Universidad Nacional Autónoma de México

Facultad de Ciencias

"APLICACIÓN DE LA TEORÍA DE VALORES EXTREMOS EN EL ANÁLISIS DE EVENTOS HIDROMETEOROLÓGICOS"

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIO

PRESENTA:

ISAAC GONZÁLEZ GARCÍA

DIRECTOR DE TESIS:

DRA. ANA MEDA GUARDIOLA

2014

Ciudad Universitaria, D. F.

1. Datos del alumno
González
García
Isaac
58154215
Universidad Nacional Autónoma de México
Facultad de Ciencias
306662355

2. Datos del tutor
Doctora
Ana
Meda
Guardiola

3. Datos del sinodal 1
Doctor
Rodolfo
Silva
Casarín

4. Datos del sinodal 2
Doctor
Edgar Gerardo
Mendoza
Baldwin

5. Datos del sinodal 3
Doctor
Fernando
Baltazar
Larios

6. Datos del sinodal 4
Maestro en Ciencias
Fernando
Guerrero
Poblete

7. Datos del trabajo escrito
Aplicación de la teoría de Valores Extremos en el Análisis de Eventos Hidrometeorológicos
117 p
2014

Cada vez que un dueño de la tierra proclama 'para quitarme este patrimonio tendrán que pasar sobre mi cadáver' debería tener en cuenta que a veces pasan.

Cálculo de probabilidades. Mario Benedetti.

A mis padres.

Agradecimientos

Agradezco en primer lugar a la Universidad Nacional Autónoma de México por abrirme sus puertas y permitirme comenzar y culminar una carrera universitaria. Particularmente agradezco a la Facultad de Ciencias y a la Doctora Ana Meda Guardiola por su paciencia, guía, atenciones y apoyo a lo largo de todo este proceso. Al Instituto de Ingeniería de la UNAM y al Doctor Rodolfo Silva Casarín, al Doctor Edgar Mendoza Baldwin, a la Maestra Mireille Del Carmen Escudero Castillo, a la Coordinación de Hidráulica y al Laboratorio de Costas y Puertos por todo el apoyo brindado a lo largo de la realización de este trabajo, y por proporcionar los recursos, el espacio para trabajar, la información necesaria, la guía, y todo aquello sin lo cual no habría sido posible la conclusión de la presente tesis.

También agradezco el apoyo incondicional de mis padres Guillermo González Jiménez y María Teresa García Cabañas a lo largo de mi carrera y de toda mi existencia. A la compañía y ánimo brindado por mi hermana Maura, mi abuelita Higinia Cabañas Ponce y todos mis demás familiares. A Mayela, Andrea, Helver, Omar, Samuel, Ivonne, Fernando y a mis demás amigos y compañeros quienes a lo largo de los semestres se convirtieron en parte fundamental de mi desarrollo y de mi vida, que aún si me faltó mencionar a muchos de ellos, todos son igual de importantes. A mis profesores que fueron mi guía e inspiración para seguir adelante, en particular al Doctor Javier Páez Cárdenas, de quien recibí apoyo siempre que lo necesité. A mis compañeros y vecinos del laboratorio de Costas y Puertos por haber hecho muy agradable esta etapa que fue la escritura de la tesis. Y a muchas personas más para quienes no bastaría este espacio para agradecerles todo lo que han hecho por mí.

Isaac González García

Índice general

Índice de figuras

Índice de tablas

Notación y símbolos

Símbolo	**Denota (se define en la p.)**
$[x]$	Parte entera de x (pág. 8)
$\mathbb{R}$	Los números reales
$\mathbb{R}_+$	Los números reales positivos
$\mathbb{N}$	Los números naturales
$\mathbb{P}(A)$	Probabilidad de A
$X, Y, \ldots$	Variables aleatorias
M_n	Máximo de un conjunto de n variables aleatorias (pág. 6)
F	Función de distribución de la variable aleatoria X
$\overline{F}$	Cola de la función de distribución F (pág. 6)
$F^n(x)$	Función de distribución F elevada a la potencia n
ω_F	Punto final derecho de la función de distribución F (pág. 6)
$F^{\leftarrow}$	Función inversa generalizada de F (pág. 7)
F_u	Función de distribución de los excesos de X sobre el umbral u (pág. 25)
$\tilde{F}$	Función de distribución empírica de F (pág. 34)
$\hat{F}$	Función de distribución estimada de F (pág. 38)
$\mathcal{C}(F)$	Conjunto de puntos de continuidad de la función F (pág. 7)
$a_n \to a$	a_n converge a a cuando n tiende a infinito
$a_n \nrightarrow a$	a_n no converge a a cuando n tiende a infinito
$X_n \xrightarrow{\mathrm{d}} X$	X_n converge en distribución a X cuando n tiende a infinito (pág. 73)
$\lim\inf$	Límite inferior (pág. 74)
$\lim\sup$	Límite superior (pág. 74)
Φ_α	Función de distribución Fréchet estándar con parámetro α (pág. 8)
Ψ_α	Función de distribución Weibull estándar con parámetro α (pág. 8)

Λ	Función de distribución Gumbel estándar (pág. 8)
ϕ	Función de densidad Fréchet estándar con parámetro α (pág. 13)
ψ	Función de densidad Weibull estándar con parámetro α (pág. 13)
λ	Función de densidad Gumbel estándar (pág. 13)
$\mathcal{B}_{\mathbb{R}}$	Sigma-álgebra de Borel (pág. 81)
$F \in \mathcal{D}(G)$	F pertenece al dominio de atracción de G (pág. 16)
$F \in VR_\rho$	F es de variación regular en infinito con índice ρ (pág. 18)
$\mathcal{L}(x)$	Función de variación lenta en infinito (pág. 18)
G_ξ	Distribución generalizada de valores extremos estándar con parámetro ξ (pág. 22)
$G_{\xi,\mu,\sigma}$	Distribución generalizada de valores extremos con parámetros ξ, μ y σ (pág. 23)
H_ξ	Distribución generalizada de Pareto estándar con parámetro ξ (pág. 26)
$H_{u,\xi,\beta}$	Distribución generalizada de Pareto con parámetros u, ξ y β (pág. 28)
$E(X)$	Esperanza o media de la variable aleatoria X
$e(u)$	Función media de excesos sobre u (pág. 29)
$\mathbb{I}_A(x)$	Función indicadora del conjunto A
$\mathcal{F}$	Familia de distribuciones
$L(\theta)$	Función de verosimilitud (pág. 37)
$l(\theta)$	Función de log-verosimilitud (pág. 37)
$T(p)$	Periodo de retorno asociado a p (pág. 42)
x_p	Nivel de retorno asociado a la probabilidad de excedencia p (pág. 42)
x_T	Nivel de retorno asociado al tiempo T (pág. 46)
$f \sim g$	f y g son asintóticamente equivalentes (pág. 17)
$\square$	Fin de una demostración

Introducción

Actualmente existe un constante interés por el desarrollo sostenible de los recursos naturales, económicos y culturales de las zonas costeras. Para ello se han realizado diversos tipos de análisis de riesgo en dichas zonas, ya que éstas se encuentran amenazadas por una amplia gama de fenómenos de varios tipos, entre ellos los eventos hidrometeorológicos naturales, especialmente aquellos considerados como 'extremos' debido al alto impacto que tienen sobre las costas y sus recursos.

Hoy en día existe una extensa literatura disponible dedicada al análisis de riesgos costeros, en particular sobre el riesgo de inundación y erosión como resultado directo de la ocurrencia de fenómenos tales como la altura de la marea y el impacto del viento. Sin embargo, según [23], en México el análisis de este tipo de riesgos apenas ha comenzado en años recientes a pesar de que nuestro país cuenta con más de 11 mil kilómetros de costa, la cual presenta diversos escenarios al estar expuesta a amenazas provenientes del Océano Pacífico, del Golfo de México y del Mar Caribe, que han determinado importantes modificaciones y afectaciones a los paisajes geográficos, ecosistemas, asentamientos humanos e infraestructura.

Es relevante y necesario tomar medidas de planificación, evaluación y análisis de riesgos costeros con una perspectiva a largo plazo, esto es, la caracterización del comportamiento de fenómenos naturales a largo plazo con el fin de optimizar la toma de decisiones y gestión de áreas susceptibles a riesgos hidrometeorológicos con el fin de evitar pérdidas humanas, minimizar las pérdidas económicas y el daño, y mejorar las medidas preventivas ante desastres naturales.

Para contribuir con este tipo de análisis se busca utilizar herramientas probabilísticas y estadísticas que respondan una variedad de preguntas de investigación tales como –¿Qué al-

tura de marea se espera que se exceda con probabilidad 1/100 en un año?– o –¿Cuál es la probabilidad de que ocurra una cierta velocidad de viento sobre un nivel dado en algún lugar en cierto año?–. La Teoría de Valores Extremos se ha utilizado para responder ese tipo de preguntas, las cuales se relacionan con la distribución de eventos extremos. Actualmente exisiten técnicas y resultados que se centran en el estudio de eventos máximos entre un grupo y excedencias sobre umbrales altos.

El objetivo de la presente tesis es mostrar los principales resultados de la Teoría de Valores Extremos y su aplicación en herramientas estadísticas para realizar un análisis de los eventos hidrometeorológicos (altura de ola y velocidad de viento) extremos cerca de la costa de Campeche. El Instituto de Ingeniería de la UNAM proporcionó los datos utilizados para este trabajo mediante el Atlas de Clima Marítimo de la Vertiente Atlántica Mexicana y contaba con un análisis previo con el que se compararon los resultados obtenidos. La información consiste en una base de datos que contiene registros horarios de la altura de ola y velocidad de viento del primero de enero de 1948 al 31 de diciembre de 2010 de una zona del Golfo de México cercana a la costa de la Isla del Carmen. La elección de estos datos en específico se hizo con el fin de comparar el modelo que se tenía ajustado previamente en la zona antes mencionada, y saber si se pudo obtener un mejor ajuste o no.

Esta tesis consta de tres capítulos, en el primero de los cuales se enuncian los principales resultados de la Teoría de Valores Extremos utilizando principalmente [7], [19] y [14]. En el segundo capítulo se muestran las herramientas utilizadas para la estimación de los parámetros de cada modelo usando [5] como referencia principal. En el tercer capítulo se muestra la aplicación de la teoría mostrada en el capítulo 1, junto con las herramientas mencionadas en el capítulo 2, haciendo el análisis de los eventos hidrometeorológicos altura de ola y velocidad de viento para la zona mencionada así como una comparación con los modelos que se tenían previamente realizados. Posteriormente se encuentran las conclusiones a las que se llegaron tras hacer el ajuste de los modelos y al hacer el comparativo con el modelo anterior, seguidas por un apéndice donde se encuentran algunos resultados y demostraciones utilizadas a lo largo del trabajo.

Las estimaciones, así como las gráficas fueron realizadas un entorno de programación para análisis estadístico y gráfico llamado *R*, el cual es un proyecto de software libre que

puede descargarse en http://cran.r-project.org/. Los paquetes estadísticos utilizados fueron *evd, evir, ismev, POT, stats y ADGofTest*, así como funciones propias programadas en el mismo entorno. En el apéndice se incluye también el código utilizado para el análisis, con el fin de que el proceso realizado sea repetible en otras zonas costeras.

Capítulo 1

Teoría de Valores Extremos

La Teoría de Valores Extremos es una parte de la probabilidad que surge de la necesidad de modelar aquellas observaciones que se separan mucho de la media, y que tienen baja probabilidad de ocurrencia. Es decir, se concentra en responder preguntas probabilísticas y estadísticas sobre valores muy altos o muy bajos en sucesiones de variables aleatorias. Actualmente existe una extensa literatura en materia de Valores Extremos, los cuales tienen una amplia variedad de aplicaciones en distintas áreas tales como las finanzas, los seguros, el análisis hidrometeorológico, entre otras.

El objetivo de este capítulo es presentar una breve introducción a esta teoría, así como mostrar los principales resultados en los que se basa la metodología estadística que se utiliza para modelar los eventos extremos. Las principales fuentes utilizadas en esta sección fueron [7], [19] y [14].

1.1. Métodos de Análisis de Valores Extremos

En la teoría de Valores Extremos existen dos formas diferentes de catalogar a una observación o variable como extremo. Una de ellas es el modelo de Máximos por Bloque, y la otra es utilizando el modelo de Picos o Excedencias sobre un Umbral (POT por sus siglas en inglés de *Peaks Over Thresholds*). El modelo de Máximos por Bloque consiste en registrar el mayor valor en un intervalo dado de tiempo, por ejemplo un año o mes, mientras que en el modelo de Picos sobre un Umbral se registran todos aquellos valores que se encuentren por encima de un nivel o umbral dado. Una ejemplificación del registro de valores extremos de

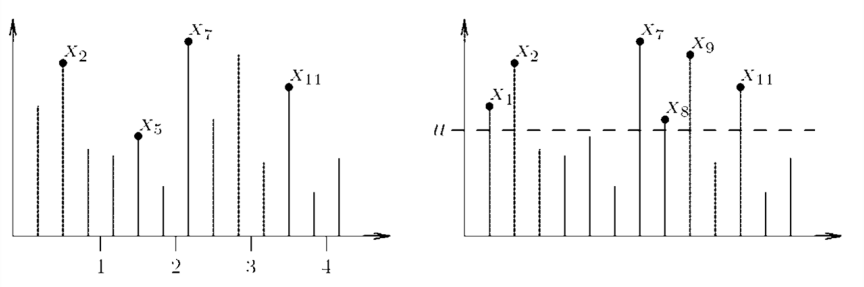

Figura 1.1: Registro de Máximos por Bloque (izquierda) y registro de Excedencias sobre un umbral (derecha).

cada uno de los dos modelos se muestra en la Figura 1.1 (tomada de [9]).

