformación o normalización con parámetros de localización y escala, más aun, su distribución no depende de la distribución de las variables originales X_i, vea Figura (1.5).

1.10. Introducción a la teoría de valores extremos

Una distribución de valor extremo permite modelar la distribución de los máximos y mínimos de un conjunto de datos observados, por lo cual, permite predecir qué tan grandes o pequeños serán los datos en determinado tiempo. Se presentan a continuación conceptos básicos acerca de la TVE para mayor comprensión de la misma. La información que se desarrolla en esta sección, se obtuvieron principalmente de [8], [16] y [33].

Definición 1.10.1. Sean $X_1, X_2, \ldots, X_n$ una sucesión de v.a.i.i.d., con función de distribución F, se definen el mínimo y máximo respectivamente como:

$$m_n = \text{mín}\{X_1, X_2, \ldots, X_n\}, \qquad y$$

$$M_n = \text{máx}\{X_1, X_2, \ldots, X_n\}.$$

Generalmente las variables X_i representan valores que son tomados en intervalos de tiempo de la misma longitud (días, semanas, meses, o años).

Considerando que las variables son independientes e idénticamente distribuidas, la función de distribución de las variables aleatorias m_n y M_n respectivamente, se pueden calcular de la siguiente forma,

$$\begin{aligned} P(m_n \leq x) &= 1 - P(X_1 > x, \ldots, X_n > x) \\ &= 1 - P(X_1 > x) \ldots P(X_n > x) \\ &= 1 - (1 - F(x))^n, \end{aligned}$$

$$\begin{aligned} P(M_n \leq x) &= P(X_1 \leq x, \ldots, X_n \leq x) \\ &= P(X_1 \leq x) \ldots P(X_n \leq x) \\ &= [F(x)]^n. \end{aligned} \tag{1.9}$$

De este modo, las funciones de densidades para m_n y M_n respectivamente son:

$$f_{m_n}(x) = n f_X(x)(1 - F_X(x))^{n-1},$$

$$f_{M_n}(x) = n f_X(x)(F_X(x))^{n-1}.$$

Con estos resultados, si bien, conociendo la distribución F se puede obtener fácilmente la función de distribución de los extremos, sin embargo, en las aplicaciones es común que F sea una función desconocida. Más aún, aunque se pudiera pensar que proporcionar un estimador para F puede ser una solución, esta propuesta no se considera ya que al elevar a la n-ésima potencia a $F(x)$, el cual puede ser pequeño o grande, se genera grandes cambios en $F(x)^n$. Por tal motivo, se recurre al estudio de las distribuciones asintóticas de los valores extremos m_n y M_n, por lo cual se proporcionan otros conceptos importantes de conocer.

Definición 1.10.2. Sea X una variable aleatoria con función de distribución F. Se definen al extremo izquierdo y al extremo derecho de la función de distribución F respectivamente, como los puntos:

$$\alpha(F) = \text{ínf}\{x : F(x) > 0\} \geq -\infty, \quad y$$

$$\omega(F) = \sup\{x : F(x) < 1\} \leq \infty.$$

Definición 1.10.3. (Función de distribución degenerada) Una variable aleatoria X es degenerada en un valor real $c \in \mathbb{R}$, si toma dicho valor con probabilidad 1, es decir $P(X = c) = 1$. Así, su función de distribución esta dada por:

$$F_X(x) = \begin{cases} 0, & si \quad x < c, \\ \\ 1, & si \quad x \geq c. \end{cases}$$

Observación 1.11. La sucesión de los máximos M_n, $n \in \mathbb{N}$ es creciente con límite en $\omega(F)$ y con probabilidad 1. En efecto, sean $X_1, X_2, \ldots, X_n, X_{n+1}$ una sucesión de v.a.'s. Considere una realización $x_1, x_2, \ldots, x_n, x_{n+1}$ de dichas v.a.'s. Luego,

$$M_n = \{x_1, x_2, \ldots, x_n\} \quad y$$

$$M_{n+1} = \{x_1, x_2, \ldots, x_n, x_n + 1\}.$$

Nótese que, puede ocurrir que $M_n \geq x_{n+1}$, de donde $M_n = M_{n+1}$. En caso contrario si $M_n < x_{n+1}$, entonces $M_n < M_{n+1}$. Por tanto $M_n \leq M_{n+1}$ y la sucesión es creciente.

Por otra parte, puede verificarse formalmente que la distribución asintótica del máximo es hacia una función de distribución degenerada, véase [33]. Así, para todo x en el dominio de F, tal que: $x < \sup\{x : F(x) < 1\} = \omega(F)$, ocurre que $F(x) < 1$, de donde $F(x)^n$ converge a 0 cuando $n \to \infty$, esto es, la función de distribución límite del máximo M_n converge a la función degenerada en $\omega(F)$:

$$P(M_n \leq x) = F(x)^n \to \begin{cases} 0, & si \quad x < \omega(F), \\ \\ 1, & si \quad x \geq \omega(F). \end{cases}$$

Por lo tanto, $M_n \to \omega(F)$ en probabilidad y dado que la sucesión es creciente, la convergencia en probabilidad implica convergencia con probabilidad 1, ver Proposición 1.7.1. Esto significa que la función de distribución del máximo converge a una distribución degenerada a un punto, sea cual sea la función de distribución de las variables aleatorias, Figura (1.6). De forma análoga se determina para la distribución del mínimo. Para ver la demostración formalmente puede consultar [8].

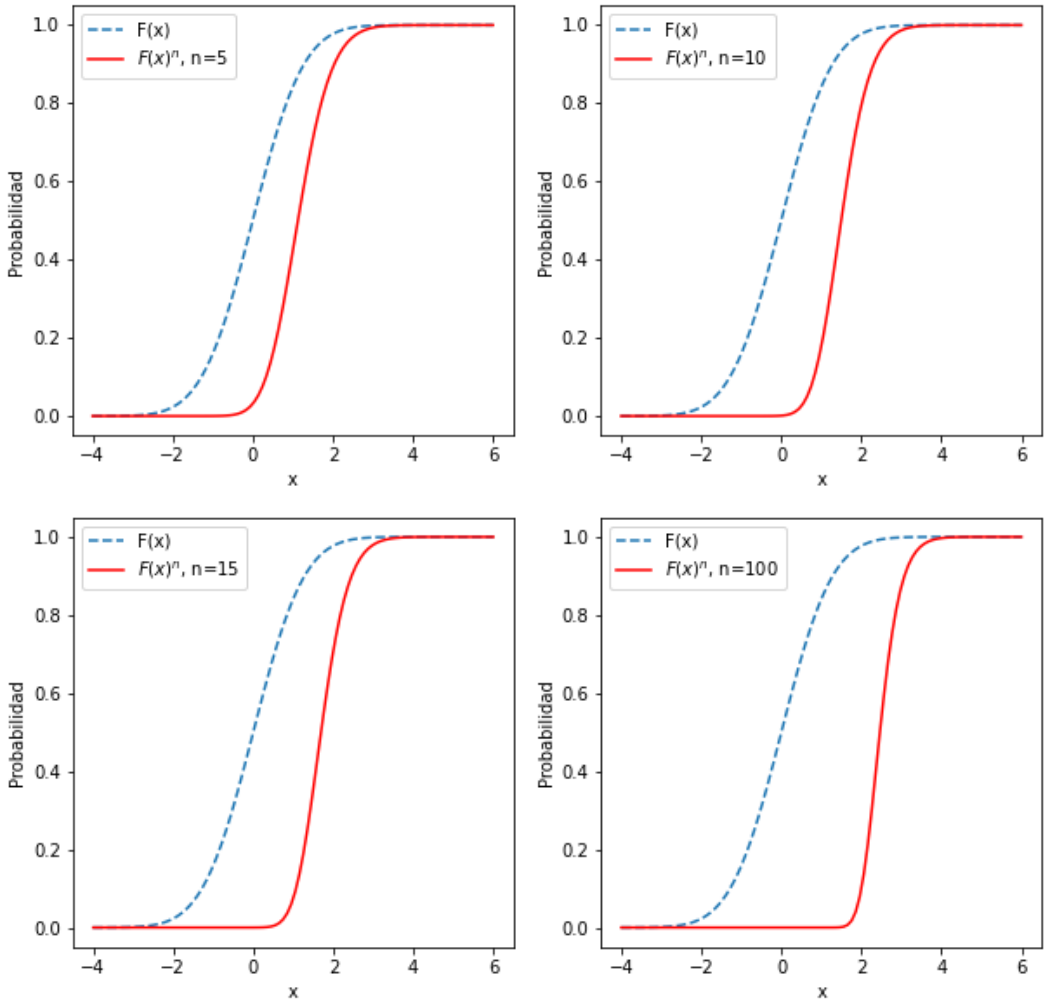


Figura 1.6: Linea azul discontinua: Función de distribución de una normal. Linea roja continua: función de distribución del máximo para $n = 5, 10, 15$ y 100.

Para atender el problema de la distribución degenerada, Fisher y Tippet en 1928 presentan un teorema análogo al teorema del límite central, el cual aproxima la distribución de la media muestral, Teorema 1.9.1. Este teorema, posteriormente es demostrado por Gnedenko en 1943, de aquí el nombre del teorema Fisher-Tippett-Gnedenko. La idea principal, reside en realizar una

transformación o normalización con parámetros de localización y escala que permitan obtener una distribución no degenerada en el límite. En términos generales, se buscan constantes $a_n > 0$ y b_n tal que,

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) = F^n(a_n x + b_n) \to G(x),$$

en distribución, cuando $n \to \infty$.

Para estudiar las características o propiedades que deben satisfacer estas constantes se analizan y describen algunos resultados.

Teorema 1.11.1. *(**Teorema de Fisher-Tippett-Gnedenko**). Sean $X_1, \ldots, X_n$ una sucesión de v.a.i.i.d. y $M_n = \max\{X_1, X_2, \ldots, X_n\}$. Si existen un par de sucesiones $\{a_n\}_{n\in\mathbb{N}}$ y $\{b_n\}_{n\in\mathbb{N}}$, donde $a_n > 0$, y $b_n \in \mathbb{R}$, tales que,*

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) = F^n(a_n x + b_n) \to G(x), \text{ cuando } n \to \infty,$$

con $G(x)$ una función de distribución no degenerada, entonces $G(x)$ pertenece a alguna de las siguientes tres familias paramétricas:

Distribución de Gumbel:

$$G(x) = exp\left(-exp\left(-\frac{x-\mu}{\sigma}\right)\right), \quad x \in \mathbb{R}. \tag{1.10}$$

Distribución de Fréchet:

$$G(x) = \begin{cases} 0, & si \quad x \leq \mu, \\ exp\left(-\left(\frac{x-\mu}{\sigma}\right)^{-\beta}\right), & si \quad x > \mu. \end{cases} \tag{1.11}$$

Distribución de Weibull:

$$G(x) = \begin{cases} exp\left(-\left(-\left(\frac{x-\mu}{\sigma}\right)\right)^{\beta}\right), & si \quad x < \mu, \\ 1, & si \quad x \geq \mu. \end{cases} \tag{1.12}$$

donde $\alpha > 0$, $\beta > 0$ y $\mu \in \mathbb{R}$.

Así, las funciones de densidades (vea Figura (1.7)) están dadas por:

Distribución de Gumbel:

$$g(x; \mu, \sigma) = \frac{1}{\sigma} \exp\left\{-\exp\left(\frac{\mu - x}{\sigma}\right) + \frac{\mu - x}{\sigma}\right\}, \qquad x \in \mathbb{R}.$$

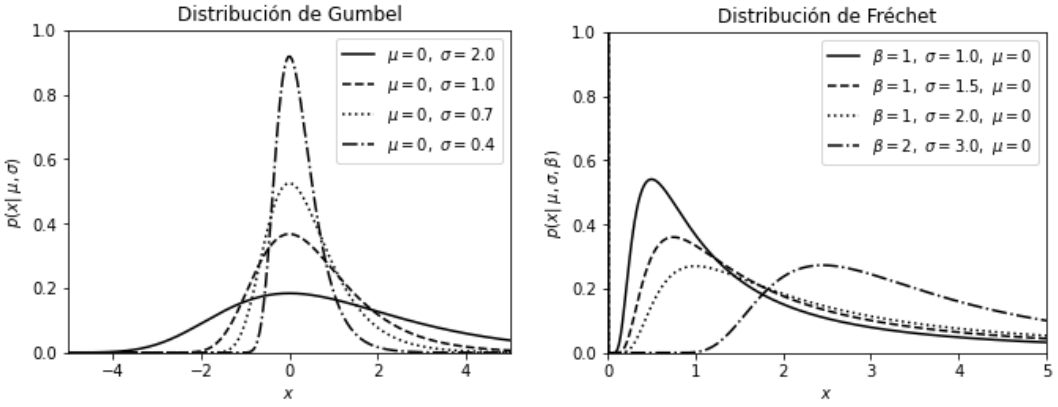
Distribución de Fréchet:

$$g(x; \mu, \sigma, \beta) = \frac{\beta}{\sigma}\left(\frac{x-\mu}{\sigma}\right)^{-(1+\beta)} \exp\left[-\left(\frac{x-\mu}{\sigma}\right)^{\beta}\right], \qquad x > \mu.$$

Distribución de Weibull:

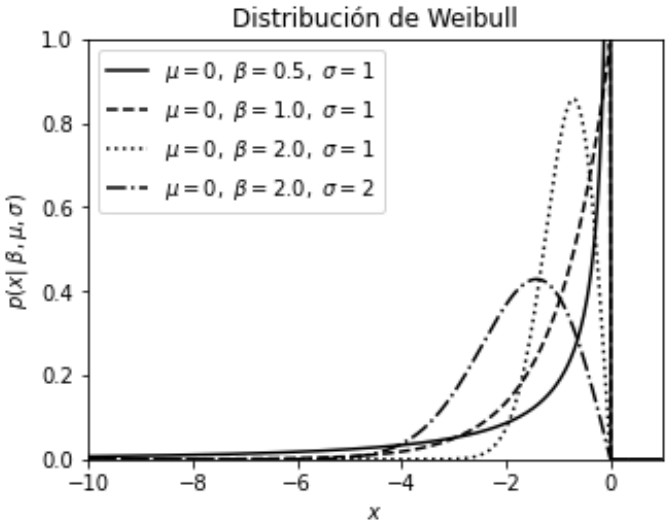
$$g(x;\mu,\sigma,\beta) = \frac{\beta}{\sigma}\left(\frac{\mu - x}{\sigma}\right)^{\beta-1} \exp\left[-\left(\frac{\mu - x}{\sigma}\right)^{\beta}\right], \qquad x < \mu.$$

Para ver la prueba formal de este Teorema (1.11.1) puede consultar [16], Teorema 1.1.3.



(a)

(b)



(c)

Figura 1.7: Función de densidad para el máximo normalizado: (a) Gumbel, (b) Fréchet y (c) Weibull.

A las tres distribuciones asintóticas del Teorema (1.11.1) se les denomina Distribuciones de Valores Extremos (DVE). Nótese que el teorema no dice algo acerca de la distribución del máximo, sin embargo, establece que, si la variable M_n puede ser normalizado mediante las sucesiones $\{a_n\}$ y $\{b_n\}$, entonces la correspondiente normalización de M_n tiene una función de distribución que converge a sólo una de las tres distribuciones para valores extremos, lo cual en la práctica es sumamente útil. De igual manera, cabe resaltar que nada dice

acerca de cómo determinar a las sucesiones $\{a_n\}_{n\in\mathbb{N}}$ y $\{b_n\}_{n\in\mathbb{N}}$, las cuales no son únicas, como se verá más adelante.

Ejemplo 1.12. Considere $X_1, X_2, \ldots, X_n$ una sucesión de v.a.i.i.d. con función de distribución exponencial de parámetro $\lambda > 0$:

$$F(x) = 1 - \exp(-\lambda x), \ x > 0.$$

Considere las variables aleatorias $M_n = \text{máx}\{X_1, X_2, \ldots, X_n\}$. Tomando las sucesiones $a_n = \lambda^{-1}$ y $b_n = \lambda^{-1} \ln n$, estas satisfacen:

$$\begin{aligned} P\left(\frac{M_n - b_n}{a_n} \leq x\right) &= P\left(M_n \leq a_n x + b_n\right) \\ &= P\left(M_n \leq \lambda^{-1}x + \lambda^{-1}\ln n\right), \end{aligned}$$

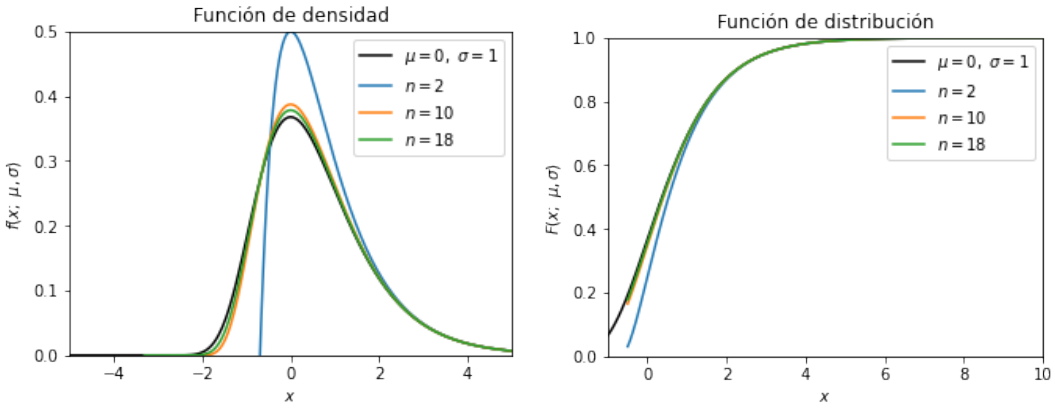
y por Teorema (1.11.1),

$$\begin{aligned} P\left(\frac{M_n - b_n}{a_n} \leq x\right) &= \left(F(\lambda^{-1}x + \lambda^{-1}\ln n)\right)^n \\ &= \left(1 - \exp(-\lambda(\lambda^{-1}x + \lambda^{-1}\ln n))\right)^n \\ &= \left(1 - (\exp(-x))(\exp(\ln n^{-1}))\right)^n \\ &= \left(1 - \frac{\exp(-x)}{n}\right)^n, \end{aligned}$$

de donde, para cada x fijo,

$$\lim_{n\to\infty} P(M_n^* \leq x) = \exp(-\exp(-x)) = G(x),$$

donde $M_n^* = \frac{M_n - b_n}{a_n}$, y $G(x)$ es una distribución de Gumbel de parámetros $\mu = 0$ y $\beta = 1$. Vea Figura (1.8).



(a) (b)

Figura 1.8: a) Función de densidad de Gumbel y función de densidad de M^* para $n = 2, 10, 18$. b) Función de distribución de Gumbel y función de distribución de M^* para $n = 2, 10$ y 18.

Ejemplo 1.13. Sea $X_1, X_2, \ldots, X_n$ una sucesión de v.a.i.i.d. con función de distribución uniforme de parámetros $a = 0$ y $b = 1$. Para cada valor fijo $z < 0$, supóngase que $n > -z$. Tomando las sucesiones $a_n = 1/n$ y $b_n = 1$, estas satisfacen,

$$\begin{aligned} P\left(\frac{M_n - b_n}{a_n} \leq x\right) &= P(M_n \leq a_n x + b_n) \\ &= P\left(M_n \leq \frac{z}{n} + 1\right). \end{aligned}$$

Por otra parte,

$$\begin{aligned} P\left(\frac{M_n - b_n}{a_n} \leq x\right) &= \left(F\left(\frac{x}{n} + 1\right)\right)^n \\ &= \left(1 + \frac{x}{n}\right)^n. \end{aligned}$$

Luego,

$$\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = \exp(x).$$

Así, la distribución en el límite es una distribución de tipo Weibull, de parámetro $\mu = 0$, $\sigma = 1$ y $\beta = 1$.

Otro aspecto importante de resaltar sobre el Teorema de Valores Extremos, es que no proporciona herramientas para conocer con exactitud a cual de las 3 familias paramétricas converge la distribución del máximo normalizado. Debido a esto, en la práctica se asume una sola distribución, y posteriormente se estiman los parámetros del modelo seleccionado. Sin embargo, existe una reformulación del teorema de valores extremos que es de gran utilidad, la cual establece que es posible expresar estás familias en una sola distribución, esto se describe a continuación.

Distribución de valor extremo generalizado

Del Teorema de Valores Extremos, se sabe que la convergencia del máximo (o mínimo) considerando una transformación o normalización, es sólo a una de las tres familias paramétricas descritas en (1.10), (1.11) y (1.12). Una reparametrización de estas 3 familias es propuesta por Richard Von mises en 1954 y Arthur F. Jenkinson en 1955 en [27]. Estas tres distribuciones se asocian a una sola distribución denominada distribución de valores extremos generalizada (DVEG), definida como:

$$G(x) = \begin{cases} \exp\left\{-\left(1 + \kappa\frac{(x-\mu)}{\sigma}\right)^{-\frac{1}{\kappa}}\right\}, & \kappa \neq 0;\ 1 + \kappa\frac{(x-\mu)}{\sigma} > 0, \\ \exp\left\{-\exp\left[\frac{-(x-\mu)}{\sigma}\right]\right\}, & \kappa = 0, \sigma > 0, x, \mu \in \mathbb{R}. \end{cases}$$

Así, la función de densidad para $\kappa \neq 0$ y $1+\kappa\frac{(x-\mu)}{\sigma} > 0$,

$$g(x) = \frac{1}{\sigma}\left\{\left(1+\kappa\frac{(x-\mu)}{\sigma}\right)^{-\left(\frac{1}{\kappa}+1\right)}\right\}\exp\left\{-\left(1+\kappa\frac{(x-\mu)}{\sigma}\right)^{-\frac{1}{\kappa}}\right\}. \quad (1.13)$$

Para $\kappa = 0$, la función de densidad está dado por:

$$g(x;\mu,\sigma) = \frac{1}{\sigma}\exp\left\{-\exp\left(\frac{\mu-x}{\sigma}\right)+\frac{\mu-x}{\sigma}\right\}. \quad (1.14)$$

En este modelo μ representa al parámetro de localización, σ el parámetro de escala y κ es el parámetro de forma. Nótese que, si $\kappa > 0$ entonces la distribución $G(x)$ es de Fréchet, para $\kappa < 0$ se tiene la distribución de Weibull, mientras que, cuando $\kappa \to 0$ se obtiene la distribución de Gumbel [8]. Así, si X es una variable aleatoria que sigue una distribución de valores extremos generalizado de parámetros μ, σ y κ, esto se denota como $X \sim VEG(\mu,\sigma,\kappa)$.

Con todo esto, el Teorema de VE puede ser reformulado de la siguiente forma.

Teorema 1.13.1. ***(Teorema de VEG)*** *Sean $X_1,\ldots,X_n$ una sucesión de v.a.i.i.d. y $M_n = \max\{X_1,X_2,\ldots,X_n\}$. Si existen un par de sucesiones $\{a_n\}_{n\in\mathbb{N}}$ y $\{b_n\}_{n\in\mathbb{N}}$, donde $a_n > 0$, y $b_n \in \mathbb{R}$, tales que,*

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) = F^n(a_n x + b_n) \to G(x), \text{ cuando } n \to \infty,$$

con $G(x)$ una función de distribución no degenerada, entonces $G(x)$ pertenece a la familia de VEG:

$$G(x) = \begin{cases} \exp\left\{-\left(1+\kappa\frac{(x-\mu)}{\sigma}\right)^{-\frac{1}{\kappa}}\right\}, & \kappa \neq 0;\ 1+\kappa\frac{(x-\mu)}{\sigma} > 0, \\ \exp\left\{-\exp\left[\frac{-(x-\mu)}{\sigma}\right]\right\}, & \kappa = 0, \sigma > 0, x, \mu \in \mathbb{R}. \end{cases} \quad (1.15)$$

Observación 1.14. La mayor parte del estudio aborda la TVE a partir de las observaciones máximas, esto debido a que los mismos resultados pueden ser utilizados en las observaciones de los mínimos, esto, considerando la relación existente entre el máximo y el mínimo. Si $M_n = \max\{X_1,X_2,\ldots,X_n\}$ y $m_n = \min\{-X_1,-X_2,\ldots,-X_n\}$ entonces,

$$M_n = \max\{X_1,X_2,\ldots,X_n\} = -\min\{-X_1,-X_2,\ldots,-X_n\} = -m_n.$$

De esta forma, para modelar la función de distribución del mínimo, se puede hacer uso de los mismos resultados asintóticos del máximo para aproximar a su función de distribución. Esto se resume en el siguiente teorema.

Teorema 1.14.1. ***(Teorema de valores extremos para mínimos)*** *Sean* $\{X_1, \ldots, X_n\}$ *una sucesión de v.a.i.i.d. y* $m_n = \text{mín}\{X_1, X_2, \ldots, X_n\}$. *Si existen un par de sucesiones* $\{a_n\}_{n\in\mathbb{N}}$ *y* $\{b_n\}_{n\in\mathbb{N}}$, *donde* $a_n > 0$, *y* $b_n \in \mathbb{R}$, *tales que,*

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) = 1 - [1 - F(a_n x + b_n)]^n \to \widetilde{G}(x), \text{ cuando } n \to \infty,$$

con $\widetilde{G}(x)$ *una función de distribución no degenerada, entonces la función* $\widetilde{G}(x)$ *pertenece a alguna de las siguientes tres familias paramétricas:*

Distribución de Gumbel:

$$\widetilde{G}(x) = 1 - exp\left(-exp\left(-\frac{(-x-\mu)}{\sigma}\right)\right), \quad x \in \mathbb{R}.$$

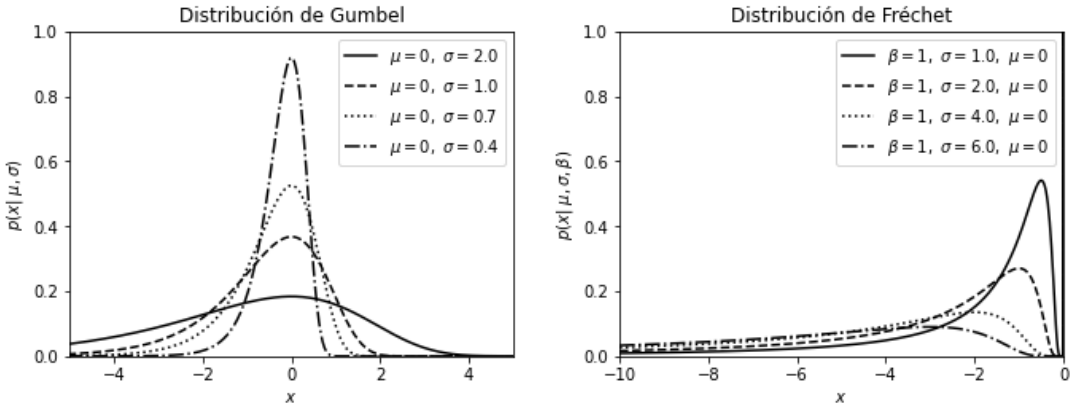
Distribución de Fréchet:

$$\widetilde{G}(x) = \begin{cases} 1, & si \quad x \geq \mu, \\ 1 - exp\left(-\left(\frac{-x-\mu}{\sigma}\right)^{-\beta}\right), & si \quad x < \mu. \end{cases}$$

Distribución de Weibull:

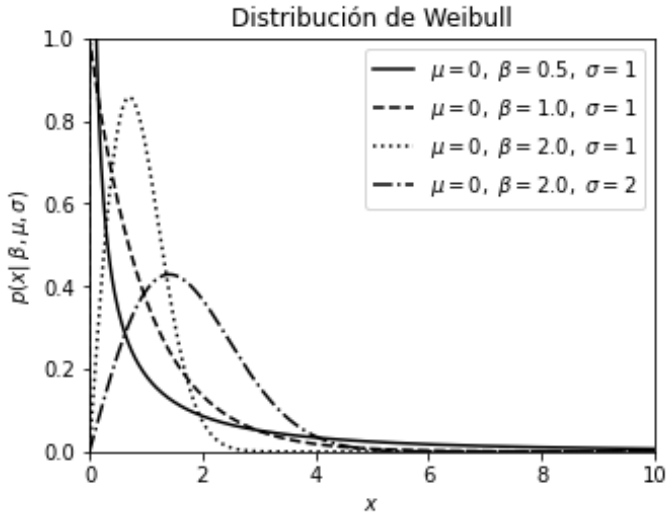
$$\widetilde{G}(x) = \begin{cases} 1 - exp\left(-\left(-\left(\frac{-x-\mu}{\sigma}\right)\right)^{\beta}\right), & si \quad x > \mu, \\ 0, & si \quad x \leq \mu. \end{cases}$$

donde $\alpha > 0$, $\beta > 0$ *y* $\mu \in \mathbb{R}$.



(a)

(b)



(c)

Figura 1.9: Función de densidad para el mínimo: (a) Gumbel, (b) Fréchet y (c) Weibull.

1.15. Demostración del teorema de valores extremos

A continuación se define el concepto de Max-estabilidad, y se enuncian algunos teoremas relacionados con la TVE necesarios para presentar una prueba del teorema de VE usando la relación existente con estos enunciados.

Definición 1.15.1. (Max-estabilidad) Una función de distribución $F(x)$ se dice que es max-estable si, para cada $n = 2, 3, \ldots$ existen sucesiones de constantes $\{a_n > 0\}$ y $\{b_n\}$ tales que,

$$F^n(a_n x + b_n) = F(x),$$

o equivalentemente, $F(a_n x + b_n) = F^{1/n}(x)$.

La relación entre el concepto de Max-estabilidad, las DVE y las DVEG, se puede ver mediante los teoremas siguientes. Cabe mencionar que dado los requerimientos para la demostración de los siguientes teoremas, estas no se probarán.

Teorema 1.15.2. *Una distribución es Max-estable si y sólo si es una distribución de valores extremos.*

Por su parte, considerando el hecho de que la distribución de valores extremos generalizados representa a las 3 funciones de distribuciones de la TVE, entonces este teorema puede enunciarse como sigue.

Teorema 1.15.3. *Una distribución es Max-estable si y sólo si es una distribución de valores extremos generalizada.*

Como puede observarse, el resultado muestra que la distribución del máximo de v.a.i.i.d. mediante una normalización o transformación, tiene la misma distribución F, siempre que F pertenezca a una de las distribuciones de valores extremos. Además, con esto puede deducirse que las distribuciones Gumbel, Fréchet y Weilbull son las únicas distribuciones Max-estables.

Por otra parte, puede verificarse que, dada una sucesión de variables aleatorias $\{\varepsilon_n\}_{n\in\mathbb{N}}$ y sucesiones de constantes $a_n > 0$ y $b_n \in \mathbb{R}$, se cumple que,

$$\frac{\varepsilon_n - b_n}{a_n} \xrightarrow{d} Y, \qquad \text{cuando } n \to \infty,$$

donde Y es una v.a. no degenerada.

Bajo estos supuestos,

$$\lim_{n\to\infty} P\left(\frac{\varepsilon_n - b_n}{a_n} \leq y\right) = G(y),$$

donde $G(y)$ es la función de distribución no degenerada de la variable aleatoria Y. Así, para n suficientemente grande se cumple que:

$$P\left(\frac{\varepsilon_n - b_n}{a_n} \leq y\right) \approx P(Y \leq y) = G(y).$$

Luego, si se considera $z = a_n y + b_n$, entonces $y = (z - b_n)/a_n$, así, sustituyendo:

$$P\left(\varepsilon_n \leq a_n y + b_n\right) \approx P\left(Y \leq \frac{z - b_n}{a_n}\right) = G\left(\frac{z - b_n}{a_n}\right).$$

Como puede deducirse, esto permite aproximar la distribución de la variable ε_n por una familia de distribuciones considerando parámetros de localización y escala.

Definición 1.15.4. Dos distribuciones F y G son del mismo tipo o pertenecen a la misma familia de distribución si para algunas constantes $a > 0$, $b \in \mathbb{R}$,

$$G(x) = F(ax + b), \qquad x \in \mathbb{R}.$$

En términos de variables aleatorias, si $X \sim F$ y $Y \sim G$ entonces,

$$Y \xrightarrow{d} \frac{X - b}{a}.$$

Teorema 1.15.5. ***(Convergencia a familias, Gnedenko y Khinchin)*** *Sean $G(x)$ y $H(x)$ dos funciones de distribución propias no degeneradas a un punto. Sean, X_n $n = 1, 2, \ldots$, variables aleatorias con funciones de distribución F_n y constantes $a_n > 0$, $b_n \in \mathbb{R}$, $\alpha_n > 0$ y $\beta_n \in \mathbb{R}$;*

- *a) Si,*

$$\begin{aligned} F_n(a_n x + b_n) &\to G(x), \ y \\ F_n(\alpha_n x + \beta_n) &\to H(x), \end{aligned} \tag{1.16}$$

entonces existen constantes $A > 0$ y $B \in \mathbb{R}$ tales que, cuando $n \to \infty$,

$$\frac{\alpha_n}{a_n} \to A > 0, \tag{1.17}$$

$$\frac{\beta_n - b_n}{a_n} \to B, \ y \tag{1.18}$$

$$H(x) = G(Ax + B). \tag{1.19}$$

- *b) Recíprocamente, si* (1.17) *y* (1.18) *son validas, entonces cualquiera de las dos relaciones en (1.16) implica la otra y* (1.19) *también es valida.*

La prueba de este teorema (1.15.5) puede consultarlo en [31].

Corolario 1.15.6. *Sean F_n una sucesión de funciones de distribuciones, $a_n > 0$ y b_n sucesiones de constantes tales que,*

$$F_n(a_n x + b_n) \to G(x), \tag{1.20}$$

en todo punto de continuidad de G, que es una función de distribución propia y no degenerada en un punto. Sean $c_n > 0$ y d_n sucesiones de constantes tales que,

$$\frac{a_n}{c_n} \to 1, \qquad \frac{d_n - b_n}{a_n} \to 0.$$

Entonces (1.20) se cumple con c_n y d_n en lugar de a_n y b_n.

Como se mencionó anteriormente, las sucesiones de constantes a_n y b_n del TVE no son únicas, ya que del teorema de convergencia a familias se deduce que estas constantes de normalización están determinadas excepto por equivalencias asintóticas, y la distribución límite está determinada excepto por los parámetros de localización y escala.

Teorema 1.15.7. *Suponga que existen sucesiones de constantes $a_n > 0$ y $b_n \in \mathbb{R}$ para todo $n \in \mathbb{N}$ de tal forma que,*

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G(x), \text{ para todo } x \in \mathbb{R}, \tag{1.21}$$

donde $G(x)$ es una función de distribución no degenerada. Entonces, para algún entero positivo m se cumple,

$$\lim_{n\to\infty} F^{nm}(a_{nm} x + b_{nm}) = G(x), \text{ para todo } x \in \mathbb{R}, \tag{1.22}$$

o de forma equivalente,

$$\lim_{n\to\infty} F^{n}(a_{nm} x + b_{nm}) = G^{1/m}(x), \text{ para todo } x \in \mathbb{R}, \tag{1.23}$$

donde $G^{1/m}(x)$ es también una función de distribución no degenerada.

A continuación, se presenta un análisis de [33], en donde se puede ver la relación entre las distribuciones de VE y las distribuciones Max-estables.

Considere una sucesión de variables aleatorias independientes $X_1, X_2, \ldots, X_n, X_{n+1}, \ldots, X_{nk}$. Se definen las siguientes variables aleatorias, para $n \in \mathbb{N}$ y k entero positivo:

$$\begin{aligned}
M_{n1} &= \text{máx}\{X_1, X_2, \ldots, X_n\}, \\
M_{n2} &= \text{máx}\{X_{n+1}, X_{n+2}, \ldots, X_{2n}\}, \\
M_{n3} &= \text{máx}\{X_{2n+1}, X_{2n+2}, \ldots, X_{3n}\}, \\
&\vdots \\
M_{nk} &= \text{máx}\{X_{(k-1)n+1}, X_{(k-1)n+2}, \ldots, X_{kn}\}.
\end{aligned}$$

Las variables aleatorias independientes M_{ni}, $i = 1, 2, \ldots, k$ representan a los valores máximos por bloques de longitud n. A continuación se define también la siguiente variable aleatoria:

$$M_k = \text{máx}\{M_{n1}, M_{n2}, \ldots, M_{nk}\}, \tag{1.24}$$

de este modo el máximo M_k representa al máximo de las nk variables aleatorias o al máximo por bloques. Luego, supóngase que existen sucesiones $a_n > 0$ y $b_n \in \mathbb{R}$ tales que,

$$\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = G(x), \tag{1.25}$$

Luego, para n suficientemente grande,

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \approx G(x), \tag{1.26}$$

así, para cualquier entero positivo k, se cumple también que nk es grande, y por el teorema (1.15.7) se tiene,

$$P\left(\frac{M_k - b_{nk}}{a_{nk}} \leq x\right) \approx G(x). \tag{1.27}$$

Por otra parte, nótese que la variable M_k tiene la misma distribución que las variables M_{ni} y las v.a's M_{ni} son independientes para cada $i = 1, \ldots, k$. De esto se deduce entonces:

$$\begin{aligned}
P\left(\frac{M_k - b_n}{a_n} \leq x\right) &= P(M_k \leq a_n x + b_n) \\
&= P(\text{máx}\{M_{n1}, \ldots, M_{nk}\} \leq a_n x + b_n) \\
&= P(M_{n1} \leq a_n x + b_n, \ldots, M_{nk} \leq a_n x + b_n) \\
&= \prod_{i=1}^{k} P(M_{ni} \leq a_n x + b_n) \\
&= P(M_n \leq a_n x + b_n)^k \\
&= \left(P\left(\frac{M_n - b_n}{a_n} \leq x\right)\right)^k.
\end{aligned}$$

Por lo tanto,

$$P\left(\frac{M_k - b_n}{a_n} \leq x\right) \approx G^k(x). \tag{1.28}$$

Por otra parte, si X es la variable aleatoria cuya función de distribución es $G(x)$, y la distribución límite de $M_n - b_n/a_n$ es $G(x)$, por el teorema de convergencia la familia $M_n - b_n/a_n$ converge en probabilidad a la variable aleatoria X no degenerada.

Luego, de (1.27) y tomando $x_1 = a_{nk}x + b_{nk}$ se deduce lo siquiente:

$$\begin{aligned}
& P\left(\frac{M_k - b_{nk}}{a_{nk}} \leq x\right) \approx P(X \leq x) \\
\iff\ & P\left(\frac{M_k - b_{nk}}{a_{nk}} \leq \frac{x_1 - b_{nk}}{a_{nk}}\right) \approx P\left(X \leq \frac{x_1 - b_{nk}}{a_{nk}}\right) \\
\iff\ & P(M_k \leq x_1) \approx G\left(\frac{x_1 - b_{nk}}{a_{nk}}\right)
\end{aligned}$$

y de (1.28), considere $x_1 = a_n x + b_n$, se obtiene:

$$
\begin{aligned}
& P\left(\frac{M_k - b_n}{a_n} \leq x\right) \approx G^k(x) \\
\iff \quad & P\left(\frac{M_k - b_n}{a_n} \leq x\right) \approx [P(X \leq x)]^k \\
\iff \quad & P\left(\frac{M_k - b_n}{a_n} \leq \frac{x_1 - b_n}{a_n}\right) \approx \left[P\left(X \leq \frac{x_1 - b_n}{a_n}\right)\right]^k \\
\iff \quad & P(M_k \leq x_1) \approx \left[P\left(X \leq \frac{x_1 - b_n}{a_n}\right)\right]^k \\
\iff \quad & P(M_k \leq x_1) \approx G^k\left(\frac{x_1 - b_n}{a_n}\right)
\end{aligned}
$$

Así,

$$G\left(\frac{x_1 - b_{nk}}{a_{nk}}\right) \approx G^k\left(\frac{x_1 - b_n}{a_n}\right).$$

Con esta última expresión y tomando $x_1 = a_{nk}x_2 + b_{nk}$ se tiene lo siguiente:

$$
\begin{aligned}
& P\left(X \leq \frac{x_1 - b_{nk}}{a_{nk}}\right) \approx P\left(X \leq \frac{x_1 - b_n}{a_n}\right)^k \\
\iff & P\left(X \leq \frac{a_{nk}x_2 + b_{nk} - b_{nk}}{a_{nk}}\right) \approx P\left(X \leq \frac{a_{nk}x_2 + b_{nk} - b_n}{a_n}\right)^k \\
\iff & P(X \leq x_2) \approx \left[P\left(X \leq \left(\frac{a_{nk}}{a_n}\right)x_2 + \left(\frac{b_{nk} - b_n}{a_n}\right)\right)\right]^k.
\end{aligned}
$$

De donde:

$$G(x_2) \approx G^k\left(\left(\frac{a_{nk}}{a_n}\right)x_2 + \left(\frac{b_{nk} - b_n}{a_n}\right)\right),$$

o bien,

$$G(x_2) \approx G^k(a_m x_2 + b_m),$$

luego, considere $a_m = a_{nk}/a_n$ y $b_m = (b_{nk} - b_n)/a_n$. Finalmente, por el Teorema (1.15.7) se tiene que la distribución $G(x)$ es Max-estable y por (1.15.2) es de valores extremos.

1.16. Distribución de valores extremos no estacionario

Hasta ahora se ha supuesto que el fenómeno en estudio está representado por una sucesión de variables independientes e idénticamente distribuidas. Sin embargo, en muchos fenómenos de la realidad, estos supuestos difícilmente se satisfacen, sin embargo, se puede tener un comportamiento estacionario en el

tiempo. Más aún, el supuesto más natural de una sucesión de variables aleatorias independientes es la de una serie estacionaria. La estacionariedad es más realista ya que este corresponde a una serie cuyas variables pueden ser mutuamente dependientes, pero cuyas propiedades estocásticas son homogéneas a lo largo del tiempo [8].

La dependencia en las series estacionarias puede tomar muchas formas diferentes y es imposible desarrollar una caracterización general del comportamiento de los valores extremos a menos que se impongan algunas restricciones [8]. Así, en la practica, lo habitual es suponer una condición que limite el grado de dependencia, esto es, considerar los eventos extremos que son aproximadamente independientes para tiempos suficientemente distantes. Muchas series estacionarias satisfacen esta propiedad. Más aún, dentro de la teoría de extremos, se puede demostrar que, en cierto sentido y sujeto a limitaciones específicas, los modelos de límite de valores extremos habituales, siguen siendo aplicables en presencia de dependencia temporal.