1.2. Máximos por Bloque

El método de observación de Máximos por Bloque consiste en tomar el máximo valor de un conjunto de variables aleatorias independientes e idénticamente distribuidas $X_1, X_2, \ldots, X_n$ con función de distribución común F no degenerada. Denotamos a este máximo como

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

Definida de ese modo, la distribución de M_n está dada por

$$\mathbb{P}(M_n \leq x) = \mathbb{P}(X_1 \leq x, \ldots, X_n \leq x) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq x) = F^n(x).$$

Donde la notación $F^n(x)$ denota a $(F(x))^n$. Como los eventos extremos suceden con poca probabilidad, digamos "cerca" del extremo superior del soporte de la distribución, intuitivamente vemos que el comportamiento asintótico de M_n esá relacionado con F en su cola derecha $\overline{F} = 1 - F$, cerca del Punto Final Derecho, el cual se define como

$$\omega_F = \sup(x \in \mathbb{R} : F(x) < 1).$$

Tenemos entonces que, para todo $x < \omega_F$,

$$\mathbb{P}(M_n \leq x) = F^n(x) \to 0,\ n \to \infty,$$

y para el caso en el que $\omega_F < \infty$ tenemos que, para $x \geq \omega_F$,

$$\mathbb{P}(M_n \leq x) = F^n(x) = 1.$$

Tenemos entonces que $M_n \xrightarrow{P} \omega_F$ cuando $n \to \infty$. Además puede demostrarse que M_n converge casi seguramente a ω_F, lo cual no aporta información suficiente acerca de la distribución límite de M_n pues se trata de una distribución degenerada por lo que se necesita recurrir a una normalización adecuada, algo similar a la forma en la que se procede usando el Teorema Central del Límite con el fin de encontrar una función de distribución no degenerada a la cual converja M_n normalizada. Buscamos entonces probabilidades de la forma:

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

cuyo límite o distribución límite exista cuando $n \to \infty$.

Denotaremos como $\mathcal{C}(F)$ para cualquier función monótona F, al conjunto de puntos de continuidad de ésta, y a lo largo de este capítulo se hará referencia a la convergencia débil o en distribución, la cual ocurre cuando una secuencia de funciones de distribución $\{F_n, n \geq 1\}$ converge a F_0 cuando $n \to \infty$, es decir

$$F_n(x) \to F_0(x)$$

para todo $x \in \mathcal{C}(F_0)$.

Las funciones inversas generalizadas, también llamadas funciones cuantil, son otra herramienta importante utilizada a lo largo de este trabajo.

DEFINICIÓN 1.1. *Supongamos que F es una función no decreciente en $\mathbb{R}$. Con la convención de que el ínfimo de un conjunto vacío es $+\infty$ se define a la* ***inversa generalizada de*** *F o* ***función cuantil de*** *F como*

$$F^{\leftarrow}(y) := \text{ínf}\{s : F(s) \geq y\}.$$

En el Apéndice A, sección A.2 se pueden encontrar algunos resultados relacionados con las funciones inversas o cuantiles, los cuales se utilizarán más adelante. Lo siguiente es definir una relación importante entre dos funciones de distribución, que utilizaremos en la enunciación de algunos resultados.

DEFINICIÓN 1.2. *Dos distribuciones son del mismo tipo o pertenecen a la misma familia si para algunas constantes $a > 0$, $b \in \mathbb{R}$,*

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

Alexandre Khintchine propuso un teorema conocido como El Teorema de Convergencia a Tipos (Ver Teorema A.3), el cual nos dice que si una sucesión de funciones de distribución converge normalizada a un cierto tipo de distribución, entonces el uso de otras constantes de normalización diferentes no cambia el tipo de la distribución límite y, además, nos da una relación entre las constantes utlizadas. Usaremos este resultado para la demostración del Teorema siguiente.

1.2.1. Distribuciones de Valores Extremos (DVE)

Uno de los resultados centrales de la Teoría de Valores Extremos es el Teorema de Fisher-Tippet y Gnedenko, pues en él se muestran las posibles formas de distribución límite para el máximo M_n con una normalización adecuada. A continuación se enuncia y demuestra dicho teorema, siguiendo [7] y [19].

Teorema 1.1. (Fisher-Tippet, Gnedenko: Distribuciones límite para máximos)[1]. *Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes e idénticamente distribuidas, y sea $M_n = max\{X_1, X_2, \ldots, X_n\}$. Si existen sucesiones $\{a_n\}$, $a_n > 0 \forall n$ y $\{b_n\}$, $b_n \in \mathbb{R}$ y alguna función de distribución no degenerada G, tales que:*

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = F^n(a_n x + b_n) \to G(x), \quad \forall \ x \in \mathcal{C}(G) \tag{1.1}$$

entonces G es del tipo de una de las tres funciones de distribución siguientes:

$$\textit{Fréchet:} \quad \Phi_\alpha(x) = \begin{cases} 0, & x \leq 0 \\ \exp\{-x^{-\alpha}\}, & x > 0 \end{cases} \qquad \alpha > 0.$$

$$\textit{Weibull:} \quad \Psi_\alpha(x) = \begin{cases} \exp\{-(-x)^{\alpha}\}, & x \leq 0 \\ 1, & x > 0 \end{cases} \qquad \alpha > 0.$$

$$\textit{Gumbel:} \quad \Lambda(x) = \exp\{-e^{-x}\}, \quad x \in \mathbb{R}\,.$$

Demostración. A partir de (1.1), tenemos que, $\forall\, t > 0$:

$$F^{[nt]}(a_{[nt]} x + b_{[nt]}) \to G(x),$$

[1] Este teorema fue originalmente propuesto por Fisher y Tippett en 1928 y demostrado rigurosamente por Gnedenko en 1943 (Según [14]).

donde $[x] = \text{máx}\{n \in \mathbb{N} | n \leq x\}$, es decir la parte entera de x. Por otro lado, también tenemos que

$$F^{[nt]}(a_n x + b_n) = (F^n(a_n x + b_n))^{\frac{[nt]}{n}} \to G^t(x).$$

Entonces, por el Teorema de Convergencia a Tipos (Ver Teorema A.3) G y G^t son del mismo tipo, existen $A(t) > 0$ y $B(t) \in \mathbb{R}$ tales que

$$\lim_{n\to\infty} \frac{a_n}{a_{[nt]}} = A(t) \quad y \quad \lim_{n\to\infty} \frac{b_n - b_{[nt]}}{a_{[nt]}} = B(t) \tag{1.2}$$

para $t > 0$ y además

$$G^t(x) = G\left(A(t)x + B(t)\right). \tag{1.3}$$

Si definimos $h_n(t) = a_{[nt]}$ para cada n, obtenemos una sucesión de funciones cuyo rango es el conjunto discreto $\{a_i, i \geq 1\}$. Suponiendo que los valores a_i son distintos tenemos que

$$h_n^{-1}(\{a_i\}) = \{t : a_{[nt]} = a_i\} = \left[\frac{i}{n}, \frac{i+1}{n}\right), \tag{1.4}$$

y en caso de haber valores repetidos de a_i, este conjunto es

$$h_n^{-1}(\{a_i\}) = \bigcup_{j:a_j=a_i} \left[\frac{j}{n}, \frac{j+1}{n}\right). \tag{1.5}$$

Con la definición y resultados sobre medibilidad que pueden encontrarse en el Apéndice A, sección A.4, observamos que al ser una unión a lo más numerable de intervalos, tenemos que los conjuntos en (1.4) y (1.5) pertenecen a $\mathcal{B}_{\mathbb{R}}$, la sigma-álgebra de Borel (ver definición A.9) y por lo tanto las funciones h_n son Borel medibles. Como la suma y el cociente de funciones medibles, así como el límite de sucesiones medibles es medible, se verifica con las relaciones en (1.2) la medibilidad de la función $A(\cdot)$ y la de $B(\cdot)$. Esta propiedad se utilizará más adelante en la demosración.

Ahora, utilizando (1.3), para $t > 0$ y $s > 0$ tenemos que

$$G^{ts}(x) = G\left(A(ts)x + B(ts)\right), \tag{1.6}$$

y por otro lado,

$$\begin{aligned} G^{ts}(x) = \left(G^s(x)\right)^t &= G^t\left(A(s)x + B(s)\right) \\ &= G\left(A(t)(A(s)x + B(s)) + B(t)\right) \\ &= G\left(A(t)A(s)x + A(t)B(s) + B(t)\right) \end{aligned} \tag{1.7}$$

Sabemos que si F es una función de distribución no degenerada, y para $a > 0$, $c > 0$, $b \in \mathbb{R}$, $c \in \mathbb{R}$ se cumple que $F(ax+b) = F(cx+d)$ para todo $x \in \mathbb{R}$, entonces $a = c$ y $b = d$. Usando este hecho para G que es no degenerada y las relaciones en (1.6) y (1.7), tenemos que

$$A(ts) = A(t)A(s) \tag{1.8}$$

$$B(ts) = A(t)B(s) + B(t) = A(s)B(t) + B(s). \tag{1.9}$$

Donde la igualdad de la derecha en (1.9) se da pues $B(ts) = B(st)$. Ya que $A(\cdot)$ es una función medible, tenemos entonces que la solución a la ecuación funcional para (1.8) (ver Teorema A.4) es de la forma

$$A(t) = t^{-\theta}, \qquad \theta \in \mathbb{R}\,.$$

Consideramos ahora tres casos: (i) $\theta = 0$, (ii) $\theta > 0$, (iii) $\theta < 0$.

Caso (i) $\theta = 0$. En este caso tenemos que $A(t) = 1$ para toda $t \in \mathbb{R}$, así que la ecuación (1.9) se convierte en

$$B(ts) = B(t) + B(s),$$

la cual es una variante de (1.8) (Teorema A.4), y por la medibilidad de $B(\cdot)$, la solución es de la forma

$$B(t) = -c \ln t, \quad t > 0, \quad c \in \mathbb{R}$$

por lo que (1.3) se vuelve

$$G^t(x) = G(x - c \ln t). \tag{1.10}$$

Observamos que $c \neq 0$, pues si c fuera 0, tendríamos que $G^t(x) = G(x)$, lo cual significaría que G es degenerada. También que para x fijo, $G^t(x)$ es no-creciente en t, por lo que $c > 0$, pues de otro modo, el lado derecho de (1.10) seria creciente en t.

Tenemos también que $G(x) \in (0,1)$ para toda $x \in \mathbb{R}$ pues si existiera x_0 tal que $G(x_0) = 0$ entonces por (1.10) tendríamos que

$$0 = G(x_0 - c \ln t)$$

para toda t, y mediante un cambio de variable, que $G(u) = 0$ para toda $u \in \mathbb{R}$, lo cual es falso. De manera similar, si suponemos que $G(x_0) = 1$ para algún x_0, llegamos a que $G(u) = 1$ para toda u, lo cual es una contradicción.

Ahora, con $x = 0$ en (1.10) obtenemos

$$G^t(0) = G(-c \ln t) \quad t > 0. \tag{1.11}$$

Escribimos ahora a $G(0) \in (0, 1)$ como $G(0) = \exp\{-e^{-p}\}$ y hacemos el cambio de variable $u = -c \ln t$. Como $t > 0$, entonces el rango de u es $(-\infty, \infty)$ y aplicando los cambios de variable en (1.11) tenemos que

$$\begin{aligned} G(u) &= G^t(0) = \left(\exp\left\{-e^{-p}\right\}\right)^t = \exp\left\{-te^{-p}\right\} \\ &= \exp\left\{-e^{-p}e^{-u/c}\right\} = \exp\left\{-e^{-(p+u/c)}\right\} \\ &= \Lambda(p + u/c). \end{aligned}$$

Caso (ii) $\theta > 0$. Tenemos que, para $\theta \neq 0$, por (1.9)

$$B(ts) = A(t)B(s) + B(t) = A(s)B(t) + B(s) = B(st),$$

por lo que, para $t \neq 1$ y $s \neq 1$ tenemos

$$\frac{B(s)}{1 - A(s)} = \frac{B(t)}{1 - A(t)},$$

es decir, la función $B(t)/(1 - A(t))$ es igual a una constante c. Entonces, para $t \neq 1$,

$$B(t) = \frac{B(s)}{1 - A(s)}(1 - A(t)) = c(1 - t^{-\theta})$$

y (1.3) se vuelve

$$\begin{aligned} G^t(x) &= G(t^{-\theta}x + c(1 - t^{-\theta})) \\ &= G(t^{-\theta}(x - c) + c) \end{aligned}$$

es decir,

$$G^t(x + c) = G(t^{-\theta}x + c).$$

Sea $H(x) = G(x + c)$. Entonces, como H y G son del mismo tipo, es suficiente resolver para H, la cual es no-degenerada y cumple que

$$H^t(x) = H(t^{-\theta}x) \tag{1.12}$$

Evaluando en $x = 0$ tenemos que $H^t(0) = H(0)$ para toda t, lo cual significa que $H(0) = 0$, o bien $H(0) = 1$. Si $H(0) = 1$ entonces existe $x < 0$ tal que $H(x) < 1$ para el cual $H^t(x)$ es

decreciente en t pero $H(t^{-\theta}x)$ es creciente, por lo tanto $H(0) = 0$. Sustituyendo ahora $x = 1$ en (1.12), obtenemos

$$H^t(1) = H(t^{-\theta}). \tag{1.13}$$

Observamos que $H(1) \in (0, 1)$, pues si $H(1) = 0$, entonces $H(t^{-\theta}) = 0$ para toda t, es decir $H \equiv 0$, y si $H(1) = 1$, entonces $H(t^{-\theta}) = 1$ para toda t, es decir $H \equiv 1$, lo cual es falso.

Sea $\alpha = \theta^{-1}$ y escribimos a $H(1) \in (0, 1)$ como $H(1) = \exp\{-p^{-\alpha}\}$, y hacemos el cambio de variable $u = t^{-\theta}$. Como $t > 0$, entonces el rango de u es $(0, \infty)$ y aplicando los cambios de variable en (1.13) tenemos que

$$\begin{aligned} H(u) &= H^t(1) = \left(\exp\left\{-p^{-\alpha}\right\}\right)^t = \exp\left\{-p^{-\alpha}t\right\} \\ &= \exp\left\{-p^{-\alpha}u^{-\alpha}\right\} = \exp\left\{-(pu)^{-\alpha}\right\} \\ &= \Phi_\alpha(pu). \end{aligned}$$

Caso (iii) $\theta < 0$. Del mismo modo que en (ii), llegamos a (1.12) y a que $H(0) = 0$ o bien $H(0) = 1$. Si $H(0) = 0$, entonces existe $x > 0$ tal que $H(x) < 1$ para el cual $H^t(x)$ es decreciente en t pero $H(t^{-\theta}x)$ es creciente. Por lo tanto $H(0) = 1$. Y sustituyendo $x = -1$ en (1.12) tenemos que

$$H^t(-1) = H(-t^{-\theta}) \tag{1.14}$$

Observamos también que $H(-1) \in (0, 1)$, pues si $H(-1) = 0$ entonces $H \equiv 0$, y si $H(-1) = 1$ entonces $H \equiv 1$, lo cual contradice la hipótesis de que H es no degenerada.

Sea $\alpha = -\theta^{-1}$, y $H(-1) = \exp\{-p^{\alpha}\}$, y hacemos el cambio de variable $u = -t^{-\theta}$. Como $t > 0$, el rango de u es $(-\infty, 0)$ y aplicando los cambios de variable en (1.14) tenemos que

$$\begin{aligned} H(u) &= \left(\exp\{-p^{\alpha}\}\right)^t = \exp\{-p^{\alpha}t\} \\ &= \exp\{-p^{\alpha}(-u)^{\alpha}\} = \exp\{-(-pu)^{\alpha}\} \\ &= \Psi_\alpha(pu). \end{aligned}$$

□

DEFINICIÓN 1.3. *Las funciones de distribución Λ, Φ_α y Ψ_α que aparecen en el Teorema de Fisher-Tippet y Gnedenko son llamadas Distribuciones de Valores Extremos Estándar. Las*

funciones de distribución del mismo tipo que Λ, Φ_α *y* Ψ_α *son llamadas Distribuciones de Valores Extremos (DVE).*

Hacemos la observación de que el Teorema de Fisher-Tippet y Gnedenko no garantiza la existencia de un límite no degenerado para cualquier función F, simplemente que cuando el límite existe, entonces se trata de una DVE. También notemos que como las DVE son continuas en $\mathbb{R}$ entonces $(M_n - b_n)/a_n \to G$ (débilmente o en distribución) es equivalente a

$$\lim_{n\to\infty} \mathbb{P}(M_n \leq a_n x + b_n) = \lim_{n\to\infty} F^n(a_n x + b_n) = G(x), \quad x \in \mathbb{R}\,.$$

Podemos dar una expersión para las densidades de las DVE en términos de sus funciones

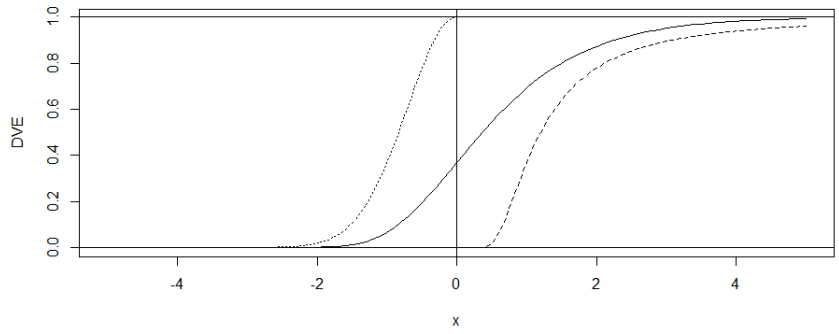

Figura 1.2: Distribuciones Gumbel (línea continua), Fréchet (guiones) y Weibull (puntos) con $\alpha = 2$ en estas dos últimas.

de distribución como sigue:

$$\text{Fréchet:} \quad \phi_\alpha(x) = \alpha\Phi_\alpha(x)x^{-(1+\alpha)} \qquad x \geq 0$$

$$\text{Weibull:} \quad \psi_\alpha(x) = \alpha\Psi_\alpha(x)(-x)^{\alpha-1} \qquad x \leq 0$$

$$\text{Gumbel:} \quad \lambda(x) = \Lambda(x)e^{-x} \qquad x \in \mathbb{R}$$

Las DVE se caracterizan por una propiedad llamada max-estabilidad que se define como sigue.

Definición 1.4. *Se dice que una función de distribución no degenerada* F *es max-estable si*

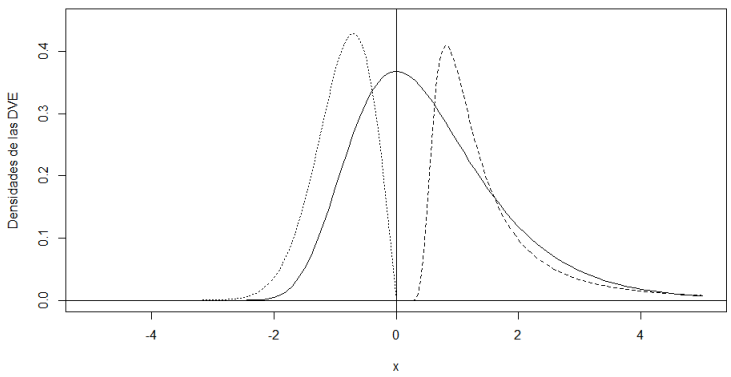

Figura 1.3: Densidades Gumbel (línea continua), Fréchet (guiones) y Weibull (puntos) con $\alpha = 2$ en estas dos últimas.

para cada n se puede hacer una adecuada elección de constantes $a_n > 0$ y $b_n \in \mathbb{R}$ tales que

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

O bien, se dice que una distribución es max-estable si la distribución de su máximo M_n para variables aleatorias independientes e idénticamente distribuidas (F^n) cumple con la propiedad anterior.

La caracterización de las DVE por la max-estabilidad se muestra en el siguiente resultado.

Teorema 1.2. *La clase de las distribuciones max-estables coincide con la clase de las posibles distribuciones límite no degeneradas para máximos, es decir, las DVE.*

Demostración. Si F es una distribución max-estable, entonces $F^n(a_n x + b_n) = F(x)$ para cada n y constantes normalizantes adecuadas. Concluimos entonces que todas las distribuciones max-estables son distribuciones límite para máximos de variables aleatorias independientes e idénticamente distribuidas. Y el Teorema de Fisher-Tippet y Gnedenko nos dice que las únicas distribuciones límite posibles para máximos son las DVE. Por tanto, sólo resta demostrar que dichas distribuciones límite son max-estables. Supongamos entonces que para constantes adecuadas, tenemos que

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G(x), \quad x \in \mathbb{R}\,.$$

Sabemos por el Teorema de Fisher-Tippet y Gnedenko que las posibles funciones de distri-

bución límite son continuas, por lo que para toda $k \in \mathbb{N}$,

$$\lim_{n\to\infty} F^{nk}(a_n x + b_n) = \Big(\lim_{n\to\infty} F^n(a_n x + b_n)\Big)^k = G^k(x), \quad x \in \mathbb{R}$$

y por otro lado tenemos que

$$\lim_{n\to\infty} F^{nk}(a_{nk} x + b_{nk}) = G(x), \quad x \in \mathbb{R}\,.$$

Ahora, por el Teorema de Convergencia a Tipos tenemos que existen constantes $\tilde{a}_k > 0$ y $\tilde{b}_k \in \mathbb{R}$ tales que

$$\lim_{n\to\infty} \frac{a_{nk}}{a_n} = \tilde{a}_k \quad y \quad \lim_{n\to\infty} \frac{b_{nk} - b_n}{a_n} = \tilde{b}_k$$

y

$$G^k(\tilde{a}_k x + \tilde{b}_k) = G(x)$$

para cada k. Por lo tanto, G es una distribución max-estable. □

Ejemplificaremos ahora la propiedad max-estable de las DVE Estándar.

EJEMPLO 1.1. *Distribución Fréchet. Sea Φ_α la función de distribución Fréchet, entonces*

$$\begin{aligned} \Phi_\alpha^n(x) &= \left(\exp\left\{-x^{-\alpha}\right\}\right)^n = \exp\left\{-x^{-\alpha} n\right\} \\ &= \exp\left\{-(xn^{-1/\alpha})^{-\alpha}\right\} \\ &= \Phi_\alpha(n^{-1/\alpha} x) \end{aligned}$$

Entonces, la distrubición Fréchet es max-estable.

EJEMPLO 1.2. *Distribución Weibull. Tomamos a la función de distribución Weibull Ψ_α, para la cual se tiene que*

$$\begin{aligned} \Psi_\alpha^n(x) &= \left(\exp\left\{-(-x)^{\alpha}\right\}\right)^n = \exp\left\{-(-x)^{\alpha} n\right\} \\ &= \exp\left\{-(-xn)^{1/\alpha})^{\alpha}\right\} \\ &= \Psi_\alpha(n^{1/\alpha} x) \end{aligned}$$

Lo cual muestra que la distribución Weibull también es max-estable.

EJEMPLO 1.3. *Distribución Gumbel. Tenemos para la función de distribución Gumbel Λ que*

$$\begin{aligned} \Lambda^n(x) &= \left(\exp\left\{-e^{-x}\right\}\right)^n = \exp\left\{-e^{-x} n\right\} \\ &= \exp\left\{-e^{-x+\ln n}\right\} \\ &= \Lambda(x - \ln n) \end{aligned}$$

Por lo que la distribución Gumbel es max-estable.

Con estos ejemplos podemos dar una forma explícita a las constantes de normalización a_n y b_n necesarias para que se cumpla el resultado de Fisher y Tippet, para el caso de las DVE estándar, las cuales son

$$\text{Fréchet:} \quad a_n = n^{1/\alpha} \qquad b_n = 0$$

$$\text{Weibull:} \quad a_n = n^{-1/\alpha} \quad b_n = 0$$

$$\text{Gumbel:} \quad a_n = 1 \qquad b_n = \ln n$$

Cada una de las DVE estándar representa una familia de distribuciones que son del mismo tipo, a las que llamamos Distribuciones de Valores Extremos, las cuales podemos escribir con los parámetros de ubicación μ y escala σ como sigue

$$\text{Fréchet:} \quad \Phi_\alpha(x) = \begin{cases} 0, & x \leq \mu \\ \exp\{-((x-\mu)/\sigma)^{-\alpha}\}, & x > \mu \end{cases} \quad \alpha > 0.$$

$$\text{Weibull:} \quad \Psi_\alpha(x) = \begin{cases} \exp\{-(-(x-\mu)/\sigma)^{\alpha}\}, & x \leq \mu \\ 1, & x > \mu \end{cases} \quad \alpha > 0.$$

$$\text{Gumbel:} \quad \Lambda(x) = \exp\{-e^{-(x-\mu)/\sigma}\}, \quad x \in \mathbb{R}\,.$$

1.2.2. Dominios de Atracción de las DVE

El Teorema de Fisher-Tippet y Gnedenko nos proporciona las posibles distribuciones límite de máximos debidamente normalizados en caso de existir. Por otro lado, podemos preguntarnos qué condiciones debe cumplir una función de distribución F para que el máximo normalizado M_n converja débilmente a una DVE G dada, así como la manera de elegir a las constantes de normalización adecuadas.

DEFINICIÓN 1.5. *Decimos que una función de distribución F pertenece al* ***Dominio de Atracción*** *del Máximo de la distribución de valores extremos G si existen constantes $a_n > 0$ y $b_n \in \mathbb{R}$ tales que*

$$F^n(a_n x + b_n) = \mathbb{P}(M_n \leq a_n x + b_n) \to G(x), \quad x \in \mathbb{R}\,. \tag{1.15}$$

Y se denota por $F \in \mathcal{D}(G)$.

Directamente de esta definición podemos dar otro criterio para determinar los dominios de atracción mediante la siguiente proposición.

Proposición 1.1. *Sea F función de distribución. Entonces $F \in \mathcal{D}(G)$ con constantes de normalización $a_n > 0$ y $b_n \in \mathbb{R}$ si y sólo si*

$$\lim_{n\to\infty} n\,\overline{F}(a_n x + b_n) = -\ln G(x) \qquad x \in \mathbb{R}\,. \tag{1.16}$$

Donde $\overline{F}(x) = 1 - F(x)$. Cuando $G(x) = 0$ el límite se interpreta como ∞.

Demostración. Supongamos que $G(x) > 0$. Entonces, sacando logaritmos en (1.15) tenemos que

$$-n \ln\big(1 - \overline{F}(a_n x + b_n)\big) \to -\ln G(x).$$

Y usaremos que $-\ln(1-z) \sim z$ cuando $z \to 0$ es decir, que $\lim_{z\to 0} -\ln(1-z)/z = 1$ para obtener

$$n\,\overline{F}(a_n x + b_n) \to -\ln G(x).$$

Lo cual se da pues $\overline{F}(a_n x + b_n) \to 0$ cuando $n \to \infty$. De otro modo existiría una subsucesión con índices $\{n_k\}$ y $x_0 \in \mathbb{R}$ tal que $\overline{F}(a_{n_k} x_0 + b_{n_k}) \to \xi > 0$, de modo que $F^{n_k}(a_{n_k} x_0 + b_{n_k}) = \big(1 - \overline{F}(a_{n_k} x_0 + b_{n_k})\big)^{n_k} \to 0$, que es una contradicción, pues supusimos que $F^n(a_n x + b_n) \to G(x) > 0$.

Para el recíproco, supongamos (1.16). Tenemos entonces que

$$\begin{aligned} F^n(a_n x + b_n) &= \big(1 - \overline{F}(a_n x + b_n)\big)^n \\ &= \left(1 - \frac{n\,\overline{F}(a_n x + b_n)}{n}\right)^n \to \exp\{-(-\ln G(x)\} = G(x). \end{aligned}$$

Para $G(x) = 0$ supongamos que $F^n(a_n x + b_n) \to 0$ pero que $n\overline{F}(a_n x + b_n) \nrightarrow \infty$, entonces existe una subsucesión n_k para la cual $n_k \overline{F}(a_{n_k} x + b_{n_k}) \to H(x) < \infty$ para alguna función H, lo cual implica que $F^{n_k}(a_{n_k} x + b_{n_k}) \to \exp\{-H(x)\} > 0$ que es una contradicción. De manera similar se prueba el recíproco. Supongamos (1.16) interpretando el límite como ∞, pero que $F^n(a_n x + b_n) \nrightarrow 0$, esto implica que existe una subsucesión para la cual $F^{n_k}(a_{n_k} x + b_{n_k}) \to H(x) > 0$ para alguna función H pero esto implica que $n_k\,\overline{F}(a_{n_k} x + b_{n_k}) \to -\ln H(x) < \infty$, lo cual es una contradicción.

□

Para caracterizar a los dominios de atracción de las DVE, introduciremos a las funciones de variación regular. Las funciones de variación regular son aquellas que se comportan asintóticamente como funciones de potencia.