Por otro lado, en los procesos no estacionarios, dado que estos tienen características que cambian constantemente a lo largo del tiempo, no se pueden establecer las teorías de VE para procesos no estacionarios. A pesar de ello, [8] señala que resultados están disponibles para algunas formas muy especializadas de no estacionariedad, pero generalmente son demasiado restrictivos para ser útiles para describir los patrones de no estacionariedad que se encuentran en los procesos reales. Más aún, se pueden considerar como modelos bases. De hecho, en casos reales, parece más realista que las observaciones extremas (máximos o mínimos) cambien durante el período de estudio, pero que en otros aspectos, la distribución no cambie.

Por ejemplo, si $VEG(\mu, \sigma, \kappa)$ denota la distribución de VEG de parámetros μ, σ y κ para la variable aleatoria X, entonces un modelo mayormente viable para X_t podría ser:

$$X_t \sim VEG(\mu(t), \sigma, \kappa), \tag{1.29}$$

donde, $\mu(t) = \beta_0 + \beta_1 t$ de parámetros β_0 y β_1. Esto es, las variaciones a lo largo del tiempo en el proceso observado se modelan como una tendencia lineal en el parámetro de ubicación. Un modelo más complejo podría ser la función cuadrática,

$$\mu(t) = \beta_0 + \beta_1 t + \beta_2 t^2. \tag{1.30}$$

Más aún, los parámetros de la distribución de valores extremos pueden tomar diversas formas,

$$\theta(t) = h(X^T \beta), \tag{1.31}$$

donde $\theta(t)$ representa cualquieras de los parámetros μ, σ y κ, h es una función especifica que suele denominarse función de enlace inverso, β es el vector de parámetros y X es un vector modelo. Si bien, puede notar que existe una analogía con este modelo y los modelos lineales generalizados (MLG), la diferencia

reside en que los MLG está restringida a distribuciones que pertenecen a la familia de la distribución exponencial, mientras que el modelo de VE quedan fuera de esta familia. Además, al ser una familia bastante amplia representa de forma adecuada la no estacionariedad en conjuntos de datos de valor extremo. Por tanto, cuando se observa que los datos de extremos pueden ser no estacionarios, el modelo a estimar es de la forma,

$$X_t \sim VEG(\mu(t), \sigma(t), \kappa(t)),$$

para $t = 1, 2, \ldots, n$.

1.17. Métodos para obtener valores extremos

Actualmente los enfoques que son mayormente utilizados para la obtención de los máximos son: el método de bloque máximo y picos por encima o por debajo de un umbral, conocido también como enfoque de excedentes.

1. Método del bloque máximo
   Este enfoque fue propuesto por Gilli y Kellezi (2006), consiste en dividir el conjunto de datos en bloques del mismo tamaño (semanal, mensual, anual, etc.), luego, se selecciona el valor máximo (mínimo) de cada uno de los bloques, estos, son los valores extremos. Posteriormente a esto, se ajusta la distribución de Valores Extremos Generalizado al conjunto de máximos (mínimos).

   Formalmente, si $x_1, x_2, \ldots, x_n$ es una muestra aleatoria, esta se divide en m bloques (renglones) de tamaño k,

$$\begin{array}{cccc} x_1 & x_2 & \ldots & x_k \\ x_{k+1} & x_{k+2} & \ldots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{(m-1)k+1} & x_{(m-1)k+2} & \ldots & x_{mk} \end{array} \tag{1.32}$$

   Posteriormente, se obtiene el máximo de cada uno de los bloques que comúnmente son denotados por x_i^*, para cada $i = 1, 2, \ldots, k$, luego, se asume que $x_i^* \sim DVEG$. Finalmente, se busca la aproximación de la distribución con el nuevo conjunto de máximos.

   Por otra parte, cabe señalar que, los bloques se eligen para que correspondan a un período de tiempo de duración de una semana, un mes, un año, etc., en cuyo caso se dice que el tamaño del bloque m es el número de observaciones en una semana, un mes, un año respectivamente y los máximos de cada bloque son denominados máximos semanales, mensuales o anuales según corresponda.

2. Enfoque de excedentes
   Este método es útil cuando el tamaño de los datos es grande. Se establece un umbral, y los valores que excedan o estén por debajo de dicho umbral,

forman el conjunto de datos de valores extremos. Para realizar la elección del umbral adecuado, se realiza un análisis gráfico.

1.18. Estimación de parámetros

Es común que una vez propuesta la función paramétrica para modelar un problema en particular, es necesario el uso de métodos o algoritmos para obtener estimaciones de los parámetros del modelo. El método frecuentemente utilizado es maximizar la función verosimilitud, la cual se describe a continuación.

Función verosimilitud

Dado una realización de una sucesión de variables aleatorias, la función de verosimilitud o únicamente verosimilitud se define como la probabilidad de observar dicha muestra o realización. Formalmente se define como sigue.

Definición 1.18.1. Sean $X_1, X_2, \ldots, X_n$ una sucesión de v.a.i.i.d. con función de probabilidad $P(x;\theta)$ y $\underline{x} = \{x_1, x_2, \ldots, x_n\}$ una realización. La verosimilitud de θ se define como:

$$L(\theta;\underline{x}) = \prod_{i=1}^{n} P(x_i;\theta),$$

donde $\theta = (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta$ es el vector de parámetros a estimar y Θ es el espacio paramétrico.

Definición 1.18.2. El estimador de máxima verosimilitud (EMV) de θ es el valor del parámetro en el espacio paramétrico $\widehat{\theta} \in \Theta$ que satisface,

$$L(\underline{x};\widehat{\theta}) = \sup_{\theta \in \Theta} L(\theta;\underline{x}). \tag{1.33}$$

Nótese que $0 \leq L(\theta;\underline{x}) \leq 1$, sin embargo, puede darse el caso en que el EMV no exista, y si existe, puede no ser único. El EMV $\widehat{\theta}$, es el valor de θ que mejor explica a la muestra observada, esto es, maximiza su probabilidad bajo el modelo propuesto.

Dependiendo del modelo propuesto, en ocasiones encontrar el valor de θ que maximiza a $L(\theta;x)$ puede ser complicado debido a la forma que toma la función de verosimilitud. Por tal motivo, usualmente es conveniente y válido trabajar con la función de log-verosimilitud de θ definida como el logaritmo de $L(\theta;x)$, esto es,

$$\log(L(\theta;x)) = \sum_{i=1}^{n} \log P(x_i;\theta).$$

Cabe aclarar que, a pesar de que la función log-verosimilitud simplifica los cálculos para obtener los valores óptimos, en su gran mayoría determinarlos sigue siendo complejo, por lo cual, se recurren a métodos numéricos para estimar

dichos valores. Sin embargo, cabe resaltar que con estos algoritmos tampoco se garantiza que se determinen los óptimos globales, pues en su caso podrían tratarse sólo de óptimos locales. Es por tal motivo, la propuesta de nuevos métodos que nos permitan estimar parámetros son de vital importancia, ya que esto a su vez garantiza la calidad del modelo propuesto.

1.18.1. Verosimilitud de la DVEG no estacionario

Para estimar la verosimilitud de un modelo de VEG no estacionario de la variable Y_t,

$$Y_t \sim GEV(\mu(t), \sigma(t), \kappa(t)),$$

para $t = 1, 2, \ldots, n$, donde cada parámetro $\mu(t), \sigma(t), \kappa(t)$ tienen una expresión en términos de un vector de parámetros y covariables del tipo (1.31), denotaremos a $\mu(t), \sigma(t)$ y $\kappa(t)$ por μ_t, σ_t y κ_t respectivamente.

Considere la función de densidad $g(y)$ de la distribución VEG definida en (1.13) y en (1.14). Sea $\underline{y} = (y_1, \ldots, y_n)$ una muestra de n extremos; la probabilidad para la DVEG no estacionaria se define entonces como la densidad conjunta en función de los parámetros de la siguiente forma:

$$\begin{aligned} L(\mu_t, \sigma_t, \kappa_t \mid \underline{y}) &= \prod_{t=1}^{n} \frac{1}{\sigma_t} \left[1 + \kappa_t \left(\frac{y_t - \mu_t}{\sigma_t}\right)\right]^{-\left(1+\frac{1}{\kappa_t}\right)} \\ &\quad \times \exp\left\{-\left[1 + \kappa_t \left(\frac{y_t - \mu_t}{\sigma_t}\right)\right]^{-\frac{1}{\kappa_t}}\right\}, \end{aligned}$$

cuando $\kappa \neq 0$ y $1 + \kappa\left(\frac{y_t - \mu}{\sigma}\right) > 0$, para $t = 1, 2, \ldots, n$.

En la práctica, se estiman los parámetros usando la función de logaritmo de la verosimilitud, debido a que la probabilidad y el logaritmo de verosimilitud tienen los mismos puntos críticos y es una función numéricamente más simple de optimizar. La función logaritmo de verosimilitud para la DVEG no estacionario es la siguiente:

$$\begin{aligned} \ell(\mu_t, \sigma_t, \kappa_t \mid \underline{y}) &= -n \log \sigma_t - \sum_{t=1}^{n} \left[1 + \kappa_t \left(\frac{y_t - \mu_t}{\sigma_t}\right)\right]^{-\frac{1}{\kappa_t}} \\ &\quad - \sum_{t=1}^{n} \left(1 + \frac{1}{\kappa_t}\right) \log\left[1 + \kappa_t \left(\frac{y_t - \mu_t}{\sigma_t}\right)\right]. \end{aligned} \tag{1.34}$$

Para simplificar la notación definimos,

$$\begin{aligned} \ell_t(\mu_t, \sigma_t, \kappa_t \mid \underline{y}) &= -\log \sigma_t - \left[1 + \kappa_t \left(\frac{y_t - \mu_t}{\sigma_t}\right)\right]^{\frac{1}{\kappa_t}} \\ &\quad - \left(1 + \frac{1}{\kappa_t}\right) \log\left[1 + \kappa_t \left(\frac{y_t - \mu_t}{\sigma_t}\right)\right], \end{aligned}$$

y $\ell_n(\mu_t, \sigma_t, \kappa_t \mid \underline{y}) = \sum_{t=1}^{n} \ell_t(\mu_t, \sigma_t, \kappa_t \mid \underline{y})$, y en consecuencia, podemos reescribir la ecuación (1.34) como:

$$\ell_n(\mu_t, \sigma_t, \kappa_t \mid \underline{y}) = \sum_{t=1}^{n} \ell_t(\mu_t, \sigma_t, \kappa_t \mid \underline{y}).$$

Por tanto, el gradiente de la verosimilitud viene dado por:

$$\frac{\partial \ell_n(\mu_t, \sigma_t, \kappa_t \mid \underline{y})}{\partial \mu_t} = \sum_{t=1}^{n} \frac{\ell_t(\mu_t, \sigma_t, \kappa_t \mid \underline{y})}{\partial \mu_t}, \tag{1.35}$$

$$\frac{\partial \ell_n(\mu_t, \sigma_t, \kappa_t \mid \underline{y})}{\partial \log \sigma_t} = \sum_{t=1}^{n} \frac{\ell_t(\mu_t, \sigma_t, \kappa_t \mid \underline{y})}{\partial \log \sigma_t}, \tag{1.36}$$

$$\frac{\partial \ell_n(\mu_t, \sigma_t, \kappa_t \mid \underline{y})}{\partial \kappa_t} = \sum_{t=1}^{n} \frac{\ell_t(\mu_t, \sigma_t, \kappa_t \mid \underline{y})}{\partial \kappa_t}, \tag{1.37}$$

donde,

$$\frac{\ell_t(\mu_t, \sigma_t, \kappa_t \mid \underline{y})}{\partial \mu_t} = -\frac{1}{\sigma_t \left[1 + \kappa_t \left(\frac{y-\mu_t}{\sigma_t}\right)\right]^{\left(\frac{1}{\kappa_t}+1\right)}} + \frac{\kappa_t \left(1 + \frac{1}{\kappa_t}\right)}{\sigma_t \left(1 + \kappa_t \left(\frac{y-\mu_t}{\sigma_t}\right)\right)},$$

$$\frac{\ell_t(\mu_t, \sigma_t, \kappa_t \mid \underline{y})}{\partial \log \sigma_t} = -1 - \frac{\left(\frac{y-\mu_t}{\sigma_t}\right)}{\left[1 + \kappa_t \left(\frac{y-\mu_t}{\sigma_t}\right)\right]^{\left(\frac{1}{\kappa_t}+1\right)}} + \frac{\left(1 + \frac{1}{\kappa_t}\right)\left(\kappa_t \left(\frac{y-\mu_t}{\sigma_t}\right)\right)}{\left[1 + \kappa_t \left(\frac{y-\mu_t}{\sigma_t}\right)\right]},$$

$$\begin{aligned}\frac{\ell_t(\mu_t, \sigma_t, \kappa_t \mid \underline{y})}{\partial \kappa_t} = & \frac{-\left(\frac{1}{\kappa_t}+1\right)\left(\frac{y-\mu_t}{\sigma_t}\right)}{1 + \kappa_t \left(\frac{y-\mu_t}{\sigma_t}\right)} + \frac{\log\left(1 + \kappa_t \left(\frac{y-\mu_t}{\sigma_t}\right)\right)}{\kappa_t^2} \\ & + \frac{\left(\frac{y-\mu_t}{\sigma_t}\right)}{\kappa_t \left[1 + \left(\kappa_t \left(\frac{y-\mu_t}{\sigma_t}\right)\right)\right]^{\left(\frac{1}{\kappa_t}+1\right)}} + \frac{\log\left(1 + \left(\kappa_t \left(\frac{y-\mu_t}{\sigma_t}\right)\right)\right)}{\kappa_t^2 \left[1 + \left(\kappa_t \left(\frac{y-\mu_t}{\sigma_t}\right)\right)\right]^{\frac{1}{\kappa_t}}}.\end{aligned}$$

Para el caso $\kappa = 0$, el logaritmo de la verosimilitud es:

$$\begin{aligned}\ell(\mu_t, \sigma_t, \kappa_t \mid \underline{y}) &= \sum_{t=1}^{n} \log\left(\frac{1}{\sigma_t}\right) + \sum_{t=1}^{n} \left[-\exp\left(\frac{\mu_t - y_t}{\sigma_t}\right) + \frac{\mu_t - y_t}{\sigma_t}\right] \\ &= -\sum_{t=1}^{n} \log(\sigma_t) - \sum_{t=1}^{n} \exp\left(\frac{\mu_t - y_t}{\sigma_t}\right) + \sum_{t=1}^{n} \frac{\mu_t - y_t}{\sigma_t}\end{aligned}$$

de donde,

$$\begin{aligned}\frac{\ell(\mu_t,\sigma_t,\kappa_t \mid \underline{y})}{\partial\mu_t} &= -\frac{1}{\sigma_t}\exp\left(\frac{\mu_t-y_t}{\sigma_t}\right)+\frac{1}{\sigma_t}\\ \frac{\ell(\mu_t,\sigma_t,\kappa_t \mid \underline{y})}{\partial\sigma_t} &= -\frac{1}{\sigma_t}+\left(\frac{\mu_t-y_t}{\sigma_t^2}\right)\exp\left(\frac{\mu_t-y_t}{\sigma_t}\right)-\left(\frac{\mu_t-y_t}{\sigma_t^2}\right).\end{aligned}$$

Como puede ver, resolver los sistemas obtenidos al igualar a cero las derivadas parciales para obtener los valores óptimos de los parámetros es bastante complejo, por lo cual, en la actualidad la forma usual de determinar los parámetros es mediante el uso de métodos numéricos. Más aún, el estudio esta mayormente enfocado a un modelo de DVEG estacionario como puede ver en [8].

1.19. Bondad de ajuste del modelo

El proceso de selección del mejor modelo que represente a los datos observados es el proceso de validación o diagnóstico de bondad de ajuste del modelo propuesto. Esto se realiza con la finalidad de determinar si se obtuvo una respuesta adecuada por parte del modelo. Existen actualmente diversos métodos gráficos y analíticos para tal propósito. Sin embargo, cabe resaltar que en teoría de valores extremos es común el uso de métodos gráficos, ya que los datos en este caso son bidimensionales. Así, para el objetivo de esta tesis, se describen algunos de los métodos comúnmente utilizados.

1.19.1. Histogramas

Una técnica muy útil para verificar la elección del modelo es mediante el uso de histogramas. La idea consiste en superponer en la misma gráfica el histograma de los datos y la función de densidad estimada del modelo que representa al "mejor candidato". Si el histograma y la función estimada son similares, entonces podemos decir que el modelo es adecuado. Sin embargo, como puede notarse, este enfoque es bastante subjetivo a la hora de ver que tan "bueno" es el modelo propuesto, además, considerar esta práctica es bastante restrictiva, ya que es más común que los datos sean multidimensionales.

1.19.2. Gráfica PP

Otra gráfica ampliamente utilizada para verificar la bondad del ajuste del modelo propuesto es la gráfica de probabilidad o gráfica PP, la cual se define como sigue.

Considere una sucesión de variables aleatorias i.i.d., $X_1, X_2, \ldots, X_n$, con función de distribución $F(x)$. Se definen las variables aleatorias:

$$\begin{aligned}
X_{(1)}(\omega) &= \text{mín}\{X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\}, \\
X_{(2)}(\omega) &= \text{mín}\{X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\} - \{X_{(1)}(\omega)\}, \\
X_{(3)}(\omega) &= \text{mín}\{X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\} - \{X_{(1)}(\omega), X_{(2)}(\omega)\}, \\
&\vdots \\
X_{(n)}(\omega) &= \text{máx}\{X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\}.
\end{aligned}$$

Al conjunto $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ se le denomina estadísticas de orden de las variables $X_1, X_2, \ldots, X_n$. Bajo esta definición se satisface:

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)},$$

de donde, a diferencia de las v.a.'s X_i estas no son independientes, ni tampoco son idénticamente distribuidas.

Definición 1.19.1. Sean $X_1, X_2, \ldots, X_n$ una sucesión de variables aleatorias i.i.d. con función de distribución $F(x)$ y $x_1, x_2, \ldots, x_n$ una realización de la sucesión. Si $X_{(1)} \leq X_{(2)} \leq, \cdots, \leq X_{(n)}$ son las estadísticas de orden, la función de distribución empírica se define como,

$$\widetilde{F}(x) = \frac{i}{n+1} \quad \text{para} \quad x_{(i)} \leq x \leq x_{(i+1)}.$$

Considere además que $\widehat{F}(x)$ es un estimador para $F(x)$, así, dado que $\widetilde{F}(x)$ es también un estimador para la función de distribución verdadera $F(x)$, debe satisfacerse entonces que, $\widehat{F}(x)$ y $\widetilde{F}(x)$ son similares.

Definición 1.19.2. Sean $X_1, X_2, \ldots, X_n$ una sucesión de variables aleatorias i.i.d. con función de distribución $F(x)$ y $x_1, x_2, \ldots, x_n$ una realización de la sucesión. Si $X_{(1)} \leq X_{(2)} \leq, \cdots, \leq X_{(n)}$ son las estadísticas de orden y $\widehat{F}(x)$ es una función de distribución estimada para $F(x)$, la gráfica probabilidad-probabilidad o gráfica PP consiste en el siguiente conjunto de puntos:

$$\left\{ \left(\widehat{F}(x_{(i)}), \frac{i}{n+1} \right) : i = 1, 2, \ldots, n \right\}.$$

De esta forma, si el estimador $\widehat{F}(x)$ de $F(x)$ es considerado como un modelo razonable para representar a los datos, entonces, los puntos de la gráfica PP deben estar cerca de la recta identidad, de tal forma que, mientras más se aleje de ella, se concluiría que el modelo no es un buen estimador.

1.19.3. Gráfica cuantil-cuantil

Una gráfica cuantil-cuantil es otro método gráfico de diagnóstico que permite verificar si un modelo es adecuado para la muestra observada, esta se define a continuación.

Definición 1.19.3. Sean $X_1, X_2, \ldots, X_n$ una sucesión de variables aleatorias i.i.d. con función de distribución $F(x)$ y $x_1, x_2, \ldots, x_n$ una realización de la sucesión. Si $X_{(1)} \leq X_{(2)} \leq, \cdots, \leq X_{(n)}$ son las estadísticas de orden y $\widehat{F}(x)$ es una función de distribución estimada para $F(x)$, la gráfica cuantil, gráfica cuantil-cuantil o gráfica QQ consiste en el siguiente conjunto de puntos,

$$\left\{ \left(\widehat{F}^{-1}\left(\frac{i}{n+1} \right), x_{(i)} \right) : i = 1, 2, \ldots, n \right\}.$$

De manera análoga a la gráfica PP, si el estimador $\widehat{F}(x)$ es un buen estimador de $F(x)$, entonces los puntos deberán estar muy cerca de la recta identidad.

Observación 1.20. Si $x_{(1)} \leq x_{(2)} \leq, \cdots, \leq x_{(n)}$ son las estadísticas de orden de los máximos de los bloques, la función de distribución empírica en $x_{(i)}$ es,

$$\widetilde{G}(x) = \frac{i}{n+1} \quad \text{para} \quad x_{(i)} \leq x \leq x_{(i+1)}.$$

La función estimada,

$$\widetilde{G}(x) = \exp\left\{ -\left(1 + \widehat{\kappa}\frac{(x-\widehat{\mu})}{\widehat{\sigma}} \right)^{-\frac{1}{\widehat{\kappa}}} \right\},\ \widehat{\kappa} \neq 0;\ 1 + \widehat{\kappa}\frac{(x-\widehat{\mu})}{\widehat{\sigma}} > 0,$$

y

$$\widehat{G}^{-1}\left(\frac{i}{n+1} \right) = \widehat{\mu} - \frac{\widehat{\sigma}}{\widehat{\kappa}}\left[1 - \left\{ -\log\left(\frac{i}{n+1} \right) \right\}^{-\widehat{\kappa}} \right].$$

A pesar de que los métodos gráficos son ampliamente utilizados, son un tanto subjetivos a la hora de comparar los modelos propuestos, por tal motivo, los criterios basados en un análisis numérico también se vuelven indispensables. A continuación se describe brevemente el criterio de información de Akaike, el cual en estadística es probablemente uno de los métodos más utilizados.