DEFINICIÓN 1.6. *Una función medible $F : \mathbb{R}_+ \to \mathbb{R}_+$ es de variación regular en ∞ con índice ρ si para $x > 0$*

$$\lim_{t \to \infty} \frac{F(tx)}{F(t)} = x^\rho.$$

Y se denota por $F \in VR_\rho$.

A ρ se le conoce como el exponente de variación. Si $\rho = 0$ decimos que F es de variación lenta y dichas funciones generalmente se denotan por $\mathcal{L}(x)$. Si $F \in VR_\rho$, entonces $F(x)/x^\rho \in VR_0$, y podemos escribir $\mathcal{L}(x) = F(x)/x^\rho$. Notemos entonces que siempre es posible representar una función de variación regular con índice ρ como $x^\rho \mathcal{L}(x)$.

Dominio de Atracción de la Distribución Fréchet $\Phi_\alpha(x) = \exp\{-x^{-\alpha}\}, \ \ x > 0$

El Dominio de Atracción de $\Phi_\alpha(x)$ está relacionado con aquellas funciones cuya cola derecha es de variación regular. Notemos que la cola de la distribución Fréchet

$$1 - \Phi_\alpha(x) \sim x^{-\alpha}, \quad x \to \infty$$

es decir, $\lim_{x \to \infty}(1 - \Phi_\alpha(x))/x^{-\alpha} = 1$. Esto indica que la cola de Φ_α decrece como una función potencia.

Proposición 1.2. *$F \in \mathcal{D}(\Phi_\alpha)$ si y sólo si $\overline{F} \in VR_{-\alpha}$. En este caso*

$$F^n(a_n x) \to \Phi_\alpha(x) \tag{1.17}$$

con

$$a_n = F^{\leftarrow}\left(1 - \frac{1}{n}\right). \tag{1.18}$$

Demostración. Primero observemos que sólo distribuciones con extremo derecho infinito pueden pertenecer al dominio de atracción de Fréchet, pues si $\omega_F < \infty$, entonces podemos encontrar una $x_0 > 0$ tal que $a_n x_0 > \omega_F$ para toda n, lo cual contradice (1.17). Además, $a_n \to \infty$, pues de otro modo existiría $K < \infty$ tal que para alguna subsucesión con índices $\{n_k\}$, $a_{n_k} \leq K$ y

$$0 < \Phi_\alpha(1) = \lim_{n_k \to \infty} F^{n_k}(a_{n_k}) \leq \lim_{n_k \to \infty} F^{n_k}(K) = 0,$$

pues $F(K) < 1$, lo cual es una contradicción.

Supongamos que $F \in \mathcal{D}(\Phi_\alpha)$, es decir que para constantes adecuadas $a_n > 0$ y $b_n \in \mathbb{R}$ se tiene que

$$F^n(a_n x + b_n) \to \Phi_\alpha(x) \tag{1.19}$$

para $x > 0$ y $\alpha > 0$. Supongamos primero que $b_n = 0$, entonces por la Proposición 1.1 podemos reescribir (1.19) como

$$n\,\overline{F}(a_n x) \to x^{-\alpha} \tag{1.20}$$

Como $a_n \to \infty$, para cada $t > 0$ existe $n(t) < \infty$ definido como

$$n(t) = \text{mín}\{m : a_{m+1} > t\}$$

entonces

$$a_{n(t)} \leq t < a_{n(t)+1}$$

que por la monotonía de $\overline{F}$ implica que

$$\left(\frac{n(t)}{n(t)+1}\right) \frac{(n(t)+1)\,\overline{F}(a_{n(t)+1}x)}{n(t)\,\overline{F}(a_{n(t)})} \leq \frac{\overline{F}(tx)}{\overline{F}(t)} \leq \left(\frac{n(t)+1}{n(t)}\right) \frac{n(t)\,\overline{F}(a_{n(t)}x)}{(n(t)+1)\,\overline{F}(a_{n(t)+1})}.$$

Hacemos la observación de que $n(t) \to \infty$ cuando $t \to \infty$, pues es una función monótona no decreciente con imagen en $\mathbb{N}$, y por ende que $n(t) \sim n(t)+1$. Entonces, tomando el límite cuando $t \to \infty$ y por (1.20), tenemos que

$$\frac{x^{-\alpha}}{1^{-\alpha}} \leq \liminf_{t\to\infty} \frac{\overline{F}(tx)}{\overline{F}(t)} \leq \limsup_{t\to\infty} \frac{\overline{F}(tx)}{\overline{F}(t)} \leq \frac{x^{-\alpha}}{1^{-\alpha}},$$

para $x > 0$, lo cual implica que $\overline{F} \in VR_{-\alpha}$.

Si $b_n \neq 0$, tenemos que demostrar que $b_n/a_n \to 0$ cuando $n \to \infty$, si esto se cumple, podemos reemplazar b_n por 0 en (1.19) y repetir el argumento anterior. Esta demostración puede encontrarse en [2], o en la Proposición 2.3.1 de [19] en la cual se da un argumento alternativo.

Para caracterizar a a_n, definimos a la función $U = 1/\overline{F}$, entonces reescribimos (1.20) como

$$\frac{U(a_n x)}{n} \to x^\alpha \qquad x > 0$$

que por la Proposición A.2 implica que

$$\frac{U^{\leftarrow}(ny)}{a_n} \to y^{1/\alpha}, \qquad y > 0$$

y con $y = 1$, tenemos que $U^{\leftarrow}(n) \sim a_n$ cuando $n \to \infty$, y observamos que

$$\begin{aligned} U^{\leftarrow}(n) = \left(\frac{1}{\overline{F}}\right)^{\leftarrow}(n) &= \text{ínf}\left\{s : \frac{1}{\overline{F}(s)} \geq n\right\} \\ &= \text{ínf}\left\{s : \overline{F}(s) \leq \frac{1}{n}\right\} \\ &= \text{ínf}\left\{s : F(s) \geq 1 - \frac{1}{n}\right\} \\ &= F^{\leftarrow}\left(1 - \frac{1}{n}\right). \end{aligned}$$

con lo cual queda definida a_n.

Recíprocamente, supongamos que $\overline{F} \in VR_{-\alpha}$, entonces si definimos a la función U como anteriormente, tenemos que $U \in VR_{\alpha}$. Definimos a a_n como en (1.18) y entonces, usando la Proposición A.1 (Apeéndice A) para $U^{\leftarrow}$, tenemos que $U^{\leftarrow}(n) \leq t$ si y sólo si $n \leq U(t)$ y que $t < U^{\leftarrow}(n)$ si y sólo si $U(t) < n$ y por lo tanto,

$$\frac{U(U^{\leftarrow}(n))}{U(U^{\leftarrow}(n)(1+\epsilon))} \leq \frac{U(U^{\leftarrow}(n))}{n} \leq \frac{U(U^{\leftarrow}(n))}{U(U^{\leftarrow}(n)(1-\epsilon))}$$

para $\epsilon > 0$. Tomando el límite cuando $n \to \infty$,

$$(1+\epsilon)^{-\alpha} \leq \liminf_{n\to\infty} \frac{U(U^{\leftarrow}(n))}{n} \leq \limsup_{n\to\infty} \frac{U(U^{\leftarrow}(n))}{n} \leq (1-\epsilon)^{-\alpha}$$

y como $\epsilon > 0$ es arbitrario, tenemos que $U(U^{\leftarrow}(n)) \sim n$ es decir, $U(a_n) \sim n$. Lo cual implica por como tomamos U que $\overline{F}(a_n) \sim n^{-1}$ cuando $n \to \infty$ por lo tanto $\overline{F}(a_n) \to 0$ dando que $a_n \to \infty$. Tenemos entonces que

$$n\,\overline{F}(a_n x) \sim U(a_n)\,\overline{F}(a_n x) = \frac{\overline{F}(a_n x)}{\overline{F}(a_n)} \to x^{-\alpha}$$

para $x > 0$ cuando $n \to \infty$, pues $F \in VR_{-\alpha}$. Para $x < 0$ tenemos que $\overline{F}(0) > 0$ y por tanto que $F(0) < 1$ porque así lo requiere la variación regular, por lo que $F^n(a_n x) \leq F^n(0) \to 0 = \Phi_\alpha(x)$. Entonces por la Proposición 1.1, $F \in \mathcal{D}(\Phi_\alpha)$. □

Dominio de Atracción de la Distribución Weibull $\Psi_\alpha(x) = \exp\{-(-x)^\alpha\}, \;\; x < 0$

El Dominio de Atracción de $\Psi_\alpha(x)$ está relacionado también con las funciones de variación regular.

Proposición 1.3. *$F \in \mathcal{D}(\Psi_\alpha)$ si y sólo si $\omega_F < \infty$ y $1 - F(\omega_F - x^{-1}) \in VR_{-\alpha}$. En este caso*

$$F^n(\omega_F + (\omega_F - \gamma_n)x) \to \Psi_\alpha(x), \quad x < 0$$

con

$$\gamma_n = F^{\leftarrow}\left(1 - \frac{1}{n}\right).$$

En este caso podemos observar que las constantes normalizantes que hacen que el Teorema de Fisher-Tippet y Gnedenko se cumpla son $a_n = \omega_F - \gamma_n$ y $b_n = \omega_F$.

Demostración. Supongamos que $\omega_F < \infty$, y que $1 - F(\omega_F - x^{-1}) \in VR_{-\alpha}$. Definimos a

$$F_\star(x) = \begin{cases} 0 & , \ x < 0 \\ F(\omega_F - x^{-1}) & , \ x \geq 0 \end{cases}$$

Entonces $1 - F_\star(x) \in VR_{-\alpha}$ y por la proposición 1.2, tomando $a_n = (1/(1 - F_\star))^{\leftarrow}(n)$ tenemos que

$$F_\star^n(a_n x) \to \Phi_\alpha(x), \quad x > 0$$

es decir,

$$F^n(\omega_F - (a_n x)^{-1}) \to \exp\left\{-x^{-\alpha}\right\}, \quad x > 0$$

y con el cambio de variable $y = x^{-1}$,

$$F^n(\omega_F + a_n^{-1} y) \to \exp\left\{-(-y)^{\alpha}\right\}, \quad y < 0.$$

Observamos que

$$\begin{aligned} a_n &= \text{ínf}\left\{s : \frac{1}{1 - F_\star(s)} \geq n\right\} \\ &= \text{ínf}\left\{s : \frac{1}{1 - F(\omega_F - s^{-1})} \geq n\right\} \\ &= \text{ínf}\left\{\frac{1}{\omega_F - s} : \frac{1}{1 - F(s)} \geq n\right\} \\ &= 1/\left(\omega_F - \text{ínf}\left\{s : 1/(1 - F(s)) \geq n\right\}\right) = 1/(\omega_F - \gamma_n). \end{aligned}$$

Por lo tanto

$$F^n(\omega_F + (\omega_F - \gamma_n) y) \to \Psi_\alpha(y), \quad y < 0.$$

Para la demostración del recíproco hacen falta más herramientas, y puede consultarse en [19], Proposición 1.13. □

Dominio de Atracción de la Distribución Gumbel $\Lambda(x) = \exp\{-e^{-x}\}$

Una forma de caracterizar a $\mathcal{D}(\Lambda)$ es mediante el siguiente resultado.

Proposición 1.4. *$F \in \mathcal{D}(\Lambda)$ si y sólo si existe una función $\tilde{a} > 0$ tal que*

$$\lim_{x \to \omega_F} \frac{\overline{F}(x + t\tilde{a}(x))}{\overline{F}(x)} = e^{-t}, \quad t \in \mathbb{R}, \tag{1.21}$$

donde una posible elección de $\tilde{a}$ es

$$\tilde{a}(x) = \int_x^{\omega_F} \frac{\overline{F}(t)}{\overline{F}(x)} dt, \quad x < \omega_F.$$

En este caso

$$F^n(a_n x + b_n) \to \Lambda(x), \quad x \in \mathbb{R}$$

donde una posible elección de las constantes de normalización es

$$b_n = F^{\leftarrow}\left(1 - \frac{1}{n}\right) \quad y \quad a_n = \tilde{a}(b_n).$$

Existen otras diferentes formas de caracterizar a los dominios de atracción de las DVE, tales como las condiciones de Von Mises, las cuales junto a la demostración de la Proposición 1.4, pueden consultarse en el capítulo 1 de [19].

1.2.3. La Distribución Generalizada de Valores Extremos (DGVE)

Los tres tipos de límite del Teorema de Fisher-Tippet y Gnedenko tienen distintas formas de comportamiento, correspondientes a los diferentes valores del parámetro de forma ξ, por lo que no es recomendable elegir una de las tres familias antes de hacer inferencia sobre los parámetros correspondientes.

Para evitar hacer una elección preliminar puede utilizarse la Distribución Generalizada de Valores Extremos (DGVE, o GVE por sus siglas en inglés), la cual es una representación que unifica los tres casos de DVE en una sola función de distribución. Esta representación fue propuesta por Jenkinson y Von Mises (según [7]), y tiene la ventaja de permitir hacer inferencia de los parámetros antes de hacer la elección de una de las tres familias de distribuciones de valores extremos.

DEFINICIÓN 1.7. **Distribución Generalizada de Valores Extremos.** *Se define a G_ξ como*

$$G_\xi(x) = \begin{cases} \exp\{-(1+\xi x)^{-1/\xi}\} & , \ \xi \neq 0 \\ \exp\{-e^{-x}\} & , \ \xi = 0 \end{cases} \tag{1.22}$$

si $1+\xi x>0$ *para* $\xi\neq 0$, $x\in\mathbb{R}$ *para* $\xi=0$ *y*

$$G_\xi(x)=\begin{cases}0 & ,\ \xi>0\\ 1 & ,\ \xi<0\end{cases}$$

en otro caso.

Tenemos entonces que el soporte de G_ξ *corresponde a*

$$\begin{array}{llll} x>-\xi^{-1} & para & \xi>0 & \textit{Caso Fréchet,}\\ x<-\xi^{-1} & para & \xi<0 & \textit{Caso Weibull,}\\ x\in\mathbb{R} & para & \xi=0 & \textit{Caso Gumbel.}\end{array}$$

Podemos observar que podemos llegar a la forma de Λ haciendo tender ξ a cero en la primera expresión de (1.22). G_ξ es llamada la DGVE estándar, y así como con cada una de las DVE, ésta representa una familia de distribuciones las cuales se pueden escribir incluyendo parámetros de localización μ y escala σ

$$G_{\xi,\mu,\sigma}(x)=\begin{cases}\exp\{-(1+\xi\left(\frac{x-\mu}{\sigma}\right))^{-1/\xi}\} & ,\ \xi\neq 0\\ \exp\{-e^{-(x-\mu)/\sigma}\} & ,\ \xi=0\end{cases}$$

La DGVE proporciona una representación conveniente, pues unifica las tres representaciones de las DVE y es principalmente utilizada para aplicaciones estadísticas. Para $\xi=0$ tenemos la misma expresión, esto es $\Lambda(x)=G_0(x)$.
Para $\xi>0$, con $\alpha=\xi^{-1}$

$$\begin{aligned} G_\xi(x) &= \exp\left\{-(1+x/\alpha)^{-\alpha}\right\}\\ &= \exp\left\{-\left(\frac{x+\alpha}{\alpha}\right)^{-\alpha}\right\}\\ &= \Phi_\alpha\left(\frac{x+\alpha}{\alpha}\right).\end{aligned}$$

Para $\xi<0$, con $\alpha=-\xi^{-1}$

$$\begin{aligned} G_\xi(x) &= \exp\left\{-(1-x/\alpha)^{\alpha}\right\}\\ &= \exp\left\{-\left(-\frac{x-\alpha}{\alpha}\right)^{\alpha}\right\}\\ &= \Psi_\alpha\left(\frac{x-\alpha}{\alpha}\right).\end{aligned}$$

Un resultado propuesto por Pickands, Balkema y de Haan [14] brinda una caracterización del dominio de atracción de la Distribución Generalizada de Valores Extremos y se enuncia a continuación.

Proposición 1.5. *Sea F una función de distribución. Para $\xi \in \mathbb{R}$, $F \in \mathcal{D}(G_\xi)$ si y sólo si existe una función $a(\cdot)$ positiva, medible tal que para $1 + \xi x > 0$,*

$$\lim_{u \uparrow \omega_F} \frac{\overline{F}(u + xa(u))}{\overline{F}(u)} = \begin{cases} (1+\xi x)^{-1/\xi} & , \ \xi \neq 0 \\ e^{-x} & , \ \xi = 0 \end{cases} \tag{1.23}$$

Demostración. Para el caso $\xi = 0$, se trata de la Proposición 1.4

Para $\xi > 0$ tenemos entonces que (1.23) es equivalente a $F \in \mathcal{D}(\Phi)$, lo cual implica que $\overline{F} \in VR_{-\alpha}$, con $\alpha = \xi^{-1}$ por la Proposición 1.2. Entonces por la Representación de Karamata para funciones de variación regular (Corolario A.2) $\overline{F}$ tiene la siguiente representación

$$\overline{F}(x) = c(x) \exp\left\{ \int_1^x \frac{\delta(t)}{t} dt \right\},$$

donde $c(x) \to c_0$ y $\delta(x) \to -\alpha$ cuando $x \to \infty$. O bien como

$$\overline{F}(x) = c(x) \exp\left\{ -\int_1^x \frac{1}{a(t)} dt \right\},$$

donde $a(x)/x \to \alpha^{-1}$, o bien $a(x) \sim \xi x$ cuanto $x \to \infty$. Entonces

$$\begin{aligned} \frac{\overline{F}(u + xa(u))}{\overline{F}(u)} &= \frac{c(u + xa(u))}{c(u)} \exp\left\{ -\int_u^{u+xa(u)} \frac{1}{a(t)} dt \right\} \\ &\sim \exp\left\{ \int_0^x \frac{a(u)}{a(u + va(u))} dv \right\} \end{aligned}$$

cuando $u \to \infty$ y haciendo el cambio de variable $t = u + va(u)$. Tenemos entonces que

$$\begin{aligned} \exp\left\{ \int_0^x \frac{a(u)}{a(u + va(u))} dv \right\} &\sim \exp\left\{ \int_0^x \frac{\xi u}{\xi(u + v\xi u)} dv \right\} \\ &= \exp\left\{ \int_0^x \frac{1}{1 + \xi v} dv \right\} \\ &= \exp\left\{ -\frac{1}{\xi} \ln(1 + \xi x) \right\} \\ &= (1 + \xi x)^{-1/\xi} \end{aligned}$$

con lo que llegamos a (1.23). Recíprocamente, si se cumple b), tomamos $a_n = (1/\overline{F})^{\leftarrow}(n)$ como en la demostración de la Proposición 1.2. Tenemos que $\overline{F}(a_n) \sim n^{-1}$, y que $a_n \to \omega_F$ cuando $n \to \infty$. Entonces

$$(1 + \xi x)^{-1/\xi} = \lim_{n \to \infty} \frac{\overline{F}(a_n + xa(a_n))}{\overline{F}(a_n)} = \lim_{n \to \infty} n\overline{F}(a_n + xa(a_n))$$

lo cual implica, por la Proposición 1.1 que $F \in \mathcal{D}(G_\xi)$. El caso para $\xi < 0$ puede demostrarse de manera similar.

□

Este resultado tiene una interpretación muy útil, que relaciona a la Distribución Generalizada de Valores Extremos con la Distribución Generalizada de Pareto, la cual se verá a continuación.

1.3. Picos sobre un Umbral

El modelo relacionado al método en el que se registran como valores extremos aquellos que excedan un nivel o umbral dado o sus excedencias es el modelo POT o Picos sobre un Umbral. Este modelo está relacionado con la Distribución Generalizada de Pareto, la cual se define más adelante.

Si consideramos una variable aleatoria X con función de distribución F decimos que ha ocurrido una excedencia si $X > u$, y el exceso sobre el umbral u es la cantidad por la cual X sobrepasa dicho umbral, es decir $X - u$. La función que nos resulta de interés es la función de distribución F_u de los valores por los que X exceda el umbral u. Esta función es conocida como la Función de Distribución de Excesos y se define como sigue.

DEFINICIÓN 1.8. *(**Función de Distrubución de Excesos**.)Sea X una variable aleatoria con función de distribución F y punto final derecho ω_F. Para $u < \omega_F$ fijo,*

$$F_u(y) := P(X - u \leq y | X > u), \quad 0 \leq y \leq \omega_F - u, \tag{1.24}$$

es la Función de Distribución de Excesos de la variable X sobre un umbral u, donde $y = x - u$ son los excesos.

Esta función también es llamada la Función de Distribución de Excesos Condicional, pues reduce el espacio muestral a aquellas variables que sobrepasen el nivel u, pero en adelante nos referiremos a ella como el nombre que se le dio en la Definición 1.8. Notemos que se es posible expresar a F_u en términos de F como sigue.

$$F_u(y) = \frac{F(y+u) - F(u)}{1 - F(u)} = \frac{F(x) - F(u)}{1 - F(u)}$$

Esta relación es de utilidad para hacer ajustes a la cola de una distribución mediante la estimación de la distribución de sus excesos.

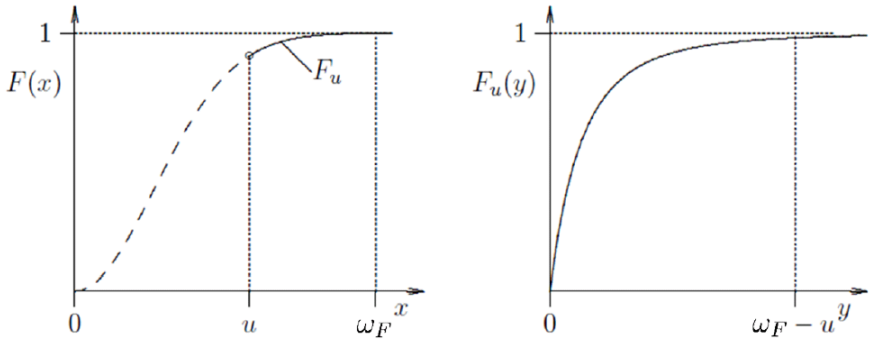

Figura 1.4: Función de Distribución F y Función de Distribución de Excesos F_u

1.3.1. La Distribución Generalizada de Pareto

La distribución asociada al modelo de Picos sobre un Umbral es la distribución generalizada de Pareto que se define a continuación.

DEFINICIÓN 1.9. *(**Distribución Generalizada de Pareto**). Se define a la función de distribución H_ξ como*

$$H_\xi(x) = \begin{cases} 1 - (1+\xi x)^{-1/\xi} & , \ \xi \neq 0 \\ 1 - e^{-x} & , \ \xi = 0 \end{cases}$$

si $x \geq 0$ para $\xi \geq 0$; $0 \leq x \leq -1/\xi$ para $\xi < 0$ y

$$H_\xi(x) = \begin{cases} 0 & , \ x < 0 \\ 1 & , \ x > -1/\xi \end{cases}$$

para $\xi < 0$ en el último caso.

Como en el caso de G_ξ, notemos que podemos llegar a la expresión de H_0 a partir de H_ξ tomando el límite en cero para ξ. Se le conoce a H_ξ como la Distribución Generalizada de Pareto estándar (DGP) a la cual se le pueden añadir parámetros de localización ν y escala β, para representar a la familia de distribuciones $H_{\xi,\nu,\beta}$ reeplazando al argumento x por $(x-\nu)/\beta$ para $\nu \in \mathbb{R}$ y $\beta > 0$ y ajustando el soporte adecuadamente.

Observación 1.1. *La DGP y la función de excesos se relacionan con el dominio de atracción de la Distrubución Generalizada de Valores Extremos mediante la Proposición 1.5 como sigue. Supongamos que X es una variable aleatoria con función de distribución $F \in \mathcal{D}(G_\xi)$ y que $a(\cdot)$ es una función positiva, entonces tenemos que*

$$\begin{aligned}
\lim_{u\uparrow\omega_F} F_u(xa(u)) &= 1 - \lim_{u\uparrow\omega_F} P\left(\frac{X-u}{a(u)} > x \;\middle|\; X > u\right) \\
&= 1 - \lim_{u\uparrow\omega_F} \frac{P(X > u + xa(u))}{P(X > u)} \\
&= 1 - \lim_{u\uparrow\omega_F} \frac{\overline{F}(u + xa(u))}{\overline{F}(u)} \\
&= 1 - (1+\xi x)^{-1/\xi} \\
&= H_\xi(x)
\end{aligned}$$

Esta relación nos dice que si F_u es una función de distribución de excesos asociada a una distribución F que cumpla con ciertas características, entonces la DGP puede ser usada como una aproximación de la disfribución de los excesos escalados sobre un umbral alto u.

La observación 1.1 puede reformularse como el siguiente resultado, el cual es uno de los centrales en la Teoría de Valores Extremos. El enunciado de este teorema fue consultado en [14].

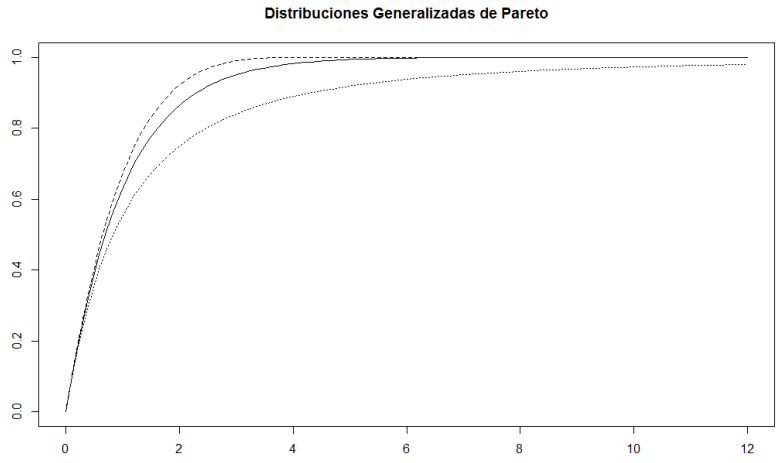

Figura 1.5: DGP con $\xi = 0$ (línea continua), $\xi = -0.2$ (guiones) y $\xi = 0.5$ (puntos) y parámetro común $\beta = 1$.

Teorema 1.3. ***(Pickands-Balkema-de Haan)***. *Para una clase grande de funciones de distribución F y un umbral alto u, si F_u es la función de distribución de excesos entonces*

$$F_u(y) \sim H_{\xi,\beta}(y), \qquad u \to \infty,$$

donde

$$H_{\xi,\beta}(y) = \begin{cases} 1 - \left(1 + \xi \frac{y}{\beta}\right)^{-1/\xi} & , \ \xi \neq 0 \\ 1 - e^{-y/\beta} & , \ \xi = 0 \end{cases} \tag{1.25}$$

para $y > 0$ *cuando* $\xi \geq 0$*;* $0 \leq y \leq -\beta/\xi$ *cuando* $\xi < 0$ *y*

$$H_{\xi,\beta}(y) = \begin{cases} 0 & , \ y < 0 \\ 1 & , \ y > -\beta/\xi \end{cases}$$

para $\xi < 0$ *en el último caso.*

La expresión en (1.25) del Teorema de Pickands-Balkema-de Haan define a la Distribución Generalizada de Pareto denotada por $H_{\xi,\beta}$. Si definimos a $x = y+u$, la DGP puede expresarse en términos de x como $H_{\xi,u,\beta}(x) = 1 - (1 + \xi(x-u)/\beta)^{-1/\xi}$. Este resultado es la base para la estimación de la distribución de los excesos sobre umbrales suficientemente altos. En la figura 1.5 se puede observar la gráfica de la DGP cuando el parámetro de forma ξ toma un valor positivo, uno negativo y cero, con el parámetro de escala β igual a 1.