1.20.1. Gráfica de niveles de retorno

En la teoría de valores extremos es de interés conocer cuál es el valor que en promedio se excede una vez cada determinado tiempo, por esta razón, se describen y enuncian conceptos relacionados a tal propósito, como son: función cuantil, nivel de retorno y período de retorno.

En probabilidad y estadística, la función cuantil, es conocida como inversa de la función de distribución. Esta función esta asociada con la función de distribución de una variable aleatoria, y representa el valor que toma la variable aleatoria para el cual la probabilidad de que esa variable aleatoria sea menor o igual a dicho valor sea la probabilidad dada. Formalmente, se define a continuación.

Definición 1.20.1. Sea X una variable aleatoria con función de distribución $F(x)$ tal que, es estrictamente monótona y continua en el intervalo $(0,1)$. La función cuantil se denota por $Q(p)$ (o bien Q_p) y se define como la función inversa de $F(x)$ en este intervalo, esto es,

$$Q(p) = F^{-1}(p) = \text{ínf}\{x \in \mathbb{R} : F(x) \geq p\},$$

para cada $p \in (0,1)$. Al valor $x_p = Q(p)$ se le denomina cuantil p de F. Sin embargo, dependiendo de la literatura al valor x_p también se le puede denominar cuantil de orden p, p-cuantil o cuantil de probabilidad acumulada p.

Los cuantiles Q_p para la distribución de VEG, descrito en la ecuación (1.13.1) esta dada por:

$$x_p = \begin{cases} \mu - \frac{\sigma}{\kappa}[1 - \{-ln(p)\}^{-\kappa}], & \text{para } \kappa \neq 0, \\ \\ \mu - \sigma ln\{-ln(p)\}, & \text{para } \kappa = 0, \end{cases} \tag{1.38}$$

los cuales se obtiene despejando Q_p de la siguiente igualdad, para $\kappa = 0$,

$$\begin{aligned} G(Q_p) &= p \\ \exp\left\{-\exp\left[\frac{-(Q_p - \mu)}{\sigma}\right]\right\} &= p \end{aligned}$$

y para $\kappa \neq 0$,

$$\exp\left\{-\left(1 + \kappa\frac{(x-\mu)}{\sigma}\right)^{-\frac{1}{\kappa}}\right\} = p.$$

De forma análoga, para el cuantil $Q_{1-p} = z_p$ con $p \in (0,1)$, que es conocido también como cuantil por exceso, resolvemos $G(z_p) = 1 - p$, obteniendo entonces,

$$z_p = \begin{cases} \mu - \frac{\sigma}{\kappa}[1 - \{-ln(1-p)\}^{-\kappa}], & \text{para } \kappa \neq 0, \\ \\ \mu - \sigma ln\{-ln(1-p)\}, & \text{para } \kappa = 0. \end{cases} \tag{1.39}$$

El período de retorno denotado comúnmente por T, es el intervalo de tiempo promedio (en años) entre la ocurrencia de un evento de una magnitud igual o mayor que un cierto valor para un año cualquiera, esto es, denota un intervalo de recurrencia [26].

Al valor z_p se le conoce como nivel de retorno, y esta asociado al período de retorno $T = 1/p$, según Stuart Coles en su libro [8] se espera que el nivel z_p se exceda en promedio una vez cada $1/p$ unidades de tiempo con cierta precisión. Es decir, z_p va a resultar excedido por el máximo de la unidad de

tiempo en cualquier unidad de tiempo particular con probabilidad p, véase [33].

Por otra parte, haciendo $y_p = -ln(1-p)$, la gráfica de niveles consiste en graficar z_p contra $ln(y_p)$. De modo que, al considerar a y_p como variable puede verse de la ecuación (1.39) que, la función es de tipo lineal en el caso en el que $\kappa = 0$, mientras que si $\kappa < 0$ la función es convexa, y para el caso $\kappa > 0$ la función es cóncava.

1.20.2. Criterio de Información de Akaike

El criterio de información de Akaike permite cuantificar la idoneidad de un modelo particular en relación con un conjunto finito de modelos candidatos, penalizando la inclusión de parámetros. Este criterio, que se enmarca en el campo de la teoría de la información, por lo cual, su fundamento teórico esta fuera del alcance de esta tesis.

El índice denotado por AIC se define como:

$$AIC = -2\ \log(L(\widehat{\theta})) + 2K, \tag{1.40}$$

donde $\log(L(\widehat{\theta}))$ es el logaritmo de la máxima verosimilitud, y K es el número de parámetros libres o parámetros estimados en el modelo.

El primer término de esta ecuación representa una medida de bondad (calidad) con la que el modelo se ajusta a los datos observados, mientras que el segundo término, es una penalización que se incrementa con la complejidad del modelo, esto es, mientras el modelo tenga un mayor número de parámetros, la penalización es mucho mayor. Por ejemplo, si a partir de un conjunto de datos, se calculan los valores AIC para dos modelos distintos, entonces el menor valor de AIC indica que, o bien el modelo se ajusta mejor a los datos o que es menos complejo en relación al otro modelo, y en realidad una combinación de ambos factores. Por lo tanto, este criterio ofrece un valor objetivo que de manera relativa, cuantifica de forma simultánea la precisión y complejidad del modelo.

El criterio para la selección del modelo es seleccionar el modelo que obtenga un menor valor AIC de entre un conjunto de modelos candidatos. En cierto modo con este método, no se busca el modelo que mejor ajusta, sino el modelo que menos información pierde de todos los modelos considerados en estudio. Sin embargo, cabe señalar que el método de selección por medio del valor AIC debe ser empleado cuidadosamente, esto debido a que existe una clara dependencia con la calidad del conjunto de modelos candidatos, puesto que el modelo que presente un menor valor AIC puede no ser un buen modelo en un sentido absoluto si el conjunto de modelos analizados no contiene buenos modelos. Por tanto, la selección del conjunto de modelos propuestos también juega un papel importante para el buen uso de este método de selección.

Capítulo 2

Árboles de decisión

Los árboles de decisión son algoritmos estadísticos (métodos no paramétricos) o del aprendizaje automático que permiten la construcción de modelos predictivos. Estos árboles se pueden clasificar en árboles de regresión y de clasificación (CART, por sus siglas en inglés, Classification and Regression Trees), los cuales fueron implementados por Leo Breiman, Jerome Friedman, Richard Olshen y Charles Stone, en el libro [4] en 1984. El tipo de árbol dependerá de la variable a predecir, esto es, si la variable respuesta es continua, entonces el problema es de regresión, en cambio, si la variable respuesta es categórica, el problema es de clasificación.

En la actualidad estos algoritmos tienen una gran importancia debido a las múltiples ventajas que ofrecen, algunos de estos son: su robustez a los valores atípicos, su fácil interpretabilidad y la variabilidad en el tipo de dato. Por tal motivo, dichos modelos suelen ser la base de otros algoritmos del aprendizaje automático como verá en la sección 2.2. Además, cabe resaltar que, aunque los árboles de regresión y los árboles de clasificación tienen algunas similitudes, también tienen diferencias, principalmente en la metodología para determinar como generar las divisiones de las ramas del árbol. Por otra parte, para el objetivo de esta tesis el interés se centra en el estudio de los árboles de regresión. Por lo cual, de aquí en adelante nos enfocaremos en este tipo de árbol.

A continuación puede ver algunas de sus principales características y elementos fundamentales, vea Figura (2.1).

1. El nodo raíz es el primer nodo del árbol.

2. Un nodo Y es descendiente directo de un nodo X, si el nodo X apunta al nodo Y, comúnmente se suele decir que Y es hijo de X.

3. Cualquier nodo que tiene descendientes y no es el nodo raíz, es nodo interno del árbol.

4. Los nodos que no tienen descendientes, se les conoce como nodos hojas.

5. El grado de un nodo, es el número de hijos del nodo. Mientras que, el grado de un árbol, es el grado máximo de cada uno de sus nodos.

6. El nivel del árbol, es el número enlaces por recorrer para llegar de un nodo a otro.

7. La altura o profundidad del árbol es el máximo nivel de cada uno de los nodos del árbol.

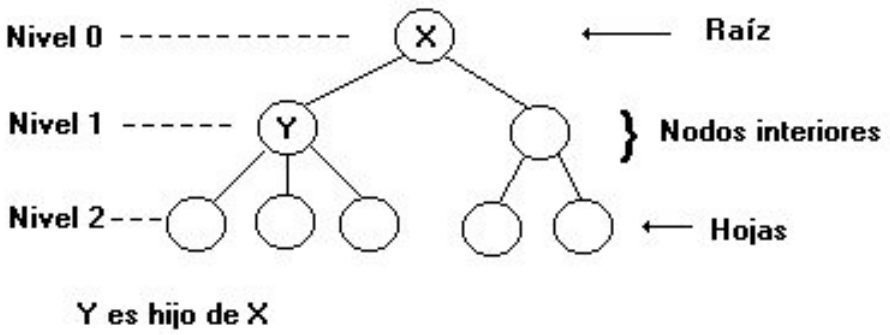


Figura 2.1: Estructura de los árboles de regresión y clasificación.

En la práctica es común que el grado de un nodo y la profundidad del árbol sean hiperparámetros por ajustar para entrenar el modelo de forma adecuada. De echo, la calidad del modelo también puede depender en gran medida de los valores de estos hiperparámetros.

La idea primordial de los árboles de regresión reside en tratar de particionar el espacio de las covariables en hiper-rectángulos, por medio de cada una de las covariables en estudio [3]. Dicha partición sobre el espacio de las covariables, se hace de manera repetitiva para cada una con el objetivo de determinar los puntos de corte óptimos, de tal manera que se minimice una función de costos. Finalmente, todas las observaciones que queden dentro de un hiper-rectángulo se les asigna el mismo valor estimado.

2.1. Método de entrenamiento de los árboles de regresión

Los procedimientos descritos en esta sección se obtuvieron principalmente de [3].

La metodología para la construcción de los arboles de regresión consta principalmente de 2 etapas. Sin embargo, cabe señalar que en la actualidad existen diversas variaciones del método en sus distintas etapas, pero en muchos casos la idea inicial del particionamiento recursivo es la misma.

La primera etapa, es la división recursiva (sucesiva) del espacio de las covariables, con el objetivo de generar regiones $R_1, R_2, R_3, \ldots, R_J$ que no se solapen

entre sí. Es usual que las regiones se establezcan como rectangulares ya que esto simplifica el proceso de construcción y facilita su interpretación. La segunda etapa es la predicción de la variable respuesta en cada una de las regiones.

Formalmente, el objetivo en regresión es encontrar J regiones $R_1, R_2, R_3, \ldots, R_J$ que minimicen la varianza sobre cada región:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_J})^2,$$

donde $\hat{y}_{R_j}$ representa la media de la variable respuesta en la región R_j. Por tanto, como puede observar, se trata de determinar la distribución sobre el espacio de las covariables de tal forma que la suma al cuadrado de la diferencia entre las observaciones y la media sobre la región a la que pertenecen sea mínima.

Sin embargo, dado que no es posible evaluar sobre todas las particiones del espacio debido al alto costo computacional, se recurre a lo que se conoce como división binaria recursiva o partición binaria recursiva, método que se describe a continuación.

División binaria recursiva

La división binaria recursiva es un algoritmo cuyo principal objetivo es encontrar en cada iteración el predictor X_i, y el punto de corte (umbral) c_i tal que si se distribuyen las observaciones sobre las regiones $R_1(i, c_i) = \{X | X_i < c_i\}$ y $R_2(i, c_i) = \{X | X_i \geq c_i\}$, entonces se consigue la mayor reducción posible en la función de costo:

$$\sum_{i \in R_1} (y_i - \hat{y}_{R_j})^2 + \sum_{i \in R_2} (y_i - \hat{y}_{R_j})^2.$$

El algoritmo consiste en los pasos siguientes:

1. Inicialmente, todas las observaciones se consideran dentro de una misma región.

2. Se calculan los puntos de corte c_j, $j = 1, \ldots, p$ para cada una de las covariables $X_1, X_2, \ldots, X_p$. En el caso de covariables discretas, el punto de corte es algunas de ellas, mientras que si la variable es continua, el punto de corte es el punto medio de los valores que puede tomar cada una de las covariables X_j, esto es:

$$c_j = \frac{\text{máx}\{X_j\} - \text{mín}\{X_j\}}{2}, \quad j = 1, \ldots, p.$$

3. Se calcula la suma de la función de costo sobre las 2 regiones generadas para cada una de las covariables,

$$\sum_{i: x_i \in R_1(j, c_j)} (y_i - \hat{y}_{R_1})^2 + \sum_{i: x_i \in R_2(j, c_j)} (y_i - \hat{y}_{R_2})^2, \quad j = 1, \ldots, p. \tag{2.1}$$

4. Se selecciona la covariable X_j con su respectivo punto de corte c_j, que genere el menor costo total. En el caso de que existan dos o más divisiones que generan el mismo costo, la selección de covariable es aleatoria.

5. Los pasos $1, 2, 3$ y 4 son repetidos, en cada una de las regiones que se van generando en el proceso anterior hasta alcanzar la condición de parada. Existen distintas condiciones de parada, sin embargo, las más comunes son; que ninguna región contenga un número mínimo de observaciones, que el árbol tenga un número máximo de nodos terminales, o que el costo se reduzca de forma significativa en la incorporación de un nuevo nodo, entre otros.

2.2. Estimación de funciones

Los temas descritos en esta sección se obtuvieron mayormente de los artículos: [12] en el tema de optimización, y [6] en el estudio de XGBoost.

En problemas de regresión o clasificación, en donde el conjunto de observaciones está dado por: $\{(\boldsymbol{x}_i, y_i)\}$, con $\boldsymbol{x_i} \in \mathbb{R}^p$ las variables independientes e y_i la variable respuesta, el objetivo principal es proporcionar un estimador $\widehat{F}(\boldsymbol{x})$ de $F^*(\boldsymbol{x})$, donde F^* es el valor óptimo, que reduce la pérdida esperada entre las observaciones predichas y observadas, de modo que está definida en términos de una función de pérdida como sigue:

$$F^*(\boldsymbol{x}) = \arg\min_F E_{y,\boldsymbol{x}}\, L(y, F(\boldsymbol{x})) = \arg\min_F E_{\boldsymbol{x}}\, [E_y(L(y, F(\boldsymbol{x}))) \mid \boldsymbol{x}],$$

el término $L(F(\boldsymbol{x}), y)$ representa la función de pérdida sobre la función de distribución conjunta de los datos $(\boldsymbol{x}, y)$. Generalmente, la función de pérdida que se considera es el error cuadrático medio $\mid y - F \mid^2$ o el error absoluto $\mid y - F \mid$ en problemas de regresión, mientras que en problemas de clasificación la función usual es $\log(1 + 2\exp(-2yF))$. Para atender este problema es usual considerar a F como una función paramétrica $F(\boldsymbol{x}; \boldsymbol{P})$, con $\boldsymbol{P} = \{P_1, P_2, \ldots, P_k\}$ el conjunto finito de parámetros. De este modo, el problema en sí implica determinar los parámetros óptimos $\boldsymbol{P}$ de la función F, los cuales usualmente son determinados mediante algún método de optimización numérica.

En el presente estudio, se considera que F es de forma aditiva,

$$F(\boldsymbol{x}; \{\beta_m, \boldsymbol{a_m}\}) = \sum_{m=1}^{M} \beta_m h(\boldsymbol{x}; \boldsymbol{a_m}), \tag{2.2}$$

en donde la expresión $h(\boldsymbol{x}; \boldsymbol{a})$ es una función paramétrica, con $\{\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots\}$ los parámetros a estimar. En el presente estudio, se considera a $h(\cdot, \cdot)$ como árboles de regresión, por lo cual, los parámetros $\boldsymbol{a_m}$ representan características propias de los mismos, esto es, las ubicaciones de las divisiones de las ramas, las medias de los nodos terminales de los árboles individuales, etc.

Entonces, considerando que el modelo es paramétrico, el objetivo se reduce a la búsqueda del parámetro $\boldsymbol{P^*}$ tal que se reduzca una función de costo:

$$\boldsymbol{P^*} = arg\min_{\boldsymbol{P}} \ \phi(\boldsymbol{P}),$$

donde

$$\phi(\boldsymbol{P}) = E_{y,x}\, L(y, F(\boldsymbol{x}; \boldsymbol{P})), \tag{2.3}$$

de esta forma se obtendría $F^*(\boldsymbol{x}) = F(\boldsymbol{x}; \boldsymbol{P}^*)$. Así, por medio de optimización numérica se tiene una solución de la forma,

$$\boldsymbol{P}^* = \sum_{m=0}^{M} \boldsymbol{p}_m, \tag{2.4}$$

en el que $\boldsymbol{p}_0$ es la conjetura inicial y los $\boldsymbol{p}_i$, $i = 1, 2, \ldots, M$ son generados de forma consecutiva por medio de los valores anteriores. Para resolver este problema de optimización (2.4), el método mayormente utilizado es el método de descenso de gradiente o descenso más pronunciado, el cual se describe a continuación.

2.2.1. Descenso por gradiente

El método de descenso de gradiente o descenso más pronunciado (o método de Augustin Louis Cauchy 1847), es uno de los algoritmos de optimización mayormente utilizados para determinar mínimos (máximos) locales de una función diferenciable. La idea del método para determinar un mínimo (máximo) radica en dar pasos consecutivos en dirección opuesta al gradiente de la función en el punto actual, ya que esta es la dirección del descenso más pronunciado, ver Figura (2.2). Por el contrario, para determinar máximos locales de la función, se avanza en la dirección del gradiente y en tal caso el procedimiento se conoce como ascenso por gradiente.

La metodología para obtener el valor óptimo mínimo se define de la siguiente forma. Considere la función definida en (2.3). Inicialmente, se calcula el gradiente $\boldsymbol{g_m}$,

$$\boldsymbol{g_m} = \left[\frac{\partial \phi(\boldsymbol{P})}{\partial P_1}, \frac{\partial \phi(\boldsymbol{P})}{\partial P_2}, \ldots, \frac{\partial \phi(\boldsymbol{P})}{\partial P_k} \right]_{\boldsymbol{P}=\boldsymbol{P}_{m-1}},$$

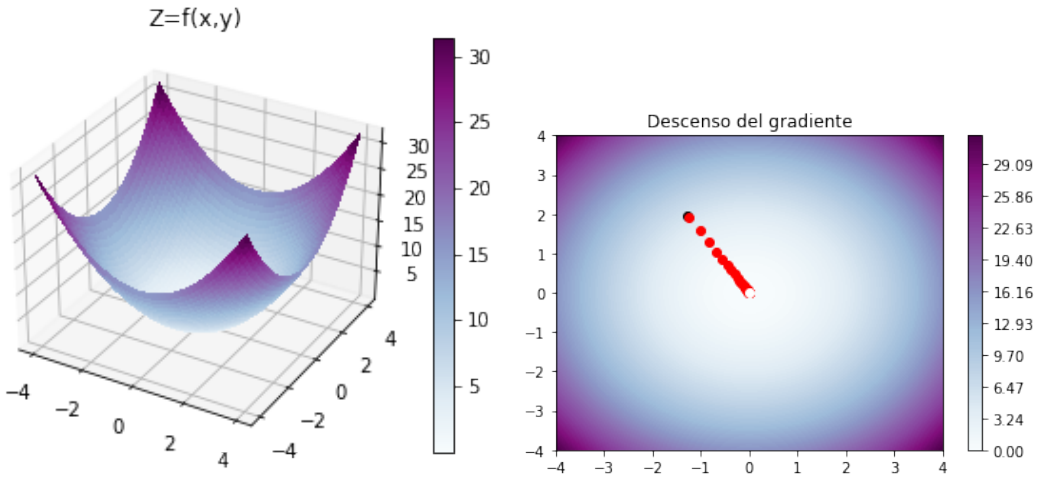
donde $\boldsymbol{P}_{m-1} = \sum_{i=0}^{m-1} \boldsymbol{P}_i$, $m = 1, \ldots, M$, y $\boldsymbol{P_0}$ la condición inicial. A los $\boldsymbol{P_i}$ se les denomina incrementos, pasos o impulsos, y cada paso $\boldsymbol{P_m}$ se calcula de la siguiente forma,

$$\boldsymbol{P}_m = -\rho_m \boldsymbol{g_m},$$

en el que,

$$\rho_m = arg\min_{\rho} \ \phi(\boldsymbol{P}_{m-1} - \rho \boldsymbol{g_m}). \tag{2.5}$$

Se dice que $-\boldsymbol{g_m}$ define la dirección del gradiente más pronunciado, y (2.5) se denomina búsqueda de línea a lo largo de esa dirección. Cabe resaltar que, el método no garantiza localizar el mínimo o máximo global, de hecho puede darse el caso en el que se encuentre sólo una solución local.



(a) (b)

Figura 2.2: (a) Función de costo a optimizar, (b) Sucesión de pasos para llegar al mínimo, el punto negro es la conjetura inicial p_0, el punto blanco es el punto que minimiza la función de costo.

2.2.2. Optimización numérica en el espacio de funciones

Un enfoque alternativo para estimar a $F^*(x)$ que se platea en [12] es considerar un modelo no paramétrico, esto es, realizar la optimización numérica en el espacio de funciones, por lo cual, para cada punto x, se considera a $F(x)$ como el parámetro a optimizar. Por tanto, se busca minimizar:

$$\phi(F) = E_{y,\boldsymbol{x}} L(y, F(\boldsymbol{x})) = E_{\boldsymbol{x}}[E_y(L(y, F(\boldsymbol{x}))) \mid \boldsymbol{x}],$$

equivalentemente,

$$\phi(F(\boldsymbol{x})) = E_y[L(y, F(\boldsymbol{x})) \mid \boldsymbol{x}]. \tag{2.6}$$

Para optimizar la ecuación anterior (2.6) por el método del gradiente, donde la solución es de la forma,

$$F^*(x) = \sum_{i=0}^{M} f_i(x),$$

con f_0 la estimación inicial, y las f_m son los pasos o impulsos, definidos el método del gradiente. Luego, se calcula el gradiente de la función $\phi(F(x))$,

$$g_m(\boldsymbol{x}) = \left(\frac{\partial \phi(F(\boldsymbol{x}))}{\partial F(\boldsymbol{x})} \right)_{F(\boldsymbol{x})=F_{m-1}(\boldsymbol{x})},$$

donde

$$F_{m-1}(x) = \sum_{i=0}^{m-1} f_i(x),$$

y los pasos,

$$f_m(\boldsymbol{x}) = -\rho_m g_m(\boldsymbol{x}).$$

En el caso del gradiente, nótese que bajo condiciones de regularidad sobre la función, es posible realizar la siguiente operación:

$$\begin{aligned} g_m(x) &= \left(\frac{\partial E_y[L(y, F(\boldsymbol{x})) \mid \boldsymbol{x}]}{\partial F(x)} \right)_{F(x)=F_{m-1}} \\ &= E_y \left(\frac{\partial L(y, F(x))}{\partial F(x)} \mid x \right)_{F(x)=F_{m-1}(x)}, \end{aligned} \tag{2.7}$$

el término ρ_m se obtiene de la siguiente forma,

$$\rho_m = arg \min_{\rho} E_{y,x} L(y, F_{m-1}(x) - \rho g_m(x)).$$

Cabe resaltar que, el modelo no paramétrico falla cuando se trata de estimar la función de distribución conjunta de los datos mediante un conjunto finito de datos, pues en este caso, no es posible estimar $E_y[\cdot|\boldsymbol{x}]$. Así, para la estimación basada en los datos se asume la forma paramétrica dado en (2.2) y se optimizan los parámetros con el método del gradiente. Por tanto, asumiendo a F como en (2.2), la expresión

$$\{\beta_m, \boldsymbol{a_m}\}_1^M = arg \min_{\{\beta'_m, \boldsymbol{a'_m}\}} \sum_{i=1}^{N} L(y_i, \sum_{m=1}^{M} \beta'_m h(\boldsymbol{x}_i; \boldsymbol{a'_m})), \tag{2.8}$$

para $m = 1, \ldots, M$. Puede reescribirse como,

$$\{\beta_m, \boldsymbol{a_m}\} = arg \min_{\{\beta, \boldsymbol{a}\}} \sum_{i=1}^{N} L(y_i, F_{m-1}(\boldsymbol{x}_i) + \beta h(\boldsymbol{x}_i; \boldsymbol{a})),$$

ya que,

$$F_m(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x}) + \beta_m h(\boldsymbol{x}; \boldsymbol{a}_m). \tag{2.9}$$

De aquí que, la expresión $h(\boldsymbol{x}; \boldsymbol{a})$ sea denominada función base, aprendiz base o aprendiz débil. Considerando cualquier estimador F_{m-1}, se dice que $\beta_m h(x; a_m)$ es el mejor paso hacia la estimación basado en datos de F^*.

Por otra parte, considere el análogo del gradiente negativo basado en los datos:

$$-g_m(\boldsymbol{x}_i) = - \left(\frac{\partial L(y_i; F(\boldsymbol{x}_i))}{\partial F(\boldsymbol{x}_i)} \right)_{F(x)=F_{m-1}(x)},$$

donde $\boldsymbol{g_m} = \{-g_m(x_i)\}_1^N$ proporciona la mejor dirección en el espacio de datos N- dimensional de $F_{m-1}(x)$. Sin embargo, nótese que el gradiente está definido sólo en el conjunto de los datos observados, por lo cual, esto no se puede extender a cualquier punto x. Por tanto, para poder generalizar a otros puntos se debe considerar a h como una función paramétrica que genera $\boldsymbol{h_m} = \{-h(\boldsymbol{x}_i; \boldsymbol{a}_m)\}_1^N$ más paralelo a $-\boldsymbol{g_m} \in \mathbb{R}^N$. Esto es, elegir el $h(\boldsymbol{x}; \boldsymbol{a})$ más correlacionado con $-g_m(x)$ sobre la distribución de los datos. Esto se puede obtener de la solución,

$$\boldsymbol{a_m} = arg \min_{\{\beta, \boldsymbol{a}\}} \sum_{i=1}^{N} [-g_m(\boldsymbol{x_i}) - \beta h(x_i; \boldsymbol{a})]^2. \tag{2.10}$$

Este gradiente negativo es utilizado en vez de (2.7) en el método de gradiente. De este modo, la búsqueda de línea se realiza:

$$\rho_m = arg \min_{\rho} \sum_{i=1}^{N} L(y_i, F_{m-1}(\boldsymbol{x_i}) + \rho h(\boldsymbol{x_i}; \boldsymbol{a_m})), \tag{2.11}$$

y la aproximación es entonces,

$$F_m(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x}) + \rho_m h(\boldsymbol{x}; \boldsymbol{a_m}).$$

Esto significa que la minimización de (2.8) que involucra los parámetros $\{\beta_m, \boldsymbol{a_m}\}$, se reduce a la minimización de mínimos cuadrados dado en (2.10) para determinar los $\boldsymbol{a_m}$ ajustando a h a las pseudorespuestas $\{\widetilde{y}_i = -g_m(x_i)\}_1^N$, seguido de la minimización del único parámetro ρ en (2.11). Luego, con esto podemos concluir que, si $h(\boldsymbol{x}; a)$ es una función tal que (2.10) se puede resolver de forma adecuada, entonces este enfoque se puede usar para minimizar cualquier función de pérdida diferenciable L y F a la vez un modelo generado de forma secuencial y que es de forma aditiva. Por tanto, el algoritmo se puede describir de la siguiente manera.

Algorithm 1 Algoritmo aumento de gradiente (Gradient-Boost)

1: **Do** $F_0(\boldsymbol{x}) = \arg \min_{\rho} \sum_{i=1}^{N} L(y_i, \rho)$

2: **For** $m = 1, \ldots, M$, **do:**

3: $\widetilde{y}_i = -\frac{\partial L(y_i, F(\boldsymbol{x_i}))}{\partial F(\boldsymbol{x_i})}_{|F(\boldsymbol{x})=F_{m-1}(\boldsymbol{x})}$, $i = 1, \ldots, N$

4: $\boldsymbol{a_m} = \arg \min_{\boldsymbol{a_m}, \beta} \sum_{i=1}^{N} [\widetilde{y}_i - \beta h(\boldsymbol{x_i}; \boldsymbol{a})]^2$

5: $\rho_m = \arg \min_{\rho} \sum_{i=1}^{N} L(y_i, F_{m-1}(\boldsymbol{x_i}; \boldsymbol{a_m}) + \rho h(\boldsymbol{x_i}; \boldsymbol{a_m}))$

6: $F_m(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x}) + \rho_m h(\boldsymbol{x}; \boldsymbol{a_m})$

7: **end for**

Ejemplo 2.3. Considere el caso en el que la función de pérdida $L(\boldsymbol{x}, F)$ es mínimos cuadrados $(y - F)^2/2$. Nótese que en este caso las pseudorespuestas son $\widetilde{y} = y_i - F_{m-1}(x_i)$, por lo cual, en las líneas del código 1, 2 y 3 se mantienen, mientras que 4 y 5 se pueden reducir a determinar,

$$\{\rho_m, \boldsymbol{a_m}\} = \arg \min_{\boldsymbol{a}\ \rho} \sum_{i=1}^{N} \left[\widetilde{y}_i - \rho h(\boldsymbol{x_i}; \boldsymbol{a})\right]^2.$$

Ejemplo 2.4. Considere la función $L(\boldsymbol{x}, F)$ como el error absoluto $|y - F|$, en este caso las pseudorespuestas están dadas por,

$$\widetilde{y}_i = -\left[\frac{\partial L(y_i, F(\boldsymbol{x_i}))}{\partial F(\boldsymbol{x_i})}\right]_{F(\boldsymbol{x})=F_{m-1}(\boldsymbol{x})} = \text{sgn}\ (y_i - F_{m-1}(\boldsymbol{x_i})).$$

donde sgn es la función signo. Luego, los parámetros $\boldsymbol{a_m}$ se calculan como en la línea 4 del código, y para obtener la línea 5 se reescribe,

$$\begin{aligned}
\rho_m &= \arg \min_{\rho} \sum_{i=1}^{N} \mid y_i - F_{m-1}(\boldsymbol{x_i}) - \rho h(\boldsymbol{x_i}; \boldsymbol{a_m}) \mid \\
&= \arg \min_{\rho} \sum_{i=1}^{N} \mid h(\boldsymbol{x_i}, \boldsymbol{a_m}) \mid \left| \frac{y_i - F_{m-1}(\boldsymbol{x_i})}{h(\boldsymbol{x_i}, \boldsymbol{a_m})} - \rho \right| \\
&= \text{median}_W \left\{ \frac{y_i - F_{m-1}(\boldsymbol{x_i})}{h(\boldsymbol{x_i}; \boldsymbol{a_m})} \right\}_1^N, \quad w_i = |h(\boldsymbol{x_i}; \boldsymbol{a_m})|,
\end{aligned}$$

la función median_W es la mediana ponderada, donde los pesos son los términos w_i.

2.4.1. Impulso de gradiente extremo

El ensamblado de algoritmos es una técnica de Machine Learning, que hace uso de modelos simples para formar un algoritmo más preciso y eficiente. Existen distintas formas de ensamble: Bagging, Boosting, Stacking, entre otros.

En Boosting (Freund y Shapire, 1996) los algoritmos más simples son usados de forma secuencial, con el objetivo principal de aprovechar esta dependencia entre los modelos que se van generando para mejorar la precisión del modelo posterior. La idea para mejorar el rendimiento del nuevo modelo, es darle un peso más grande a los errores cometidos por el modelo previo, hasta obtener un modelo estable en la precisión.

Algunos ensambles que se basan en el principio del Boosting son; AdaBoost, XGBoost, CatBoost y LightGBM. XGBoost (Extreme Gradient Boosting por sus siglas en inglés) es un algoritmo desarrollado como proyecto de investigación en la Universidad de Washington por T. Chen y C. Guestrin, los cuales presentan su artículo [6] en 2016 en la conferencia SIGKDD. Este modelo está construido mediante la suma de modelos más simples basados en árboles de

regresión. El algoritmo de optimización consiste en el descenso por gradiente (Gradient Descent) de una función objetivo compuesta por la suma de funciones de pérdida individuales de cada observación. Estas funciones de pérdida miden la distancia entre una observación y su predicción basada en la suma de su estimación en la iteración previa más una nueva función que es agregada secuencialmente en cada iteración. Estas características permiten que la estimación global del modelo pueda ser escalable.

En algunas aplicaciones recientes se ha podido ver su eficiencia y escalabilidad en comparación con otros métodos. En [22] se implementó XGBoots para predecir la concentración de PM2.5 por hora. El método se comparó con los algoritmos bosque aleatorio, regresión lineal múltiple, árbol de decisión y las máquinas de vectores de soporte para modelos de regresión. Los resultados obtenidos demostraron que el algoritmo XGBoost supera a cada uno de estos métodos.

En [21] se diseña una tarea de conducción real para extraer datos que permitan modelar la dinámica del estrés al conducir y, propone un modelo de control del estrés del conductor basado en el comportamiento de conducción, el entorno y la familiaridad de la ruta. Basado en los datos psicológicos y resultados históricos del estrés del conductor, en el estudio se utilizó un análisis de clúster con K-means para la agrupación de observaciones y XGBoost para monitorear el estrés dentro de cada grupo. Las comparaciones del rendimiento con los otros modelos mostraron que el modelo XGBoost superó significativamente a los otros algoritmos de aprendizaje automático convencionales y a la mayoría de los modelos tradicionales, incluso sin el uso de datos psicológicos.

Modelo XGBoost

Considere un conjunto de datos $\{(x_i, y_i) : i = 1, \ldots, n,\ \ x_i \in \mathbb{R}^p,\ \ y_i \in \mathbb{R}\}$. Un modelo de ensamble basado en árboles de regresión hace uso de K modelos más simples, para predecir y_i de forma aditiva:

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i), \qquad f_k \in \mathbf{F},$$

$\widehat{y_i}$ es la predicción de y_i y $\mathbf{F}$ es la clase de funciones de todos los árboles de regresiones posibles;

$$\mathbf{F} \ = \left\{ f(x) = w_{q(x)} |\ \ q : \mathbb{R}^p \to I, \ \ w \in \mathbb{R}^T \right\}, \tag{2.12}$$

donde T es el número de hojas, $I = \{1, 2, \ldots, T\}$, $q(x)$ representa el índice $q(x)$-ésimo en el vector w, $w_{q(x)}$ representa la $q(x)$-ésima componente de w, f_k representa un árbol de regresión independiente que corresponde a una estructura de árbol q con pesos de hoja w.

Observación 2.5. Los árboles de regresión tienen una puntuación en cada una de sus hojas, w_i denota la puntuación en la i-ésima hoja de un árbol. Los datos usados en el entrenamiento del modelo quedan agrupados en los nodos hojas. Así, en la predicción de un ejemplo, se consideran las reglas de decisiones generadas por estos árboles para clasificarlo en las hojas correspondientes y calcular la predicción final, sumando cada una de las puntuaciones en las hojas dadas por el vector w de cada árbol [6].

Para aprender cada una de las funciones f_i se minimiza la función regularizada:

$$\mathcal{L} = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i}\right) + \sum_{k} \Omega(f_k), \tag{2.13}$$

donde $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$. l representa la función de pérdida entre el valor observado y_i y el valor predicho $\widehat{y_i}$. Ω es el término de regularizado que penaliza la complejidad del modelo, para evitar el sobreajuste de los datos, γ penaliza el número de hojas o equivalentemente la complejidad del árbol, T es el número de hojas en el árbol, λ es el parámetro de regularización y w es el vector de puntuación en las hojas.

Al considerar que $\widehat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \left\{\sum_{k=1}^{t-1} f_k(x_i)\right\} + f_t(x_i)$, la ecuación (2.13) se puede escribir como

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \widehat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t). \tag{2.14}$$

Se calcula la expansión de Taylor de orden 2 para la función de perdida:

$$\begin{aligned} l(y_i, \widehat{y}_i^{(t-1)} + f_t(x_i)) &= l(y_i, \widehat{y}_i^{(t-1)}) + \frac{\partial\ l(y_i, \widehat{y}_i^{(t-1)})}{\partial_{\widehat{y}_i^{(t-1)}}} f_t(x_i) \\ &+ \frac{1}{2}\frac{\partial\ l(y_i, \widehat{y}_i^{(t-1)})}{\partial^2_{\widehat{y}_i^{(t-1)}}} f_t^2(x_1), \end{aligned}$$

haciendo $g_i = \partial_{\widehat{y}_i^{(t-1)}} l(y_i, y_i^{(t-1)})$ y $h_i = \partial^2_{\widehat{y}_i^{(t-1)}} l(y_i, y_i^{(t-1)})$ de (2.14) se tiene:

$$\begin{aligned} \mathcal{L}^{(t)} &= \sum_{i=1}^{n} \left\{ l(y_i, \widehat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right\} + \Omega(f_t) \\ &= \sum_{i=1}^{n} \left\{ g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right\} + \Omega(f_t) + \sum_{i=1}^{n} l(y_i, \widehat{y}_i^{(t-1)}). \end{aligned} \tag{2.15}$$

De (2.15) se puede ver que el último término es constante positiva con respecto a f_t y por tanto a w, por lo cual dicho término puede ser eliminado. Luego,

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \left\{ g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right\} + \Omega(f_t). \tag{2.16}$$

Por otra parte, los nodos del árbol se pueden agrupar de la siguiente forma. Se denota por $I_j = \{i|q(x_i) = j\}$ para alguna estructura fija $q(x)$ al conjunto de índices cuyos ejemplos corresponden a dicha hoja. Así, al reescribir (2.16) y considerar (2.12) se obtiene

$$\begin{aligned} \mathcal{L}^{(t)} &= \sum_{i=1}^{n} \left\{ g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right\} + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} \left[(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2 \right] + \gamma T. \end{aligned} \tag{2.17}$$

Las sumas internas de la ecuación (2.17) sobre el conjunto de índices I_j de la hoja j, es debido a que los ejemplos que están en dicha hoja tienen la misma puntuación.

Posteriormente, para encontrar el valor óptimo de los parámetros w_j en cada una de las hojas, al fijar $q(x)$ y derivar (2.17), se determinó que dicho valor es:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}.$$

Sustituyendo en (2.17) se obtiene que el valor mínimo se alcanza en:

$$\mathcal{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T.$$

En la implementación del método se trata de optimizar un nivel del árbol a la vez, de modo que se va agregando de forma iterativa ramas a un árbol. Así, si se divide una hoja en dos hojas y el conjunto de índices de los ejemplos que están en la hoja izquierda y en la hoja derecha son I_I e I_D respectivamente, entonces la reducción de la pérdida después de la división está dado por:

$$L_{split} = \frac{1}{2} \left[\frac{(\sum_{i \in I_I} g_i)^2}{\sum_{i \in I_I} h_i + \lambda} + \frac{(\sum_{i \in I_D} g_i)^2}{\sum_{i \in I_D} h_i + \lambda} - \frac{(\sum_{i \in I_I \cup I_D} g_i)^2}{\sum_{i \in I_I \cup I_D} h_i + \lambda} \right] - \gamma. \tag{2.18}$$

Capítulo 3

Caso de estudio

En este capítulo se describe una aplicación basado en la distribución de valores extremos generalizados, haciendo uso de modelos árboles para analizar los niveles máximos de concentración de material particulado con un diámetro de menos de 2.5 micras (PM2.5) en el área metropolitana de la Ciudad de México durante el período 2003–2021. Para lo cual, se realizó lo siguiente: las tendencias espaciales se modelaron a través de un árbol de decisión en el contexto de un modelo GEV no estacionario. Se utilizó un modelo de conjunto de árboles como predictor de los parámetros GEV para aproximar las tendencias no lineales. El árbol de decisión se construyó utilizando un enfoque voraz por etapas, cuya función objetivo era el logaritmo de verosimilitud. Se verificó la validez del modelo mediante la verosimilitud y el criterio de información de Akaike (AIC). Con todo esto, se pudo verificar que, los mapas de los parámetros de valores extremos generalizados en el plano espacial muestran la existencia de tendencias locales diferenciadas en los valores extremos de PM2.5 en el área de estudio. Los resultados indicaron una fuerte evidencia de un aumento en la dirección oeste-este del área de estudio. Se construyó un mapa espacial de riesgo con niveles máximos de concentración de PM2.5 en un periodo de 25 años.

3.1. Antecedentes

La materia particulada con un diámetro de menos de 2.5 micrones es un contaminante del aire con efectos potencialmente negativos en los humanos. Llamada también contaminación por partículas, la materia particulada está compuesto principalmente por sulfatos, nitratos y carbono. Los sulfatos constituyen del 25 % al 55 % de la composición total de las PM2.5 y, junto con los nitratos, son el resultado de la transformación de las emisiones de dióxido de azufre de las centrales eléctricas e instalaciones industriales y las emisiones de óxido de nitrógeno de los automóviles, camiones y plantas eléctricas. El carbono se libera de las emisiones de automóviles, camiones, instalaciones industriales, incendios forestales, etc. Tanto el sulfato de amonio como el nitrato de amonio presentes en la atmósfera se forman a partir de fuentes como los fertilizantes y las operaciones de alimentación animal. La materia particulada puede estar en forma sólida o líquida en partículas como polvo, suciedad, ho-

llín o humo, e incluso algunas de estas pueden cambiar de una forma a otra [29].