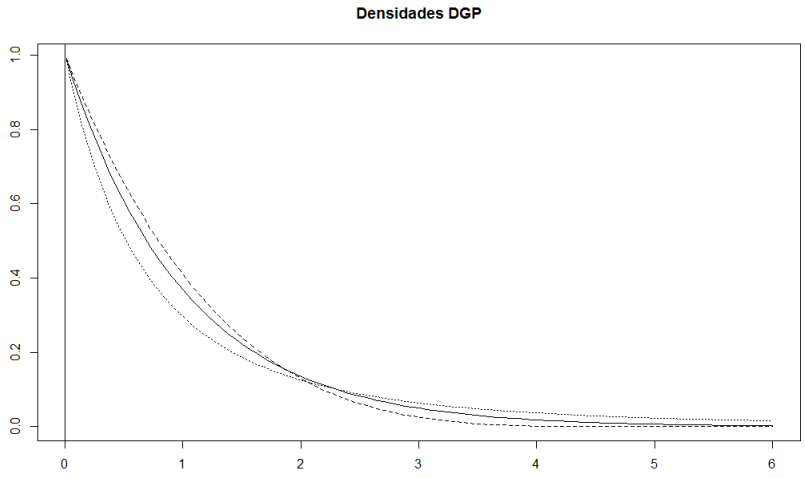

Figura 1.6: Densidades de la DGP con $\xi = 0$ (línea continua), $\xi = -0.2$ (guiones) y $\xi = 0.5$ (puntos) y parámetro común $\beta = 1$.

Una función que resulta de importancia para el análisis de extremos en el modelo de Picos sobre un Umbral es la función media de exceso, la cual es la esperanza del valor del exceso

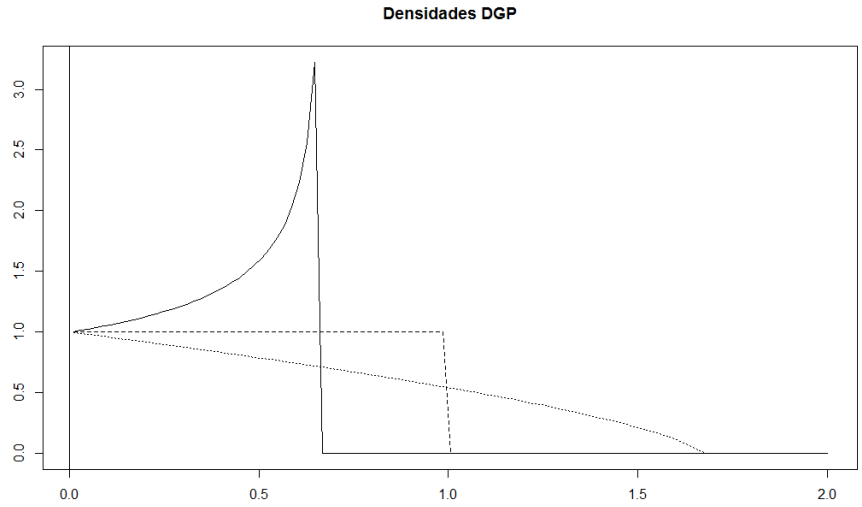

Figura 1.7: Densidades de la DGP con $\xi = -1.5$ (línea continua), $\xi = -1$ (guiones) y $\xi = -0.6$ (puntos) y parámetro común $\beta = 1$. Todos con punto final derecho $\omega_F = -1/\xi$.

de una variable aleatoria X sobre un umbral u, condicionada a que dicha variable excedió el umbral.

DEFINICIÓN 1.10. *(**Función Media de Excesos**). Sea X una variable aleatoria con función de distribución F y punto final derecho ω_F. Para $u < \omega_F$ fijo,*

$$e(u) := E(X - u | X > u) \tag{1.26}$$

es llamada la Función Media de Exceso de X.

Esta esperanza como función de u servirá como una herramienta gráfica para la selección de un umbral adecuado para realizar una aproximación con la DGP. En el Teorema siguiente se enumeran algunas de propiedades básicas de la DGP que servirán para sustentar las aproximaciones estadísticas a la distribución de los excesos.

Teorema 1.4. *(**Propiedades de la DGP**)*

a) Para toda $\xi \in \mathbb{R}$, $F \in \mathcal{D}(G_\xi)$ si y sólo si

$$\lim_{u \uparrow \omega_F} \sup_{0 < x < \omega_F - u} |F_u(x) - H_{\xi, \beta(u)}(x)| = 0 \tag{1.27}$$

para alguna función positiva β.

b) Supongamos que x_1 y x_2 están en el dominio de $H_{\xi,\beta}$, entonces

$$\frac{\overline{H}_{\xi,\beta}(x_1+x_2)}{\overline{H}_{\xi,\beta}(x_1)} = \overline{H}_{\xi,\beta+\xi x_1}(x_2) \tag{1.28}$$

c) Supongamos que X tiene DGP con parámetros $\xi < 1$ y β. Entonces para $u < \omega_F$,

$$e(u) = E(X-u|X>u) = \frac{\beta+\xi u}{1-\xi}, \quad \beta+\xi u > 0. \tag{1.29}$$

Demostración. a) Podemos expresar a (1.23) de la Proposición 1.5 como

$$\lim_{u\uparrow\omega_F}\left|\frac{\overline{F}(u+xa(u))}{\overline{F}(u)} - (1+\xi x)^{-1/\xi}\right| = 0.$$

Tenemos entonces que

$$\begin{aligned}\lim_{u\uparrow\omega_F}\left|\left(1-\frac{\overline{F}(u+x)}{\overline{F}(u)}\right) - \left(1-\left(1+\xi\frac{x}{a(x)}\right)^{-1/\xi}\right)\right| &= \lim_{u\uparrow\omega_F}|F_u(x) - H_{\xi,a(u)}(x)| \\ &= 0.\end{aligned}$$

Como la DGP es una función continua, tenemos la convergencia uniforme.

b) Se puede verificar directamente de la expresión de $H_{\xi,\beta}$.

c) Tenemos que para $\xi < 1$

$$\begin{aligned} e(u) &= \int_u^\infty (x-u)dH_{\xi,\beta}(x)/\overline{H}_{\xi,\beta}(u)du \\ &= \frac{1}{\overline{H}_{\xi,\omega}(u)}\int_u^\infty\int_u^x dydH_{\xi,\beta}(x) \\ &= \frac{1}{\overline{H}_{\xi,\omega}(u)}\int_u^\infty\int_y^\infty dH_{\xi,\beta}(x)dy \\ &= \frac{1}{\overline{H}_{\xi,\omega}(u)}\int_u^\infty \overline{H}_{\xi,\beta}(y)dy \\ &= \left(1+\xi\frac{u}{\beta}\right)^{1/\xi}\int_u^\infty\left(1+\xi\frac{y}{\beta}\right)^{-1/\xi}dy \\ &= \left(1+\xi\frac{u}{\beta}\right)\frac{\beta}{1-\xi} \\ &= \frac{\beta+\xi u}{1-\xi}\end{aligned}$$

□

Con la propiedad *a)* queda demostrado el Teorema de Pickands-Balkema-de Haan, el cual es el resultado fundamental en el la justificación del uso de la DGP para la aproximación de

la distribución de los excesos de una variable aleatoria sobre un umbral alto. La propiedad b) puede interpretarse como sigue. Supongamos que Y tiene distribución DGP(ξ, β) y sea u_0 un umbral fijo, entonces para $u > u_0$,

$$\begin{aligned} \mathbb{P}(Y > u | Y > u_0) &= \frac{\mathbb{P}(Y > u)}{\mathbb{P}(Y > u_0)} && (1.30) \\ &= \frac{\overline{H}_{\xi,\beta}(u)}{\overline{H}_{\xi,\beta}(u_0)} && (1.31) \\ &= \overline{H}_{\xi,\beta+\xi u_0}(u - u_0). && (1.32) \end{aligned}$$

Esto significa que si la DGP es un modelo adecuado para la distribución de excesos sobre un umbral u_0, entonces las distribuciones de excesos sobre umbrales $u > u_0$ también estarán bien ajustadas por la DGP, donde los parámetros de forma ξ permanecen constantes y el parámetro de escala β cambia de la forma $\tilde{\beta} = \beta + \xi u_0$. Esto en conjunto con la propiedad c) brindan un par de técnicas gráficas para la elección de un umbral adecuado para que la aproximación de F_u pueda hacerse mediante la DGP, la cual será descrita en el siguiente capítulo.

En resumen, en este capítulo se vio que la distribución generalizada de valores extremos describe las distribución límite de máximos normalizados, asociados con el modelo de Máximos por Bloque para elegir observaciones extremas, respaldado por el Teorema de Fisher-Tippet-Gnedenko; y la distribución generalizada de Pareto aparece como la distribución límite de excedencias escaladas sobre umbrales altos, en el modelo de Picos sobre un Umbral o POT respaldado por el Teorema de Pickands-Balkema-de Haan.

Capítulo 2

Modelación estadística

En el capítulo anterior se introdujeron algunos de los modelos probabilísticos para describir matemáticamente a los eventos extremos de una muestra en el caso unidimensional. En este capítulo se mostrarán algunas de esas herramientas y técnicas para dar aproximaciones adecuadas mediante dichos modelos al comportamiento de los valores extremos de una muestra. La principal fuente utilizada en esta sección fue [5].

En la práctica podemos encontrar diversos fenómenos que pueden ser estudiados bajo el enfoque de la Teoría de Valores Extremos para contestar una serie de preguntas sobre el comportamiento de sus valores máximos. Lo que se busca es hacer uso de las herramientas disponibles para encontrar un modelo que explique estadísticamente el comportamiento de las variables de interés.

2.1. Herramientas estadísticas

Para obtener información de una muestra de cierta distribución se pueden utilizar una serie de herramientas para hacer inferencia sobre la estructura probabilística de la población de donde provienen las observaciones.

2.1.1. Primeras herramientas para el análisis de datos

Un primer análisis que puede realizarse es mediante la visualización de los datos a través de herramientas gráficas exploratorias que brindan información de la distribución de una

muestra tales como una serie de tiempo, la cual nos permite localizar temporalmente la ocurrencia de eventos extremos e identificar conglomeraciones en ellos, el histograma de los datos, el cual da una aproximación visual a la forma de la función de densidad, y la Función de Distribución Empírica. Para ejemplificar estas funciones, utilizaremos una base de datos contenida en el paquete POT del software R, la cual consiste en los registros de flujos de inundación en metros cúbicos por segundo del río Ardieres en Beaujeu, Francia durante un periodo de 1970 a 2004.

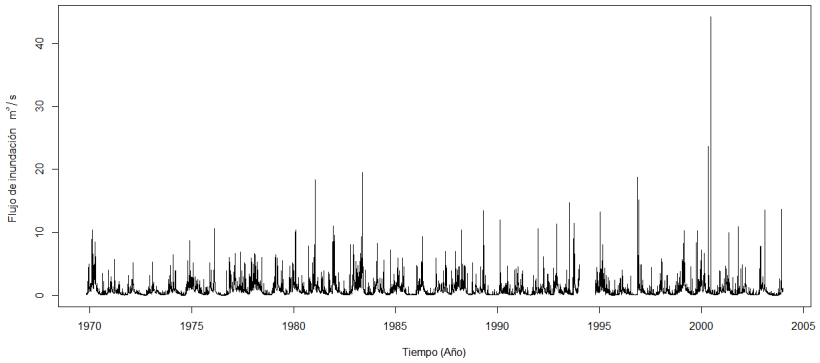

Figura 2.1: Serie de Tiempo de los flujos de inundación del río Ardieres.

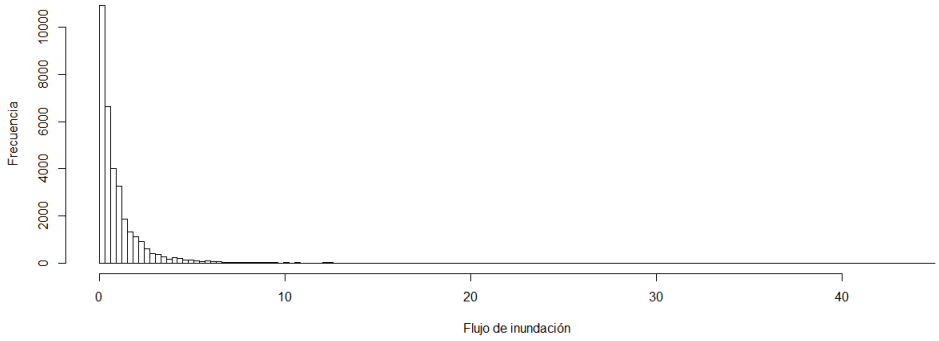

Figura 2.2: Histograma de frecuencias de flujos de inundación del río Ardieres.

DEFINICIÓN 2.1. **(Función de Distribución Empírica)** *Sea $x_1, \ldots, x_n$ una muestra de una Variable Aleatoria con función de distribución F. Definimos a la función de distribución*

empírica $\tilde{F}_n$ *por*

$$\tilde{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(-\infty,t]}(x_i).$$

Donde $\mathbb{I}_{(-\infty,t]}(x_i)$ *es la función identidad, que vale* 1 *si* $x_i \leq t$ *y* 0 *si no.*

La función de distribución empírica es una buena aproximación de la distribución real de una sucesión de variables aleatorias independientes e idénticamente distribuidas, tal como lo sustenta el Teorema de Gilvenko-Cantelli (Teorema A.2).

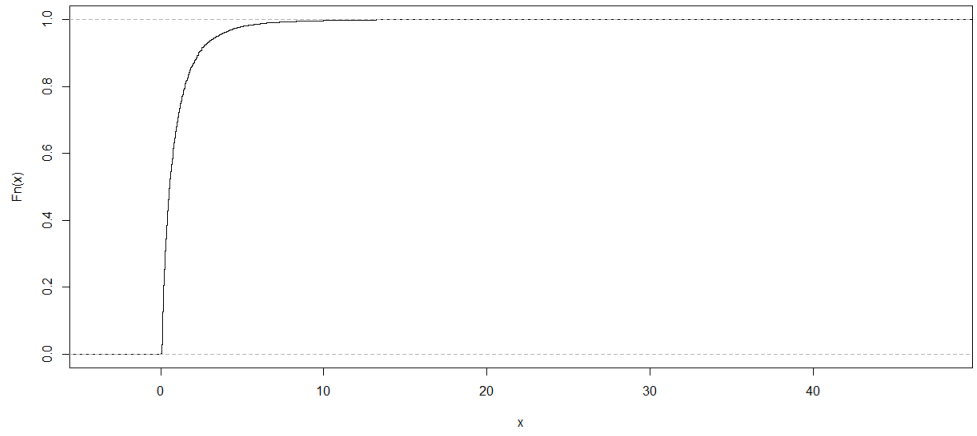

Figura 2.3: Distribución empírica de los flujos de inundación del río Ardieres.