El impacto negativo de PM2.5 en la salud humana se ha establecido en un número creciente de estudios [28], [30]. En el sistema respiratorio humano, los científicos han encontrado una correlación significativa entre PM2.5 y la morbilidad y mortalidad respiratoria [38]. Este efecto negativo en la salud de los seres humanos varía dependiendo de la concentración de PM en el aire y la susceptibilidad de la población, siendo los adultos mayores, mujeres embarazadas, adolescentes, lactantes y pacientes con problemas cardiopulmonares los más vulnerables [18], [9], [23]. En cuanto al efecto de la concentración de PM2.5, los resultados varían ligeramente; en el caso del cáncer de pulmón, un estudio de la American Cancer Society basado en una población de $500,000$ adultos reportó un aumento de mortalidad del $8\,\%$ por cada $10\ \mu g/m^3$ de aumento de PM2.5 [19], mientras que otro estudio rastreó a 1.2 millones de adultos estadounidenses y encontró que la mortalidad de cáncer de pulmón aumentó entre un $15\,\%$ y un $27\,\%$ [35]. La mortalidad global y la mortalidad por enfermedades cardiopulmonares aumentan un $4\,\%$ y un $6\,\%$, respectivamente, por cada $10\ \mu g/m^3$ de aumento de PM2.5, tras descartar ocupación, tabaquismo, dieta y otros factores de riesgo [19]. Según Zanobetti et al. [40], la mayor tasa de ingresos hospitalarios de emergencia por un aumento de $10\ \mu g/m^3$ en la concentración promedio de PM2.5 de 2 días es de $1.89\,\%$ para causas cardíacas, $2.25\,\%$ para infarto de miocardio, el $1.85\,\%$ por insuficiencia cardiaca congestiva, el $2.74\,\%$ por diabetes y el $2.07\,\%$ por ingresos respiratorios.

Al crecimiento exponencial de las ciudades le ha seguido un aumento de las emisiones de carbono y, en general, un aumento de la contaminación del aire por material particulado. Dicho crecimiento se ha dado de manera no uniforme principalmente en las grandes ciudades del mundo, generando áreas con diferentes densidades de población dentro de una misma ciudad. Esta distribución desigual de la población dentro de una misma región se refleja en una distribución similar de material particulado (PM), derivada de la correlación positiva entre la contaminación y las diferentes actividades antropomórficas. Un ejemplo de este caso se observa en el área metropolitana de la Ciudad de México, que es una de las áreas urbanas más grandes del mundo y también una región frecuentemente afectada por la contaminación del aire. Para monitorear las concentraciones atmosféricas de gases contaminantes, se han establecido varias estaciones de monitoreo. Todas estas estaciones recopilan observaciones cada hora del día sobre varios tipos de contaminantes. Sin embargo, a pesar de que esta región es una de las más desarrolladas del país, el número de estaciones de monitoreo es aún pequeño en comparación con la extensa área que cubre, por lo que extrapolar a regiones sin monitoreo es un desafío constante. Hinojosa-Baliño [17] realizó un análisis espacial de la distribución de la contaminación del aire PM2.5 en la Ciudad de México utilizando un modelo de regresión de uso de suelo, en el que se encontraron dos regiones con concentraciones altas y dos con concentraciones bajas. Si bien, el análisis permitió visualizar la distribución de las concentraciones de PM2.5, sus resultados no

permitieron hacer inferencias sobre riesgos futuros. Además, un análisis de valores extremos para determinar los riesgos futuros de concentraciones extremas de material particulado fue realizado por Aguirre-Salado et al., [1]. Este estudio se realizó para partículas de 10 μm o menos de diámetro, modelando los parámetros de una distribución de valores extremos mediante suavizado con funciones de base radial.

Se ha utilizado una amplia variedad de métodos para analizar la distribución espacial de concentraciones de PM2.5. El pronóstico a corto plazo ha sido evaluado utilizando modelos de regresión [17], modelos de series de tiempo [7], bosque aleatorio [13], máquinas de vectores de soporte [41] y redes neuronales [25] , entre otros. La teoría de los valores extremos se ha utilizado para evaluar los riesgos a largo plazo. En particular, el análisis de valores extremos con tendencias no estacionarias se ha ajustado aproximadamente bien al comportamiento no lineal observado en los extremos. Aunque la teoría de los valores extremos se basa en la distribución límite del máximo de una muestra aleatoria, conocida como distribución de valores extremos o distribución GEV, se han propuesto varios enfoques para obtener ajustes que se adapten adecuadamente a casos particulares de fenómenos observados. La mayoría de estos centran sus esfuerzos en aproximar la tendencia modelando el parámetro de ubicación de la distribución GEV. Aguirre-Salado et al., [1] realizó un estudio sobre la distribución espacial de PM10 en el que se modeló el efecto de la tendencia a través del parámetro de ubicación mediante una función suavizante de base radial. Otros estudios han propuesto modelar simultáneamente el parámetro de forma mediante una función seno [37], funciones lineales [34], splines [32], etc. En el caso del parámetro escala, ref. [39] modelos aditivos propuestos para aproximar el logaritmo del parámetro de escala.

Los árboles de regresión y clasificación son algoritmos de aprendizaje automático populares y eficientes introducidos por [4]. El algoritmo clasifica el espacio de variables independientes en una estructura de árbol y hoja mediante la optimización de alguna función objetivo, como una función de probabilidad o una función de pérdida. La estrategia de construir árboles se puede realizar de muchas maneras, algunas de las cuales se basan en la teoría de la información y la entropía; sin embargo, en muestras pequeñas, un algoritmo codicioso puede revisar secuencialmente todos los árboles posibles para determinar el óptimo. El modelo se puede regularizar agregando un término de penalización en la función objetivo [5] y ajustarse en un marco general de "impulso" de descenso de gradiente [11]. Esta característica permite que el algoritmo sea altamente paralelizable y adecuado para el análisis de Big Data [5].

El estudio pretende extender el uso de árboles de regresión al caso del análisis de valores extremos. En la regresión de valores extremos, el uso de árboles de regresión en la optimización de la probabilidad de la distribución generalizada de valores extremos es un poco más complejo, porque esta distribución tiene tres parámetros, en lugar de uno, como en la regresión con distribuciones normales y binomiales. Esta última característica crea la posibilidad de utili-

zar diferentes enfoques para su implementación. Por lo tanto, para construir un modelo parsimonioso, usamos la misma estructura de árbol para los tres parámetros con sus respectivos pesos.

3.2. Descripción de los datos

La Zona Metropolitana de la Ciudad de México (ZMC) se ubica en la región centro de México y está formada por 59 municipios del Estado de México y un municipio del Estado de Hidalgo. En la Figura (3.1) puede ver el área de estudio y los sitios primarios de muestreo tales como; Acolman (ACO), Ajusco (AJU), Ajusco Medio (AJM), Benito Juárez (BJU), Camarones (CAM), Centro de Ciencias de la Atmósfera (CCA), Coyoacán (COY), Gustavo A. Madero (GAM), Hospital General de México (HGM), Investigaciones Nucleares (INN), Merced (MER), Miguel Hidalgo (MGH), Montecillo (MON), Milpa Alta (MPA), Nezahualcóyotl (NEZ), Pedregal (PED), La Perla (PER), San Agustín (SAG), Santa FE (SFE), San Juan Aragón (SJA), Tlalnepantla (TLA), UAM Xochimilco (UAX), UAM Iztapalapa (UIZ), Xalostoc (XAL), FES Aragón (FAR) y Santiago Acahualtepec (SAC).

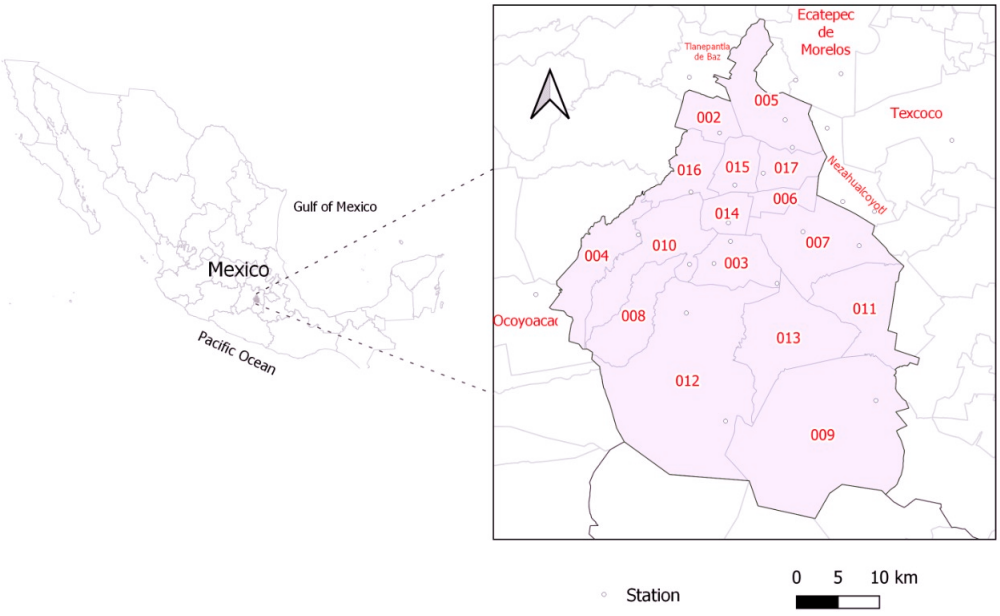


Figura 3.1: (**Izquierda**): México a nivel nacional. (**Derecha**): Ciudad de México con Alcaldías. 002: Azcapotzalco, 003: Coyoacán, 004: Cuajimalpa de Morelos, 005: Gustavo A. Madero, 006: Iztacalco, 007: Iztapalapa, 008: La Magdalena Contreras, 009: Milpa Alta, 010: Álvaro Obregón, 011: Tláhuac, 012 : Tlalpan, 013: Xochimilco, 014: Benito Juárez, 015: Cuauhtémoc, 016: Miguel Hidalgo, y 017: Venustiano Carranza.

Los datos utilizados en el presente estudio correspondieron a 2683 observaciones obtenidas por el método de los bloques máximos mensuales de PM2.5, donde cada bloque fue de 720 observaciones. El periodo de recopilación fue

entre el 2 de agosto de 2003 y el 11 de septiembre de 2021, obtenidas de 26 estaciones fijas de monitoreo del Sistema de Monitoreo Atmosférico (SIMAT) a través de la Red Automática de Monitoreo Atmosférico (RAMA), red establecida por la Secretaría del Medio Ambiente (SEDEMA) de la Ciudad de México, responsable de recopilar datos y reportar los niveles de calidad del aire. Cabe mencionar que SIMAT tiene un total de 69 estaciones, sin embargo, sólo 26 de estas estaciones miden PM2.5.

3.2.1. Enfoque propuesto

En los casos de aplicación donde se analizan valores extremos comúnmente se observan patrones complejos en el comportamiento de los datos. En consecuencia, ajustar los parámetros de la distribución VEG como si estos fueran de una sola muestra distribuida de forma idéntica no es consistente con la información que se observa en la muestra, ya que, como pudo verse en secciones anteriores la distribución VEG se construye sobre el supuesto de independencia. Una posible forma de resolver este problema es asumir que la muestra no está distribuida de manera idéntica en todas las regiones de un área espacial de estudio y ajustar los parámetros de manera similar al caso de los modelos lineales generalizados, por lo cual, esto conduce a considerar modelos de valores extremos no estacionarios. La novedad en este trabajo de tesis, está en la forma de asociar las covariables con los parámetros de la distribución VEG, ya que tradicionalmente, se asigna un predictor lineal de covariables a los parámetros de la distribución, como se describió en la sección 1.16. Sin embargo, en este estudio, se propuso un árbol de decisión basado en los resultados satisfactorios obtenidos en varias aplicaciones de aprendizaje automático. Además, cabe resaltar que dicha propuesta es una de las primeras implementaciones donde los árboles de decisión se ajustaban simultáneamente a más de un parámetro. Asumimos que las observaciones en la misma localidad espacial s tienen los mismos parámetros de forma y escala en la distribución VEG. Sin embargo, también asumimos que el parámetro de ubicación varía espacialmente según una tendencia que es modelada por su respectivo árbol de decisión. El modelo propuesto se describe a continuación.

3.2.1.1. Planteamiento del modelo

Sean $Y_1, \ldots, Y_n$ una muestra aleatoria y las variables aleatorias $M_n = \max\{Y_1, \ldots, Y_n\}$ tal que siguen una distribución de $VEG(\mu_t, \sigma_t, \kappa_t)$ no estacionario. Los parámetros de la distribución se proponen de la forma:

$$\mu_t = \sum_{k=1}^{K} u_k(x_t), \tag{3.1}$$

$$\kappa_t = \sum_{k=1}^{K} v_{s,k}(x_t), \tag{3.2}$$

$$\log \sigma_t = \sum_{k=1}^{K} w_{s,k}(x_t), \tag{3.3}$$

donde $u_{s,k}, v_{s,k}, w_{s,k} \in \mathbf{F}$, con $\mathbf{F}$ la clase de funciones de todos los árboles de regresión posibles definida en (2.12).

Nótese que el modelo propuesto asegura que localmente, la muestra proviene de la misma población, donde además se permite estimar la tendencia de forma conjunta en toda la región. Este enfoque permitió obtener un modelo regularizado, sin necesidad de incluir un término adicional en la probabilidad de regularizar el modelo. Una alternativa era dejar σ_t y κ_t libres, así como el parámetro de ubicación μ_t; sin embargo, Yee y Stephenson [39] observaron que permitir que el parámetro de forma sea libre hace que la estimación sea numéricamente inestable en los modelos parametrizados.

Por otro lado, estrategia de optimización utilizada dentro de cada árbol es descenso de gradiente, sección 2.2.1. De acuerdo con las ecuaciones (3.2) y (3.3), tanto los parámetros de forma κ_t como de escala σ_t en todas las observaciones son iguales en la misma estación de monitoreo. Sin embargo, la ecuación (3.1) muestra que se tiene el mismo valor del parámetro de ubicación para todas las observaciones que están en la misma hoja de un árbol. Se usan estas suposiciones basadas en el hecho de que las observaciones obtenidas en las mismas estaciones de monitoreo provienen de una distribución VEG con la misma forma y parámetro de escala.

Considere una muestra de n extremos $\underline{y} = (y_1, \ldots, y_n)$, el logaritmo de la verosimilitud para el modelo de VEG no estacionario $\ell_n(\mu_t, \sigma_t, \kappa_t \mid \underline{y})$ descrito en la sección 1.18.1 con la ecuación (1.34). Y considere además las derivadas parciales de dicha función ecuaciones (1.35), (1.36) y (1.37).

Sin pérdida de generalidad, se denotará;

$$\ell_n(\mu_t, \sigma_t, \kappa_t \mid \underline{y}) = \phi(F(\mathbf{x})) = \sum_{t=1}^{n} \ell(y_t, F(\underline{x}_t)),$$

siguiendo el procedimiento de optimización numérica descrito en la sección 2.2.1, considerando la solución de forma aditiva:

$$F(\underline{x}_t) = \sum_{k=1}^{K} f_k(\underline{x}_t),$$

donde,

$$f(\underline{x}_t) = \begin{pmatrix} u_k(\underline{x}_t) \\ v_{s,k}(\underline{x}_t) \\ w_{s,k}(\underline{x}_t) \end{pmatrix},$$

$u_0(\underline{x}_t)$, $v_{s,0}(\underline{x}_t)$ y $w_{s,0}(\underline{x}_t)$ son conjeturas iniciales para cada uno de los parámetros, y $\{u_k(\underline{x}_t)\}_1^K$, $\{v_{s,k}(\underline{x}_t)\}_1^K$ y $\{w_{s,k}(\underline{x}_t)\}_1^K$ son funciones incrementales ("pasos" o "impulsos") definidas por el método de optimización.

Para el descenso más empinado,

$$f_k(x_t) = -\rho_k g_k(\underline{x}_t),$$

con

$$g_k(x_t) = \left[\frac{\partial \ell(y_t, F(\underline{x}_t))}{\partial F(x_t)}\right]_{F(\mathbf{x})=F_{k-1}(\mathbf{x})}.$$

Cada una de las funciones K-simple $f_k(\underline{x}_t)$ puede obtenerse alternativamente mediante el uso de un enfoque "codicioso por etapas", por $k = 1, 2, \ldots, K$:

$$f_k(\underline{x}_t) = \arg\max_{f_k} \sum_{t=1}^{n} \ell(y_t, F_{k-1}(\underline{x}_t) + f_k(\underline{x}_t)).$$

El algoritmo anterior puede ser numéricamente eficiente; esto puede no ser óptimo, porque la suma de funciones es secuencial. Para entender esto, considere el modelo de regresión lineal múltiple. En estos modelos no podemos obtener una solución global óptima sumando las variables una a una, pero optimizando todo el conjunto de variables simultáneamente se puede obtener el óptimo. Similar a esta situación, la solución óptima, aunque en algunas situaciones inviable, es aquella que satisface la siguiente condición:

$$f_k(\underline{x}_t) = \arg\max_{f_k} \sum_{t=1}^{n} \ell\left(y_t, \sum_{k=1}^{K} f_k(\underline{x}_t)\right).$$

La estructura propuesta por el modelo permitió mejorar el ajuste a medida que aumenta el número de hojas sin incurrir en sobreajuste. Un modelo con pocas hojas tiene la ventaja de incluir las relaciones espaciales de las estaciones de monitoreo cercanas en un modelo parsimonioso; sin embargo, puede que no sea óptimo cuando se utiliza un enfoque "codicioso por etapas". Aumentar el número de pasos de "impulso" mejora el modelo sin incurrir en un sobreajuste; por lo tanto, no se requirió un estudio de simulación para el modelo propuesto.

3.2.2. Análisis y descripción de los datos

En este estudio, los valores extremos se obtuvieron utilizando el método de bloques máximos [15]. Los valores extremos se obtuvieron de un bloque grande con 720 observaciones, con datos medidos por hora. Por lo tanto, se obtuvo una muestra aleatoria de bloques máximos mensuales que es aproximadamente independiente. Además, se asume que el valor del parámetro de ubicación era espacialmente homogéneo entre estaciones cercanas, mientras que los parámetros de escala y forma son específicos para cada estación de monitoreo. La estimación del peso de las hojas de los árboles de decisión estimados se llevó a cabo secuencialmente, tomando las coordenadas de cada estación y calculando el log-verosimilitud resultante utilizando un algoritmo voraz que revisó todas las estaciones de monitoreo. La descripción gráfica del procedimiento global

se muestra en la Figura 3.2. Aunque este enfoque se consideró numéricamente eficiente, un esquema de búsqueda óptimo consistiría en formar todas las combinaciones posibles de árboles. Por lo tanto, se realizó la optimización utilizando el método de descenso de gradiente, con tamaños de paso de 0.05, 0.001 y 0.0001 para los parámetros de ubicación, escala y forma, respectivamente. Luego, se verificó que, al usar estas configuraciones el algoritmo siempre convergía a un punto crítico. La homogeneidad de los máximos en cada estación de monitoreo justificó que cada uno de estos tuviera su propio parámetro de escala y forma, dando lugar a un modelo simple y parsimonioso. Esta característica permitió que el modelo permaneciera regularizado sin necesidad de agregar términos adicionales en la función de verosimilitud. Para simplificar, se ajustaron los parámetros en cada hoja de un árbol, en lugar de usar funciones aditivas en cada hoja. Obsérvese que, aunque este enfoque era más lento, también era equivalente al otro.

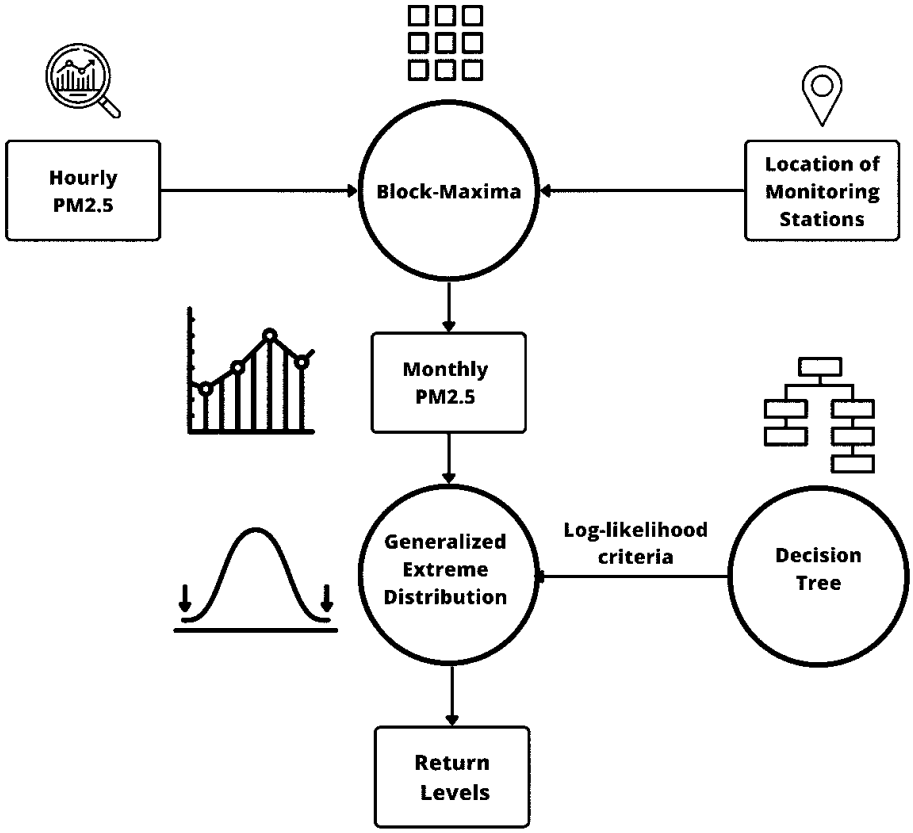


Figura 3.2: Gráfico del procedimiento global.

A continuación se presenta un breve resumen de los datos recopilados, en la tabla 3.2.2. Como puede verse, en dicha tabla se observa que cada estación de monitoreo tienen características paramétricas diferentes, además, se destaca la estación de monitoreo PER con altas concentraciones de PM2.5 medidas, en contraste con la estación AJM que tiene una concentración promedio más

baja, así como las estaciones SAC, SFE y PED. También se puede observar la existencia de valores altamente atípicos en la mayoría de las estaciones de monitoreo. La estación SAC tiene un valor atípico superior a 600 $\mu g/m^3$, que es más del doble de la magnitud del siguiente valor más alto en la misma estación de monitoreo. Sin embargo, el valor atípico más grande está en la estación XAL, con una concentración de 988 $\mu g/m^3$. Se ordenan las estaciones en el diagrama de caja de acuerdo a su ubicación geográfica y proximidad, observando que las siguientes estaciones comparten características distributivas similares. Esto justifica la elección del modelo propuesto, en el que el parámetro de ubicación se comparte entre estaciones cercanas.

Block ID	**Key**	**Long(W)**	**Lat(N)**	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
1	ACO	−98.9°	19.6°	35	62.8	87	100.5	117.8	281
14	AJU	−99.2°	19.2°	34	47	62	80.9	100	302
19	CCA	−99.2°	19.3°	30	48	56	65.3	72	302
16	INN	−99.4°	19.3°	20	42.2	51	58.8	59.5	246
18	AJM	−99.2°	19.3°	44	57	66	69.9	77.8	127
24	MGH	−99.2°	19.4°	51	64.5	71	86.1	90	267
25	BJU	−99.2°	19.4°	41	59	67	83.4	83	690
23	MER	−99.1°	19.4°	31	72.8	84.5	94.5	101	428
8	TLA	−99.2°	19.5°	33	73.5	86	93	104	294
9	FAR	−99°	19.5°	34	47.5	59	68.1	72.5	236
22	HGM	−99.2°	19.4°	36	74	90	97.2	107	346
20	PED	−99.2°	19.3°	41	59	69	73.8	78.5	179
26	COY	−99.2°	19.4°	21	72	85	95.6	105	544
3	SAC	−99°	19.3°	50	67	77	84.4	98	211
15	MPA	−99°	19.2°	46	57	65.5	84.8	103.5	211
11	SJA	−99.1°	19.5°	34	65.2	81.5	95.7	110	333
21	CAM	−99.2°	19.5°	43	71	83	95.7	99.2	777
2	MON	−98.9°	19.5°	29	50	65	73.3	83	227
6	UAX	−99.1°	19.3°	40	55	66.5	78.5	86.8	209
17	SFE	−99.3°	19.4°	28	56.5	70	72.9	81	179
5	PER	−99°	19.4°	53	87	125	167.2	200	681
10	GAM	−99.1°	19.5°	48	66	75	86.4	89	359
12	SAG	−99°	19.5°	36	65	77	98.4	101.8	698
13	XAL	−99.1°	19.5°	58	84	101	125.3	129	988
4	NEZ	−99°	19.4°	39	63.2	81	100.9	114	393
7	UIZ	−99.1°	19.4°	44	73	88	101.4	110.2	429

Tabla 3.1: Información resumida descriptiva de los máximos de $PM2.5$ en el área metropolitana de la Ciudad de México.

El análisis de la distribución de los datos es también de gran importancia para comprender mejor los datos, para ello, se realizaron diagramas de caja, Figura (3.3), esta gráfica muestra que las estaciones de monitoreo cercanas

comparten características similares, especialmente aquellas relacionadas con los cuantiles superiores de la distribución. Por tanto, estas propiedades en las observaciones hacen que la estimación de los parámetros mediante árboles de decisión pueda considerarse una elección adecuada para la estimación de los parámetros. Además, estas características, principalmente las basadas en la magnitud de los cuantiles podrían utilizarse para proponer heurísticas en el proceso de construcción de los árboles de decisión con el fin de llegar a soluciones globales.

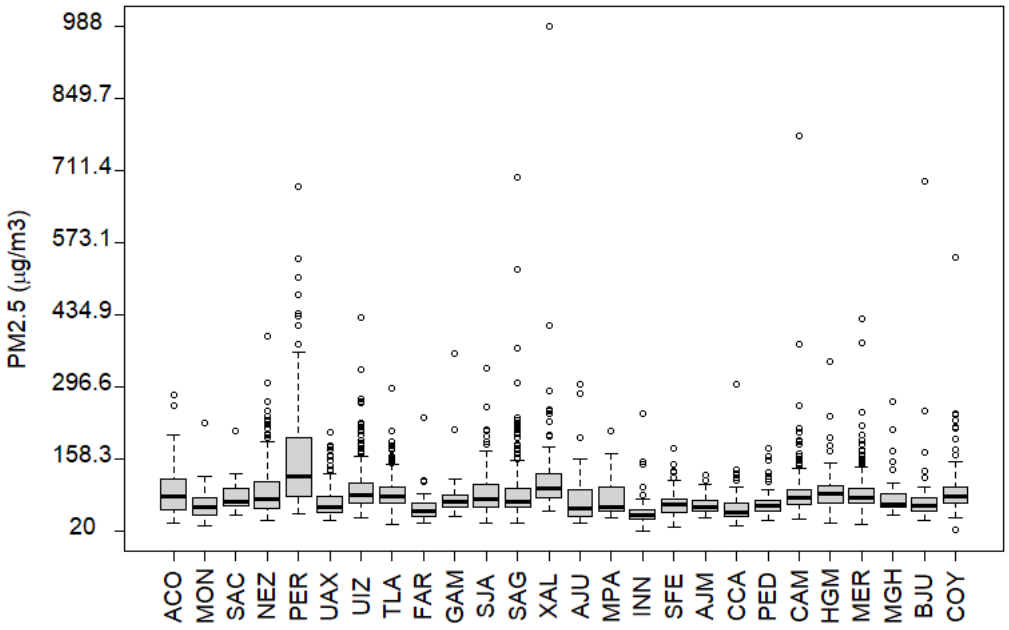


Figura 3.3: Box-plots de los máximos de PM2.5 en 26 estaciones de monitoreo en el área metropolitana de la Ciudad de México.

Por otra parte, la implementación del algoritmo codicioso por etapas, descrito en **Algorithm** 2 se realizó mediante el software estadístico *R* 4.1.2. Este algoritmo se basó en el logaritmo de la función verosimilitud y se utilizó para construir tanto el árbol como todos los demás algoritmos utilizados en esta investigación. Se denota por el símbolo I el conjunto inicial de instancias, es decir, el conjunto de todas las localidades geográficas donde se ubicaron las estaciones de monitoreo. Esto es, un elemento en I es un vector bivariado que contiene la latitud y longitud de alguna estación de monitoreo. También se denota por J el conjunto de vectores unitarios en $\mathbb{R}^2$ y la operación $\cdot$ como el producto interior habitual en $\mathbb{R}^2$. La propuesta para la división del árbol es, tomando un elemento en I y un elemento en J y evaluar el logaritmo de la verosimilitud del árbol resultante después de la división, luego se conserva el candidato con el valor del logaritmo de la verosimilitud más alta. De esta forma, puede notarse que esta etapa fue la más intensa computacionalmente, principalmente en las primeras divisiones del árbol donde cada rama tiene relativamente muchas observaciones. El algoritmo se repitió el número de veces definido en la variable *profundidad*.

Algorithm 2 Algoritmo codicioso por etapas para la búsqueda dividida

Input: I, Conjunto de instancias inicial
for i en **I**
  for j en **J**
    $L \leftarrow \max_{F} \sum_{k \in \{k | k \cdot j < i \cdot j\}} \ell\left(y_k, F\left(x_k\right)\right)$
    $R \leftarrow \max_{F} \sum_{k \in \{k | k \cdot j > i \cdot j\}} \ell\left(y_k, F\left(x_k\right)\right)$
    puntaje $\leftarrow$ L+R
  end for
end for
Output: Dividir con puntuación máxima

En términos de bondad del modelo, este se verificó mediante el uso del logaritmo de la verosimilitud y el criterio de información de Akaike (AIC) descritos en la Sección (1.40). Luego, debido a que la representación del modelo en (3.3) induce un modelo paramétrico en cada hoja, se calculó el AIC del árbol global sumando las contribuciones del AIC en cada hoja. El logaritmo de verosimilitud del modelo de valor extremo estacionario fue de $-13{,}061.27$, mientras que el logaritmo de verosimilitud del modelo final obtenido mediante el árbol de decisión fue de $-12{,}844.21$, lo que significa que el modelo final ha aprendido correctamente la distribución de los datos. Por otra parte, el AIC del modelo estacionario es de 26,128.54, mientras que el AIC del modelo propuesto es de 25,844.42. Más aún, para evaluar la bondad del modelo en términos gráficos, puede ver en Figura (3.4), el gráfico cuantil-cuantil, que demuestra un buen ajuste del modelo propuesto.

Por otra parte, cabe resaltar que en la práctica se pueden proponer varios algoritmos para obtener las reglas que dividan a un árbol de forma óptima. Chen y Guestrin [5], en modelos de un sólo parámetro, propusieron un algoritmo voraz exacto para la búsqueda dividida y un algoritmo aproximado para la búsqueda dividida basado en percentiles. En este estudio se uso un enfoque similar a un modelo multiparamétrico, usando cada componente del vector de parámetros simultáneamente como candidatos homogéneos para la búsqueda dividida. Esta configuración disminuyó el tiempo de procesamiento; sin embargo, también disminuyó el espacio de solución. Por lo tanto, con esta configuración se obtuvo un equilibrio entre el tiempo de procesamiento y la cantidad de árboles visitados.

El árbol de decisión ajustado se muestra en la Figura (3.5). Se observó que la probabilidad en el nodo raíz del árbol era igual a la probabilidad del modelo estacionario. Al final de la décima profundidad del árbol, la ramificación mejoró la probabilidad logarítmica a $-12{,}844.21$, lo que es suficiente para asegurar que el modelo mejoró estadísticamente. Luego, para dividir el árbol y agregar más ramas, se eligió el nodo que maximizaba la suma de la probabilidad logarítmica de las hojas resultantes, para cada uno de todos los posibles nodos candidatos.

Por tanto, la independencia entre los ajustes realizados en cada nodo tiene la ventaja de permitir la implementación del algoritmo en paralelo. Además, este algoritmo no garantiza que sea óptimo en un sentido global, porque el árbol resultante puede ser atraído por una solución local.

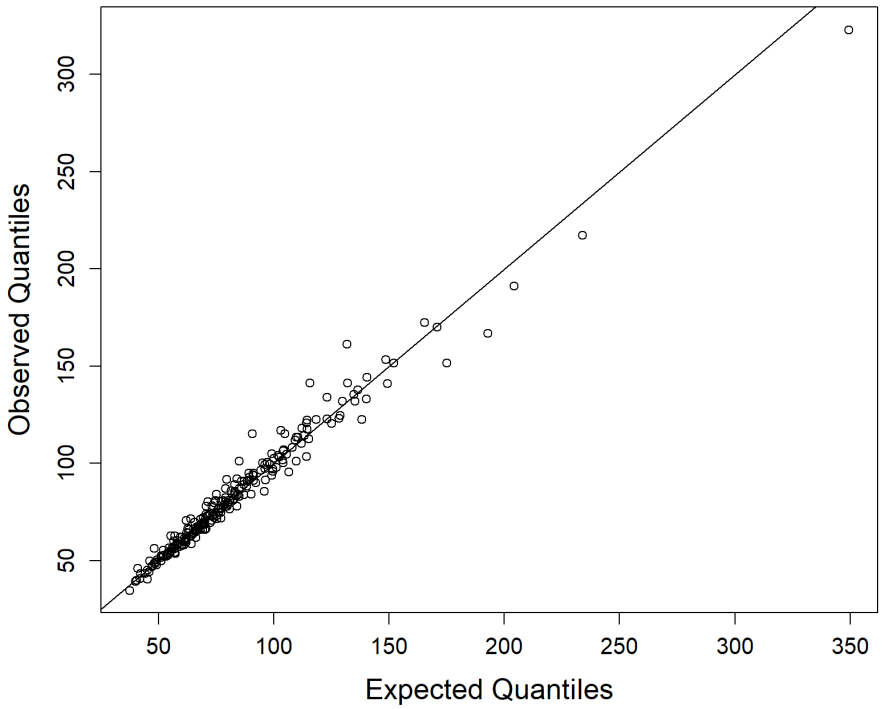


Figura 3.4: Gráfica cuantil-cuantil de PM2.5 máxima en el área metropolitana de la CDMX.

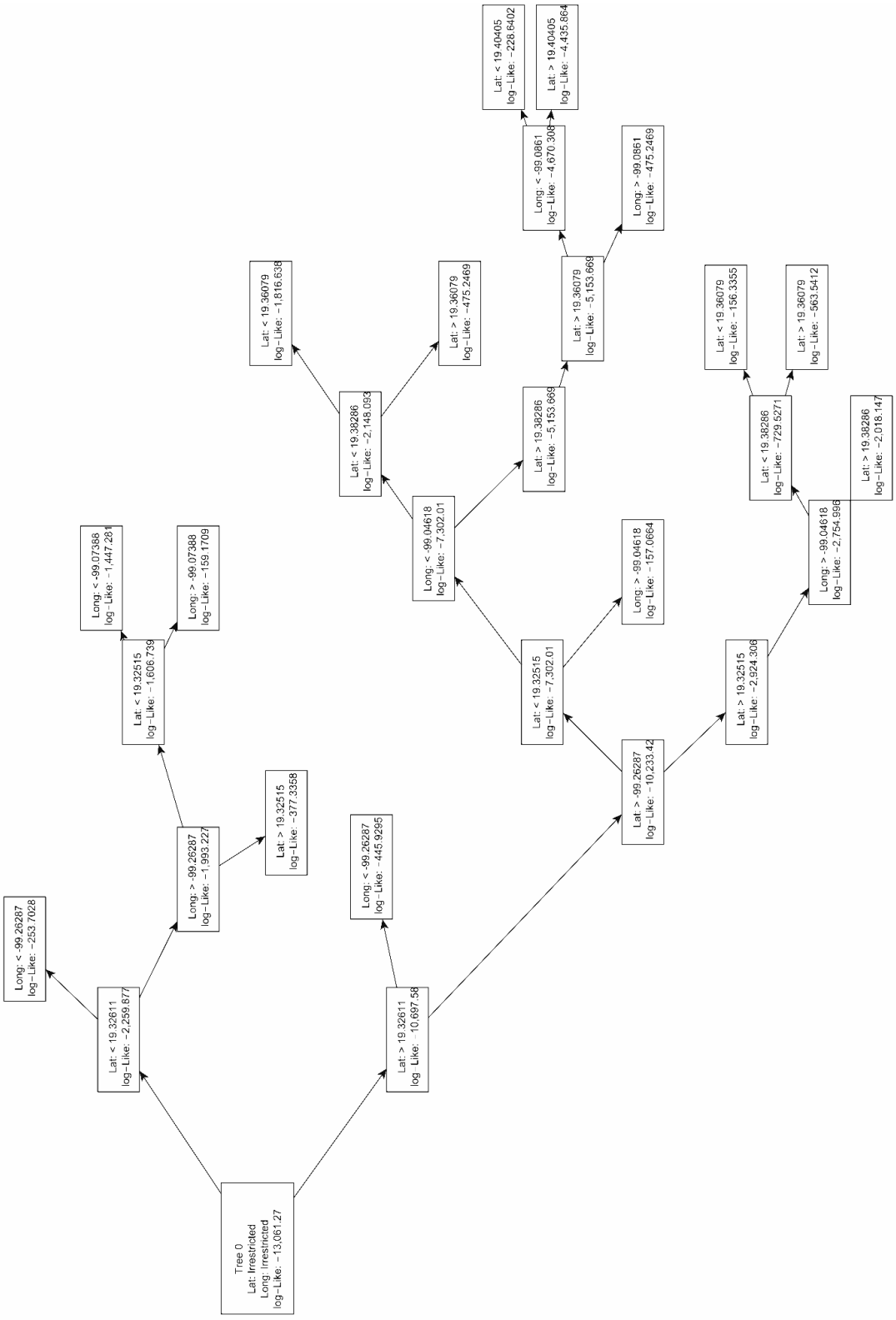


Figura 3.5: Regla de ramificación obtenidas para formar el árbol de decisión de los parámetros de distribución VEG de los máximos de PM2.5 en la ZMCM.

El árbol de decisión ajustado para cada uno de los parámetros se puede ver en la Figura 3.6. En el caso del parámetro de ubicación μ_t, Figura 3.6 (a), se puede observar que las estaciones de monitoreo han sido agrupadas espacialmente de acuerdo a su función de verosimilitud usando el algoritmo voraz por etapas (gredy stagewise). Como puede verse, se determinó que los niveles más bajos para el parámetro de ubicación de la distribución VEG se encuentran en la región suroeste. En esta región, el valor estimado para el parámetro de ubicación es de aproximadamente 50. En contraste, con los valores más altos para el parámetro de ubicación, que se ubican cerca del meridiano -99.2 W y el paralelo 19.5 N. En la Figura 3.6 (b), se muestra un mapa con las estimaciones para el parámetro de escala. Se puede observar que las características del mapa son similares al parámetro de ubicación. Aunque, estos tienen diferencias notables en la región noreste, donde el parámetro de escala tiende a aumentar en lugar de disminuir, en contraste con el parámetro de ubicación. Una situación similar ocurrió entre los paralelos 19.2 N y 19.4 N y al este del meridiano -99.3 W. Se observa un comportamiento totalmente diferente con el parámetro de forma. Por su parte, el árbol de decisión ajustado para el parámetro de forma se muestra en la Figura 3.6 (c). Esta figura muestra que, en general, el parámetro es positivo, con una tendencia creciente en dirección suroeste a noreste, concentrándose los valores más altos en la región central y noreste del área de estudio.

Los resultados de la Figura (3.6) también muestran la existencia de zonas geográficas con distribuciones de colas pesadas en áreas aledañas a las coordenadas -99 W y 19.4 N. Estos hallazgos son similares a los encontrados por Hinojosa-Baliño et al., [17] sobre las concentraciones diarias de PM2.5 en las mismas regiones de la Ciudad de México. Estos resultados muestran que, en general, en la región este del área de estudio se observaron los valores más altos de concentraciones de PM2.5. Los resultados que se encontraron en este estudio también coincidieron con los resultados obtenidos por su investigación en regiones de la parte sur y suroeste del área de estudio. Los resultados son similares en ambas investigaciones, mostrando una correlación positiva entre el parámetro de ubicación de la distribución VEG y las observaciones diarias de PM2.5 analizadas por ellos.

Por otra parte, en el análisis predictivo de valores extremos, es común el análisis de los niveles de retorno. Los niveles de retorno z_p en este caso de aplicación son niveles de concentración cuyos valores se espera superar una vez cada $1/p$ año. En este estudio, el mapa de niveles de retorno de 25 años se muestra en la Figura (3.7). Las estimaciones del nivel de retorno en las estaciones de monitoreo se obtuvieron usando el modelo (3.1), y extrapolamos estos valores en el mapa usando el algoritmo de ponderación de distancia inversa, con el objetivo de suavizar el mapa de niveles de retorno. Del mapa, puede verse que los niveles de retorno más altos se esperan en áreas cercanas a la estación PER, mientras que los niveles de retorno más bajos se esperan en áreas cercanas a la estación AJM. Además, se observa que los niveles de retorno tienden a aumentar en dirección este-oeste y disminuyen nuevamente

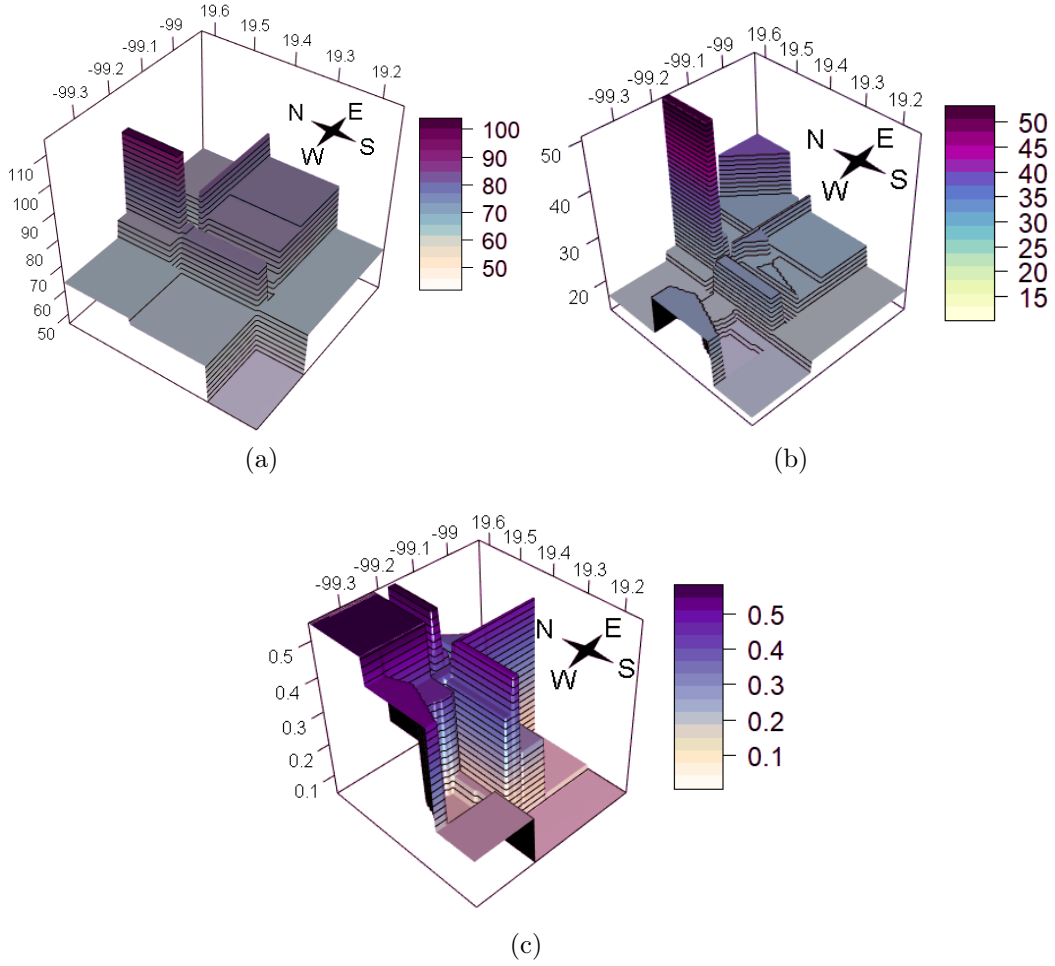


Figura 3.6: (**a**) Representación tridimensional del árbol de decisión ajustado al parámetro de ubicación, (**b**) parámetro de escala y (**c**) parámetro de forma. Los ejes X y Y están en coordenadas geográficas (grados decimales). Z es el valor calculado del parámetro correspondiente (ubicación, escala y forma), para cada posición geográfica.

después de la estación MON. El mapa también muestra la característica del modelo de agrupar estaciones cercanas según su parámetro de ubicación, lo que conlleva a un mapa más homogéneo y un modelo con menos parámetros. Una característica importante del mapa fue la suavidad de las estimaciones de una estación de monitoreo a otra, además de la estabilidad de las estimaciones. De hecho, los modelos que asocian cada observación con una única distribución tienden a sobreajustar los datos, provocando estimaciones poco realistas.