Otra forma de obtener una primera aproximación a la distribución real de una muestra de datos es la siguiente. Consideremos la muestra ordenada $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Para cada una de las observaciones $x_{(i)}$, exactamente i de las n observaciones tienen un valor menor o igual a $x_{(i)}$, así que un estimador empírico de que la probabilidad de una observación sea menor o igual a $x_{(i)}$ es $\tilde{F}(x_{(i)}) = i/n$. Se suele recurrir a un ajuste haciendo $\tilde{F}(x_{(i)}) = i/(n+1)$ para evitar que el valor máximo de la muestra funja como punto final derecho de la distribución empírica ya que de ser así se tendría que $\omega_{\tilde{F}} < \infty$ [5], lo cual nos da la definición siguiente.

Definición 2.2. *Si se tiene una muestra ordenada de tamaño* n, $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ *entonces se define a la función de distribución empírica* $\tilde{F}_n$ *como*

$$\tilde{F}_n(x) = \frac{i}{n+1}, \tag{2.1}$$

para $x_{(i)} \leq x < x_{(i+1)}$.

Existen otras versiones de la definición de $\tilde{F}$ que son usadas principalmente para dar una aproximación de las probabilidades de no excedencia, así como para aproximar la ubicación de las observaciones en gráficas de diagnóstico de ajuste de algún modelo. Una de ellas fue propuesta por Hosking en [13] que se utiliza principalmente para aproximar la posición de graficación de probabilidades empíricas, la cual consiste en calcular dichas probabilidades como

$$\tilde{F}_n(x) = \frac{i - 0.35}{n}$$

para $x_{(i)} \leq x < x_{(i+1)}$. Estas estimaciones son utilizadas en el software R en [20].

2.1.2. Estimación de parámetros

El objetivo de la modelación estadística que nos ocupa es usar la información muestral para hacer inferencia sobre la estructura probailística del fenómeno del que es obtenida dicha información. Supongamos que se busca hacer inferencia estadística sobre los parámetros representados por el vector θ, de una distribución F que pertenece a una familia de distribuciones conocida $\mathcal{F} = \{f(x; \theta : \theta \in \Theta)\}$, como se hace en [5]. Denotamos al valor real de θ como θ_0 y se hace inferencia sobre este valor en el espacio paramétrico Θ. A las funciones de variables aleatorias usadas para estimar el valor real del parámetro, o vector de parámetros θ_0 se les conoce como estimadores y se denotan generalmente por $\hat{\theta}_0$.

Como la información sobre la que se hacen las estimaciones proviene de variables aleatorias, esta aleatoriedad de los datos muestrales induce aleatoriedad en los estimadores. La distribución de probabilidad inducida en el estimador es llamada la distribución muestral, y ésta determina la variabilidad de los estimadores. La desviación estándar de esta distribución es llamada el error estándar del estimador $\hat{\theta}_0$ el cual es implícitamente una medida de la precisión de éste, por lo que se prefieren aquellos estimadores con el error estándar más pequeño. Otra forma de cuantificar la precisión de un estimador es mediante sus intervalos de confianza.

Definición 2.3. *Para un valor* $0 < \alpha < 1$*, el intervalo dado por* $[\theta_l, \theta_u]$ *es el intervalo a* $1 - \alpha$ *de confianza para* θ_0 *si*

$$\mathbb{P}(\theta_l \leq \theta_0 \leq \theta_u) = 1 - \alpha.$$

La elección de α es arbitraria, ya que valores muy pequeños dan altos niveles de confianza pero resultan en intervalos grandes, mientras que valores grandes dan intervalos pequeños pero con pocos niveles de confianza. Los valores más utilizados son $\alpha = 0.05$, 0.01 y 0.001. En este trabajo se utilizará $\alpha = 0.05$ que resultará en intervalos a 95 % de confianza.

Máxima verosimilitud

Existen varios métodos propuestos que pueden utilizarse para la estimación de los parámetros de una distribución a partir de una muestra. Uno de los más flexibles y comúnmente utilizados es el de máxima verosimilitud. Este método, en su forma más simple, supone una muestra $x_1, \ldots, x_n$ de variables aleatorias independientes e idénticamente distribuidas proviene de una familia de distribuciones conocida $F = \{f(x; \theta : \theta \in \Theta)\}$. Cada valor de $\theta \in \Theta$ define a un modelo en específico que potencialmente describe las probabilidades de la información de la muestra. A estas probabilidades, vistas como función de θ se les conoce como función de verosimilitud.

DEFINICIÓN 2.4. *Sea $x_1, \ldots, x_n$ una muestra de variables aleatorias independientes e idénticamente distribuidas con función de densidad común $f(x; \theta_0)$. Se define a la* ***Función de Verosimilitud*** *como*

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta) \tag{2.2}$$

Usualmente es más conveniente tomar logaritmos y trabajar con la función de Log-Verosimilitud

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta). \tag{2.3}$$

Aquellos valores de θ que tengan una alta verosimilitud corresponerán a aquellos modelos para los que la muestra tenga altas probabilidades de ser observada. El principio de la estimación por máxima verosimilitud consiste en adoptar el modelo con la mayor verosimilitud. El estimador de máxima verosimilitud $\hat{\theta}_0$ de θ_0 se define como el valor θ que maximiza a la función de verosimilitud. Como la función logaritmo es monótona, la log-verosimilitud alcanza su máximo en el mismo punto que la verosimilitud, así que $\hat{\theta}_0$ maximiza tanto a L como a $\log L$.

En algunos casos es posible obtener el estimador de máxima verosimilitud de manera explícita, diferenciando la función de log-verosimilitud e igualando a cero. En otros casos es necesario utilizar métodos numéricos para encontrar el estimador máximo verosímil, tal

como es el caso de los estimadores para los parámetros de la DGVE y de la DGP. Los software y paquetes estadísticos que realizan las estimaciones puntuales también proveen de los intervalos de confianza de cada parámetro estimado al nivel que se desee mediante varios métodos. Puede consultarse [5], [20] y [26] para mayor detalle. Existen otros métodos de estimación de parámetros como el métdo de los momentos, o el de probabilidades ponderadas, pero en este trabajo utilizaremos el presente método, cuyas propiedades pueden consultarse en diversas fuentes, entre ellas [5].

2.1.3. Herramientas de diagnóstico de modelos

Una vez que se obtienen los estimadores de los parámetros de un modelo probabilístico que explican el comportamiento de una muestra, se espera que dichas estimaciones representen un ajuste adecuado a los datos observados. Para verificarlo existen algunas herramientas gráficas que sirven para dar un diagnóstico visual de la precisión con la que un modelo estimado explica la información deseada. Supongamos que $x_1, x_2, \ldots, x_n$ es una muestra aleatoria de una variable con función de distribución desconocida F. Denotamos por $\hat{F}$ a un estimador de F, obtenido por ejemplo por máxima verosimiltud, y queremos evaluar qué tan plausible es que la muestra de x_i provenga de $\hat{F}$.

Una primera comparación que puede hacerse es mediante la función de distribución empírica de la muestra $\tilde{F}$, ya que ésta también representa una estimación de la distribución verdadera F. Para un ajuste correcto se espera entonces que $\hat{F}$ y $\tilde{F}$ muestren una concordancia razonable. Existen diversas herramientas de diagnóstico de ajuste basadas en comparaciones de $\hat{F}$ y $\tilde{F}$, comúmnente se usan las siguientes técnicas gráficas.

DEFINICIÓN 2.5. *Dada una muestra ordenada de observaciones independientes $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ de una variable con función de distribución estimada $\hat{F}$, una* ***gráfica de probabilidades****, o P-P Plot por su nombre en inglés (Probability Plot) consiste en los puntos*

$$\left\{ \left(p_i, \hat{F}(x_{(i)}) \right) : i = 1, \ldots, n \right\},$$

donde p_i usualmente se determina por $i/(n+1)$, o bien $(i - 0.35)/n$ según [13].

Si $\hat{F}$ es un modelo razonable para la muestra, los puntos graficados deberán asemejarse a una línea recta diagonal. Si los puntos se alejan demasiado de dicha recta significa que el modelo $\hat{F}$ no es el adecuado.

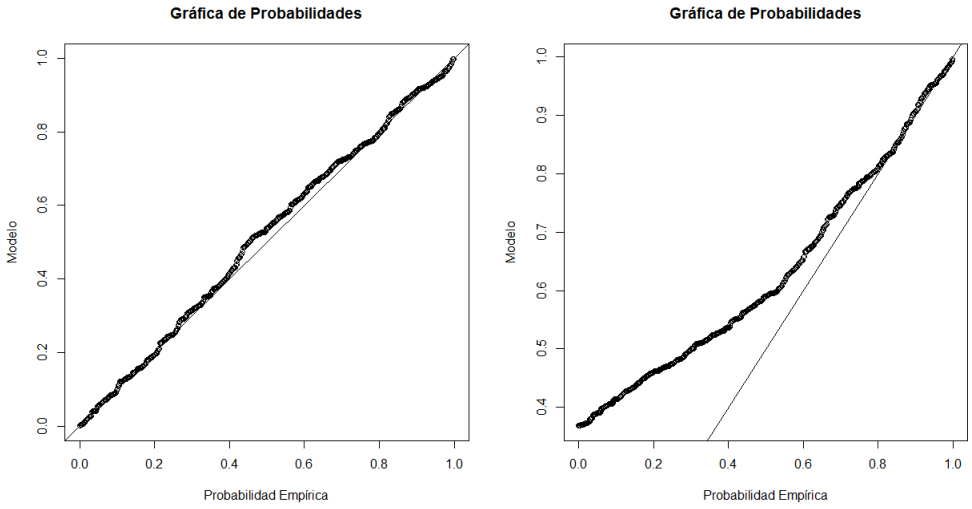

Figura 2.4: Gráficas de Probabilidades de simulaciones de variables exponenciales (izquierda) y uniformes (derecha) contra el modelo exponencial.

La siguiente herramienta gráfica hace uso de las estimaciones de los cuantiles, usando que $x_{(i)}$ provee un estimador del cuantil $i/(n+1)$ y que puede obtenerse la función de cuantiles estimada a partir de la inversa generalizada de la función de distribución estimada que denotaremos por $\hat{F}^{\leftarrow}$. Definimos entonces a la gráfica de cuantiles como sigue.

DEFINICIÓN 2.6. *Dada una muestra ordenada de observaciones independientes* $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ *de una variable con función de distribución estimada* $\hat{F}$*, una* ***gráfica de cuantiles****, o Q-Q Plot por su nombre en inglés (Quantile Plot) consiste en los puntos*

$$\left\{ \left(\hat{F}^{\leftarrow}(p_i), x_{(i)} \right) : i = 1, \ldots, n \right\},$$

donde p_i *se calcula de igual manera que en la gráfica de probabilidades.*

Del mismo modo de que la gráfica de probabilidades, los puntos en la gráfica de cuantiles deberían asemejarse a una recta diagonal si $\hat{F}$ es una estimación razonable. Ambas gráficas contienen la misma información expresada en escalas diferentes, lo cual permite tener una perspectiva más amplia de la veracidad del modelo en juicio, ya que un modelo que aparente tener un buen ajuste en una gráfica, podría no mostrar lo mismo en la otra. También se utilizan otras variantes de estas gráficas, en las que se utiliza el estimador propuesto por Hosking, donde en vez de $i/(n+1)$ se utiliza $(i-0.35)/n$.

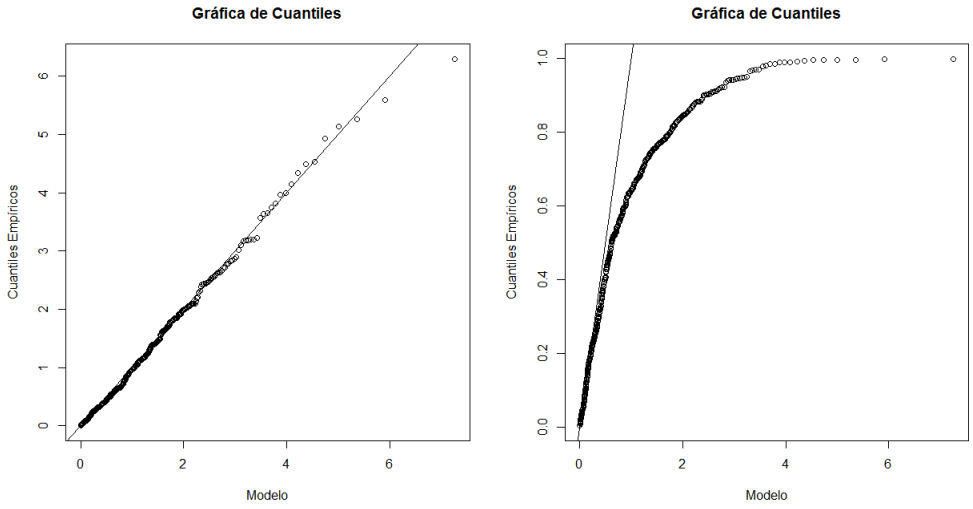

Figura 2.5: Gráficas de Cuantiles de simulaciones de variables exponenciales (izquierda) y uniformes (derecha) contra el modelo exponencial.

Otra de las gráficas que se utilizan como herramienta de diagnóstico es la gráfica de la densidad asociada a la distribución ajustada $\hat{f}$ con el histograma de la muestra. Como mencionamos anteriormente, la gráfica de un histograma sirve para estimar la función de densidad de una muestra, pues en ella se pueden observar las proporciones de los datos por clases. Un modelo adecuado hará que la curva de densidad ajustada se asemeje a las barras del histograma. Esta herramienta depende de la elección de las clases con las que se elabora el histograma de los datos, en el que usualmente se utiliza la regla de Sturges (ver [27]) para elegir la cantidad de dichas clases.

2.2. Estimación de parámetros y cuantiles en la TVE

Una vez que se han visto las herramientas utilizadas para hacer inferencia sobre una muestra, aplicaremos los resultados principales de la Teoría de Valores Extremos para obtener herramientas especiíficas para el análisis de eventos considerados como extremos en los dos modelos principales que son el modelo de Máximos por Bloque y el modelo de Picos sobre un Umbral.