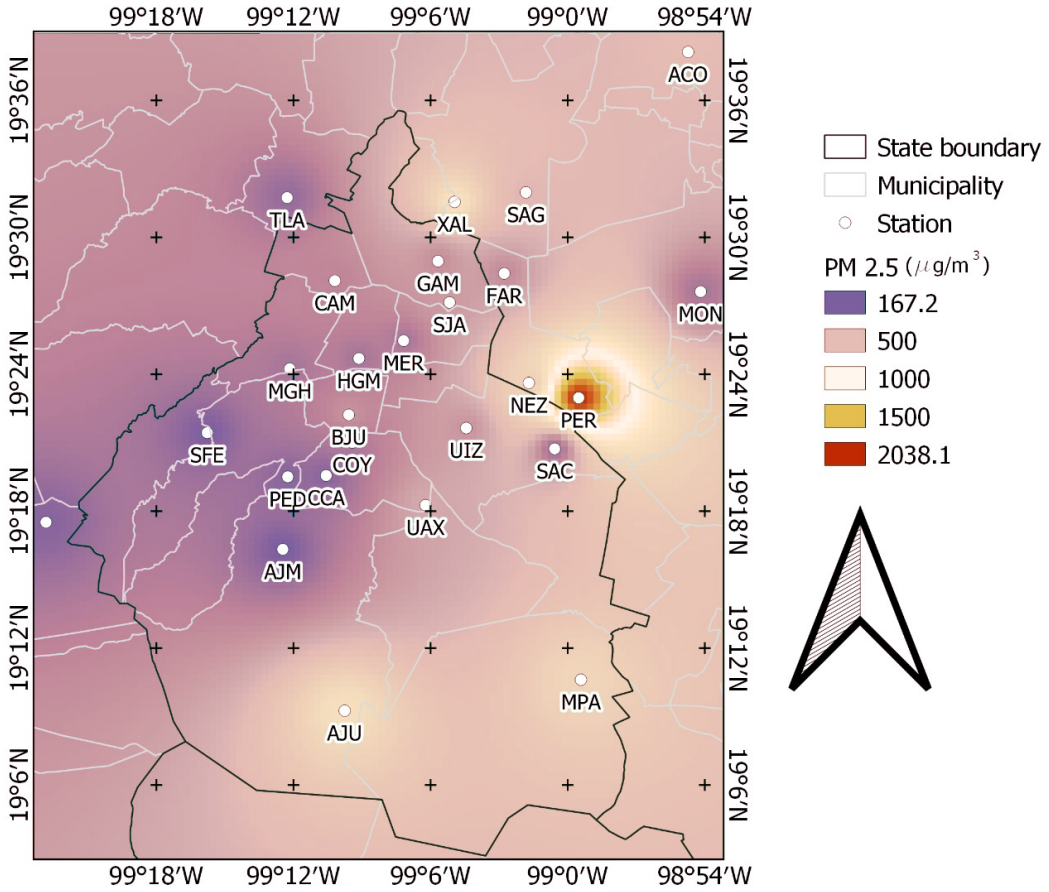


Figura 3.7: Distribución espacial de PM2.5 para un período de retorno de 25 años para la región de estudio.

Comparando los resultados obtenidos en este estudio, con los hallazgos de una investigación similar sobre PM10 considerando la misma área de estudio. En [1] desarrollaron un modelo jerárquico para el análisis espacial de los extremos de contaminación por PM10 en la Zona Metropolitana de la Ciudad de México. Basaron sus estimaciones en funciones de suavizado radiales y modelaron espacialmente sólo el parámetro de ubicación, lo que conllevo a que los parámetros de escala y forma se consideraran constantes. Por lo tanto, se modelaron los valores extremos no estacionarios, obteniendo un patrón de incremento lineal en dirección sureste–noroeste, como se muestra esquemáticamente en Figura 3.8. Aquí, se encontró un patrón similar con respecto a los valores extremos de PM2.5. Sin embargo, la tendencia se modificó levemente, resultando en una dirección de aumento en el oeste-este. Además, el árbol de

decisión propuesto es paramétricamente más simple y estable, lo que se reflejó en la adecuada convergencia obtenida en cada uno de los pasos de ajuste.

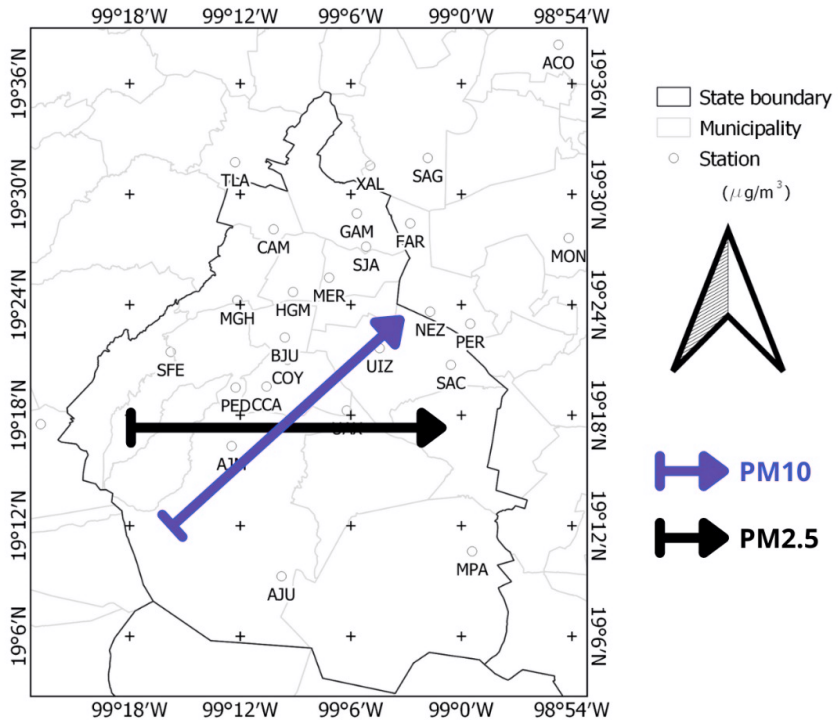


Figura 3.8: Comparación espacial de las tendencias de aumento en la región de estudio. PM10 (linea azul, Aguirre-Salado, [1]) y PM2.5 (linea negra).

La ventaja de considerar las observaciones de la misma estación de monitoreo como elementos de la misma distribución se observa en la Figura 3.7. Estudios previos sobre la distribución de valores extremos no estacionarios de material particulado en la zona metropolitana de la Ciudad de México no se restringió el modelo para evitar observaciones obtenidas en la misma ubicación geográfica [1], sin embargo, se asumió en el estudio un modelo flexible para representar las condiciones observadas en cada medición, lo cual provocó un sesgo cuando se consideran los elementos de las colas de distribución, esto a su vez resultó en estimaciones irreales e inestables. Más aún, tal situación generó modelos inestables y no robustos, y cuando se agregaban o eliminaban las observaciones, esto producía estimadores drásticamente diferentes.
En contraste, el modelo propuesto en esta investigación permitió considerar las observaciones de una misma estación de monitoreo con una distribución idéntica, dando lugar a estimadores robustos, que además no presentaban los problemas comunes de la no convergencia del algoritmo de estimación.

Se observó además que, el algoritmo voraz por etapas y cualquier otro algoritmo voraz que agrega funciones incrementales, como “pasos” o “impulso”, tiene un alto riesgo de generar árboles de decisión que satisfagan las condiciones de los óptimos locales. El trabajo adicional puede involucrar la propuesta de algoritmos, en el contexto de la teoría del valor extremo, utilizando los cuantiles superiores de los máximos como distancias para la agrupación de estaciones

de monitoreo y aumentando el número de nodos en grupos, no secuencialmente uno por uno. Adicionalmente, se debe investigar si al proponer los nodos candidatos en las hojas de cada árbol de decisión candidato, se mejora la ganancia en el log-verosimilitud. Creemos que en esta situación, la pérdida respectiva de rendimiento del algoritmo será seguida por una ganancia en la verosimilitud logarítmica.

Capítulo 4

Conclusiones

La teoría de Valores Extremos en una herramienta estadística importante utilizada para analizar problemas aplicados, en donde el objetivo es modelar o evaluar la probabilidad de eventos (o de valores) extremos que los que fueron registrados anteriormente. Por tal motivo, producir un buen modelo en términos de precisión de los valores predichos y observados es fundamental. Esto a su vez, muestra la importancia de una correcta selección de los parámetros del modelo propuesto.

En este trabajo de investigación, el objetivo principal es proponer una nueva metodología que nos permita estimar los parámetros de una DVEG no estacionario por medio de árboles de decisión, el algoritmo voraz y el método de descenso de gradiente. No obstante, cabe señalar que en la actualidad existen diversos métodos o algoritmos sobre cómo abordar dicho problema. Pese a ello, dada la gran diversidad del comportamiento de los datos, en muchos casos contar con nuevas propuestas para abordar dicha problemática permite poder obtener mejores resultados que con las metodologías existentes, más aún, proporciona una perspectiva más amplia sobre cómo abordar diversos problemas, y no sólo relacionados a la teoría de valores extremos.

Sin embargo, una vez seleccionado el modelo, cabe señalar que existen algunas consideraciones importantes que deben de tenerse en cuenta en el proceso de inferencia de la distribución de VEG para un correcto resultado, algunas de estas son:

1. En el método de bloque máximo, la elección del tamaño del bloque es sumamente importante, ya que un bloque demasiado pequeño generaría un modelo límite poco eficiente, produciendo un sesgo en la estimación y extrapolación de los datos. En cambio, un bloque excesivamente grande produciría pocos bloques, lo que a su vez conlleva a una gran varianza en la estimación.

2. La función de verosimilitud de la distribución generalizada no satisface las condiciones suficientes de regularidad que son indispensables para la estimación de los parámetros óptimos. Por lo cual, algunas alternativas que existen son; técnicas gráficas, métodos basados en momentos, basadas en probabilidad, entre otros.

3. La construcción del árbol puede producir un sobreajuste de los datos, como en el caso de árboles complejos con un parámetro de profundidad alto; mientras que uno simple, podría resultar en una capacidad predictiva pobre. Por tal motivo, proporcionar los parámetros adecuados relativos a la profundidad, números de ramas, u algún criterio de parada, son de vital importancia.

4. El criterio de división de las ramas del árbol, mostrado en la descripción del Algoritmo (2), es bastante costoso computacionalmente, sin embargo, como el número de estaciones en este trabajo es pequeño, el algoritmo converge satisfactoriamente, más aún, si el número de covariables fuera muy grande, probablemente este criterio no sería viable para la construcción del árbol.

5. En este trabajo nosotros utilizamos la teoría de valores extremos no estacionarios para modelar los máximos de PM2.5 empleando un conjunto de árboles. Los principales hallazgos encontrados fueron: el modelo tiene la ventaja de aproximar tendencias espaciales no lineales complejas de valores extremos, mediante un modelo de ensamblaje basado en árboles de decisión para los parámetros de la distribución de VEG que hace uso de un modelo K más simple. Los parámetros en cada hoja se estimaron mediante el método de descenso de gradiente, lo que tiene la ventaja de ser fácil de implementar, y por lo cual se garantiza la convergencia a la solución óptima en cada hoja. Además, las estimaciones se obtuvieron ajustando el modelo de VEG no estacionario usando un enfoque de árbol de decisión simultáneamente para los tres parámetros de la distribución de valores extremos, lo cual consideramos es la principal aportación de este trabajo, ya que proporciona una forma novedosa de realizar estimaciones en modelos de parámetros multivariados mediante el uso de árboles de decisión. Este modelo fue validado comparando el log-verosimilitud y el AIC del modelo estacionario con los obtenidos para el modelo ajustado, resultando que los mejores valores los obtiene el modelo propuesto, mostrando así el soporte y la validez de los resultados obtenidos. Un cambio importante para extender el modelo debería considerar la construcción del árbol de decisión empleando conglomerados basados en la proximidad de las estaciones de monitoreo y sus cuantiles superiores; esto debería ayudar a resolver el problema de las soluciones mínimos locales que se pueden obtener mediante el enfoque codicioso por etapas.

Bibliografía

[1] Aguirre-Salado, A. I., Vaquera-Huerta, H., Aguirre-Salado, C. A., Reyes-Mora, S., Olvera-Cervantes, A. D., Lancho-Romero, G. A., and Soubervielle-Montalvo, C. Developing a hierarchical model for the spatial analysis of pm10 pollution extremes in the mexico city metropolitan area. *International Journal of Environmental Research and Public Health 14*, 7 (2017).

[2] Aguirre-Salado, A. I., Venancio-Guzmán, S., Aguirre-Salado, C. A., and Santiago-Santos, A. A novel tree ensemble model to approximate the generalized extreme value distribution parameters of the pm2.5 maxima in the mexico city metropolitan area. *Mathematics 10*, 12 (2022).

[3] Barajas, F. H. *Modelos predictivos*. https:fhernanb.github.io/libro_mod_pred, 2023.

[4] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. *Classification And Regression Trees*. Routledge, Oct. 2017.

[5] Chen, T., and Guestrin, C. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Aug. 2016), ACM.

[6] Chen, T., and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016), pp. 785–794.

[7] Chiang, P.-W., and Horng, S.-J. Hybrid time-series framework for daily-based pm2.5 forecasting. *IEEE Access 9* (2021), 104162–104176.

[8] Coles, S., Bawa, J., Trenner, L., and Dorazio, P. *An introduction to statistical modeling of extreme values*, vol. 208. Springer, 2001.

[9] de Oliveira, B. F. A., Ignotti, E., Artaxo, P., do Nascimento Saldiva, P. H., Junger, W. L., and Hacon, S. Risk assessment of PM2.5 to child residents in brazilian amazon region with biofuel production. *Environmental Health 11*, 1 (Sept. 2012).

[10] Fisher, R., and Tippett, L. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society* (1928), 180–190.

[11] Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics 29* (1999), 1189–1232.

[12] Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[13] Geng, G., Meng, X., He, K., and Liu, Y. Random forest models for pm2.5 speciation concentrations using MISR fractional AODs. *Environmental Research Letters 15*, 3 (mar 2020), 034056.

[14] Gnedenko, B. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of mathematics* (1943), 423–453.

[15] Gumbel, E. *Statistics of extremes*. Columbia University Press, 1958.

[16] Haan, L., and Ferreira, A. *Extreme value theory: an introduction*, vol. 3. Springer, 2006.

[17] Hinojosa-Baliño, I., Infante-Vázquez, O., and Vallejo, M. Distribution of pm2.5 air pollution in mexico city: Spatial analysis with land-use regression model. *Applied Sciences 9*, 14 (2019).

[18] Huynh, M., Woodruff, T. J., Parker, J. D., and Schoendorf, K. C. Relationships between air pollution and preterm birth in california. *Paediatric and Perinatal Epidemiology 20*, 6 (Nov. 2006), 454–461.

[19] III, C. A. P. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA 287*, 9 (Mar. 2002), 1132.

[20] Karr, A. F., and Karr, A. F. Convergence of random variables. *Probability* (1993), 135–162.

[21] Lu, Y., Fu, X., Guo, E., and Tang, F. Xgboost algorithm-based monitoring model for urban driving stress: Combining driving behaviour, driving environment, and route familiarity. *IEEE Access 9* (2021), 21921–21938.

[22] Ma, J., Yu, Z., Qu, Y., Xu, J., Cao, Y., et al. Application of the xgboost machine learning method in pm2. 5 prediction: A case study of shanghai. *Aerosol and Air Quality Research 20*, 1 (2020), 128–138.

[23] Martinelli, N., Girelli, D., Cigolini, D., Sandri, M., Ricci, G., Rocca, G., and Olivieri, O. Access rate to the emergency department for venous thromboembolism in relationship with coarse and fine particulate matter air pollution. *PLoS ONE 7*, 4 (Apr. 2012), e34831.

[24] Martín Pliego, F. J., and Ruíz Pérez, L. *Fundamentos de probabilidad*. Alfa Centauro, 2006.

[25] Masinde, C. J., Gitahi, J., and Hahn, M. Training recurrent neural networks for particulate matter concentration prediction. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLIII-B2-2020* (Aug. 2020), 1575–1582.

[26] Mélice, J.-L., and Reason, C. J. Return period of extreme rainfall at george, south africa. *South African Journal of science 103*, 11-12 (2007), 499–501.

[27] Mises, R. v. La distribution de la plus grande de n valeurs. *Rev. Math. Union Interbalcanique 1* (1936), 141–160.

[28] Nemery, B., Hoet, P. H., and Nemmar, A. The meuse valley fog of 1930: an air pollution disaster. *The Lancet 357*, 9257 (Mar. 2001), 704–708.

[29] None. The particle pollution report: current understanding of air quality and emissions through 2003, Dec 2004.

[30] Orru, H., Maasikmets, M., Lai, T., Tamm, T., Kaasik, M., Kimmel, V., Orru, K., Merisalu, E., and Forsberg, B. Health impacts of particulate matter in five major estonian towns: main sources of exposure and local differences. *Air Quality, Atmosphere Health 4*, 3-4 (June 2010), 247–258.

[31] Resnick, S. *A probability path.* Springer, 2019.

[32] Rosen, O., and Cohen, A. Extreme percentile regression. *In: Härdle, W. and M.G. Schimek, (eds.) Statistical Theory and Computational Aspects of Smoothing: Proceedings of the COMPSTAT '94 Satellite Meeting held in Semmering, Austria, August 1994,Physica-Verlag, Heidelber* (1996), 27–28.

[33] Ruíz, J. L. C. Análisis de temperaturas máximas en el estado de oaxaca. *Tesis, de Universidad Tecnólogica se la Mixteca* (2020).

[34] Tawn, J. Bivariate extreme value theory: models and estimation. *Biometrika 75* (1988), 397–415.

[35] Turner, M. C., Krewski, D., Pope, C. A., Chen, Y., Gapstur, S. M., and Thun, M. J. Long-term ambient fine particulate matter air pollution and lung cancer in a large cohort of never-smokers. *American Journal of Respiratory and Critical Care Medicine 184*, 12 (Dec. 2011), 1374–1381.

[36] Wackerly, D. D., Mendenhall III, W., and Scheaffer, R. L. *Estadística matemática con aplicaciones.* México: Cengage Learning, 2008.

[37] Weissman, I. Estimation of parameters and large quantiles based on the k largest observations. *J. Am. Stat. Assoc. 73* (1978), 812–815.

[38] Xing, Y.-F., Xu, Y.-H., Shi, M.-H., and Lian, Y.-X. The impact of pm2.5 on the human respiratory system. *Journal of thoracic disease 8*, 1 (Jan 2016), E69–E74.

[39] Yee, T. W., and Stephenson, A. G. Vector generalized linear and additive extreme value models. *Extremes 10* (2007), 1–19.

[40] Zanobetti, A., Franklin, M., Koutrakis, P., and Schwartz, J. Fine particulate air pollution and its components in association with cause-specific emergency admissions. *Environmental Health 8*, 1 (Dec. 2009).

[41] Zhang, C.-J., Dai, L.-J., and Ma, L.-M. Rolling forecasting model of PM2.5 concentration based on support vector machine and particle swarm optimization. In *Hyperspectral Remote Sensing Applications and Environmental Monitoring and Safety Testing Technology* (2016), vol. 10156, International Society for Optics and Photonics, SPIE, pp. 387 – 394.