2.2.1. Inferencia en el modelo de Máximos por Bloque

Una vez realizado el análisis preliminar de los datos, procedemos a estimar los parámetros de la distribución asociada al modelo de Máximos por Bloque. Si tenemos una muestra de de m máximos de bloques de tamaño n $M_{1;n}, M_{2;n}, \ldots, M_{m;n}$ que denotaremos simplemente por $x_1, \ldots, x_m$, el Teorema de Fisher-Tippet-Gnedenko nos dice que si n es grande la distribución de la que proviene dicha muestra puede aproximarse por la DGVE.

Estimación de parámetros de la DGVE

Si suponemos que $x_1, \ldots, x_m$ es una muestra proveniente de una DGVE, tenemos que la función de log-verosimilitud para los parámetros de la DGVE cuando $\xi \neq 0$ es

$$l(\xi, \mu, \sigma) = -m \ln \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{m} \ln \left(1 + \xi \left(\frac{x_i - \mu}{\sigma}\right)\right) - \sum_{i=1}^{m} \left(1 + \xi \left(\frac{x_i - \mu}{\sigma}\right)\right)^{-1/\xi}, \tag{2.4}$$

con $1 + \xi(x_i - \mu)/\sigma > 0$ para $i = 1, \ldots, m$. Para el caso $\xi = 0$ se tiene que la función de log-verosimilitud es

$$l(\mu, \sigma) = -m \ln \sigma - \sum_{i=1}^{m} \left(\frac{x_i - \mu}{\sigma}\right) - \sum_{i=1}^{m} \exp\left\{-\left(\frac{x_i - \mu}{\sigma}\right)\right\}. \tag{2.5}$$

No existen expresiones analíticas para el vector de parámetros (ξ, μ, σ) que maximiza las ecuaciones (2.4) y (2.5) por ello los paquetes estadísticos utilizan métodos numéricos para obtener el vector de estimadores de máxima verosimilitud $\hat{\theta} = (\hat{\xi}, \hat{\mu}, \hat{\sigma})$.

Periodos de Retorno

Para estimar los periodos de retorno asociados a los cuantiles extremos de la distribución en el modelo de Máximos por Bloque podemos proceder como sigue. Sabemos por el teorema de Fisher-Tippet-Gnedenko que $G_{\xi,\sigma,\mu}(x_0)$ es la probabilidad límite de que el máximo de un bloque, digamos de un año, sea menor o igual que el valor x_0. Invirtiendo la expresión de la DGVE para $\xi \neq 0$ tenemos que

$$G^{\leftarrow}_{\xi,\mu,\sigma}(t) = \mu + \frac{\sigma}{\xi}\left((-\ln t)^{-\xi} - 1\right), \tag{2.6}$$

para $0 < t < 1$. El caso en el que $\xi = 0$ se trata de la misma manera, solamente utilizando la expresión de G_0

$$G^{\leftarrow}_{0,\mu,\sigma}(t) = \mu - \sigma \ln(-\ln t). \tag{2.7}$$

Si queremos encontrar ahora un valor x_p tal que $G_{\xi,\sigma,\mu}(x_p) = 1-p$ o bien que la probabilidad de que valor del máximo anual exceda el nivel x_p sea p basta con evaluar las expresiones (2.6) y (2.7) en $1-p$ para obtener

$$x_p = \mu + \frac{\sigma}{\xi}\left((-\ln(1-p))^{-\xi} - 1\right) \tag{2.8}$$

para $\xi \neq 0$ y

$$x_p = \mu - \sigma \ln(-\ln(1-p)) \tag{2.9}$$

para $\xi = 0$. A este valor se le conoce como el nivel de retorno asociado al periodo de retorno $T(p) = 1/p$, ya que el nivel x_p se espera que será excedido en promedio una vez cada $1/p$ años. Esta relación entre probabilidades de excedencia y periodos de retorno sirve para determinar los niveles que se espera serán excedidos en un lapso de una cierta cantidad de años. Una gráfica de los periodos de retorno contra sus niveles de retorno asociados es conocida como la gráfica de nivel de retorno, es decir

$$(T(p), \hat{x}_p)\,,$$

donde $0 < p < 1$ y $\hat{x}_p$ es el nivel de retorno calculado con los parámetros estimados. Ya que su interpretación es bastante simple, se suele graficar en escala logarítmica en el eje de las abscisas, con el fin de hacer más visibles los niveles asociados a periodos cortos sin perder de vista los periodos largos. Esta gráfica es de utilidad para presentar el modelo estimado y también como herramienta de diagnóstico de ajuste como lo son las funciones de probabilidad, de cuantiles y de densidad ajustada. Para ello se grafican sobre la curva de de nivel de retorno los puntos de la muestra ordenada $x_{(1)}, \ldots, x_{(n)}$ como

$$\left(T(q_i), x_{(i)}\right) \tag{2.10}$$

para $i = 1, \ldots, n$ donde $q_i = 1 - p_i$ la probabilidad de excedencia asociada a las p_i, que son calculadas de la misma forma que en las gráficas de probabilidad y de cuantiles. Al igual que en estas últimas, si el modelo estimado representa un buen ajuste a los datos, los puntos se asemejarán a la curva de nivel de retorno, a la cual en ocasiones se le añaden intervalos de confianza para dar mayor información. Usualmente resultan de mayor interés aquellos periodos de retorno grandes, correspondientes con valores pequeños de p.

Diagnóstico de modelos DGVE

Como se describió anteriormente, una gráfica de probabilidades es una comparación de las distribuciones empírica y estimada. Si tenemos una muestra ordenada de máximos de m bloques $x_{(1)}, \ldots, x_{(m)}$, la función de distribución empírica evaluada en $x_{(i)}$ está dada por $\tilde{G}(x_i) = p_i$. Sustituyendo los parámetros estimados en la expresión de la DGVE tenemos que el modelo estimado evaluado en las observaciones es

$$\hat{G}(x_{(i)}) = \exp\left\{-\left(1+\hat{\xi}\left(\frac{x_{(i)}-\hat{\mu}}{\hat{\sigma}}\right)\right)^{-1/\hat{\xi}}\right\}.$$

Si el modelo de la DGVE es adecuado, entonces

$$\hat{G}(x_{(i)}) \approx \tilde{G}(x_{(i)})$$

para cada i, así que la gráfica de probabilidades que consiste en los puntos

$$\left\{\left(\tilde{G}(x_{(i)}), \hat{G}(x_{(i)})\right), i = 1, \ldots, m\right\},$$

debería asemejarse a una recta diagonal. El ajuste de las estimaciones que son de mayor interés se pueden apreciar mejor en la gráfica de cuantiles que consiste en los puntos

$$\left\{\left(\hat{G}^{\leftarrow}(p_i), x_{(i)}\right), i = 1, \ldots, m\right\},$$

donde

$$\hat{G}^{\leftarrow}(p_i) = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}}\left((-\ln p_i)^{-\hat{\xi}} - 1\right).$$

2.2.2. Inferencia en modelos de Picos sobre un Umbral

Una vez realizado el análisis y estimación de parámetros bajo el modelo de Máximos por Bloque, si se cuenta con la información suficiente, la Proposición 1.5 nos da pie para continuar con el análisis y ajuste bajo el enfoque de Picos sobre un Umbral. Supongamos ahora que contamos con una muestra $x_1, x_2, \ldots, x_n$ que proviene de una serie de variables aleatorias independientes e idénticamente distribuidas. Como se mencionó anteriormente, identificamos a los excedentes sobre un umbral u como aquellas observaciones que sobrepasen dicho umbral. Denotamos a esas excedencias $x_{\langle 1\rangle}, x_{\langle 2\rangle}, \ldots, x_{\langle k\rangle}$ y definimos a los excesos sobre u como $y_i = x_{\langle i\rangle} - u$ para $i = 1, \ldots, n$. Por el Teorema de Pickands-Balkema-de Haan sabemos que los excesos y_i provienen de una distribución que puede aproximarse por la DGP con un umbral u suficientemente alto.

Elección del umbral

La elección del umbral adecuado no puede estimarse puntualmente, así que representa un problema que requiere un análisis previo a las estimaciones de los parámetros de la DGP y que debe tratarse con cuidado, ya que una elección de un umbral muy bajo violaría la suposición del límite de la aproximación, mientras que un umbral demasiado alto dejará muy pocas observaciones de excesos para hacer las estimaciones. Se busca entonces elegir un umbral lo más bajo posible para mantener un número suficiente de excesos, sujeto a que la distribución límite sea una aproximación razonable. Existen varias herramientas disponibles para determinar el umbral adecuado, dos de ellas son un análisis exploratorio haciendo uso de la función media de excesos, y la otra es una valoración de la estabilidad de los estimadores de los parámetros, basado en modelos ajustados a diferentes valores de u.

La primera de las herramientas se apoya en los resultados del Teorema 1.4 donde se enuncian algunas propiedades de la DGP, una de las cuales dice que la función media de excesos de una variable aleatoria con esta distribución y parámetros ξ y β tiene la expresión

$$e(u) = E(X - u|X > u) = \frac{\beta + \xi u}{1 - \xi}.$$

En conjunto con la propiedad de cerradura de la clase de la distribución generalizada de Pareto, la cual dice que si un modelo DGP es adecuado para aproximar la distribucón de excesos sobre un umbral u_0 entonces debe ser válido para umbrales $u > u_0$, nos indica que para umbrales $u > u_0$ la función media de excesos es lineal, si u_0 es un umbral adecuado. Tenemos entonces un primer criterio que puede usarse para la elección de un umbral adecuado para ajustar una DGP a los excesos de un conjunto de datos. Comenzando por aproximar empíricamente a $e(u)$ con la media muestral de los excesos sobre u.

DEFINICIÓN 2.7. *Sea $x_1, \ldots, x_n$ una muestra y $x_{\langle 1\rangle}, x_{\langle 2\rangle}, \ldots, x_{\langle k\rangle}$ aquellas excedencias que sobrepasan un umbral u, la* ***gráfica de medias de excesos*** *consiste en los puntos*

$$\left\{ \left(u, \frac{1}{k} \sum_{i=1}^{k} (x_{\langle i\rangle} - u) \right) : u < x_{(n)} \right\},$$

donde $x_{(n)}$ es el valor máximo de las observaciones.

Como se espera que estas estimaciones cambien linealmente con u en niveles en los que el umbral es adecuado para el ajuste de una DGP, se elige un umbral para el cual la gáfica de media de excesos tenga un comportamiento aproximadamente lineal. Otro procedimiento

consiste en hacer estimaciones de los parámetros ξ y β de la DGP en un rango de umbrales, y se describirá más adelante.

Estimación de parámetros de la DGP

Una vez elegido un umbral adecuado se pueden estimar los parámetros de la DGP por máxima verosimilitud. Supongamos que los valores $y_1, \ldots, y_k$ son los k excesos sobre el umbral u. Para $\xi \neq 0$ la función de log-verosimilitud es

$$l(\xi, \beta) = -k \ln \beta - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{k} \ln(1 + \xi y_i/\beta), \tag{2.11}$$

con $(1 + \xi y_i/\sigma) > 0$ para $i = 1, \ldots, k$. Para el caso $\xi = 0$ la log-verosimilitud queda como sigue

$$l(\beta) = -k \ln \beta - \frac{1}{\beta} \sum_{i=1}^{n} y_i. \tag{2.12}$$

Al igual que el la estimación de parámetros de la DGVE, las ecuaciones 2.11 y 2.12 no pueden maximizarse de manera analítica, por lo que se recurre también a métodos numéricos para encontrar el vector de estimadores de máxima verosimilitud $(\hat{\xi}, \hat{\beta})$.

Revisión de la elección del umbral

Como se mencionó anteriormente, realizar la gráfica de medias de excesos brinda una primera herramienta para la elección de un umbral adecuado para el ajuste de una DGP. Sin embargo existe una técnica complementaria que puede realizarse mediante el análisis de la estabilidad del modelo, ajustando una DGP a una muestra en un rango de diferentes umbrales. Como se vio en el Teorema 1.4 del Capítulo 1, si la DGP es un modelo adecuado para la distribución de excesos sobre un umbral u_0, entonces las distribuciones de excesos sobre umbrales $u > u_0$ también estarán bien ajustadas por la DGP, donde los parámetros de forma ξ permanecen constantes y el parámetro de escala β cambia de la forma $\beta_u = \beta_{u0} + \xi u_0$. Para facilitar el procedimiento se recurre a una reparametrización del parámetro de escala por

$$\beta^* = \beta_u - \xi u_0.$$

Este nuevo parámetro se le llama simplemente escala modificada y es constante respecto a u al igual que ξ, por lo tanto los parámetros estimados de ξ y de β^* deberían comportarse de manera aproximadamente constante o estable cerca de u_0 si éste representa un umbral

adecuado para excesos que siguen una DGP. Adicionalmente a la gráfica de $\hat{\xi}$ y $\hat{\beta}^*$ contra u se grafican los intervalos de confianza de los estimadores para tener un criterio sobre la variabilidad de las estimaciones, y se selecciona el umbral como el menor valor u_0 para el cual los estimadores permanezcan aproximadamente constantes.

Niveles de retorno DGP

El Teorema de PIckands-Balkema-de Haan provee una aproximación a la distribución condicional de los excesos sobre un umbral u, esto nos permite estimar los cuantiles asociados a la cola de la distribución de las excedencias. Si una distribución generalizada de Pareto con parámetros (ξ, β) es un modelo adecuado para describir la distribución de los excesos sobre u, entonces podemos aproximar a F_u con $H_{\xi,\beta}$. Entonces para $x > u$

$$\overline{F}_u(x-u) = \left(1 + \xi\left(\frac{x-u}{\beta}\right)\right)^{-1/\xi}, \tag{2.13}$$

si $\xi \neq 0$, y

$$\overline{F}_u(x-u) = e^{-(x-u)/\beta}, \tag{2.14}$$

si $\xi = 0$. Supongamos ahora que queremos encontrar un valor o nivel $x_m > u$ tal que se exceda en promedio una vez cada m excedencias sobre u, o bien que la probabilidad de que una excedencia sobrepase el nivel x_m sea $1/m$, es decir, $\overline{F}_u(x_m - u) = 1/m$. Invirtiendo las expresiones (2.13) y (2.14) llegamos a que

$$x_m = u + \frac{\beta}{\xi}\left(m^{\xi} - 1\right)$$

para $\xi \neq 0$, y

$$x_m = u + \beta \ln m$$

si $\xi = 0$.

Por construcción, a x_m se le conoce como el m-observado nivel de retorno. Para representar a los niveles de retorno en una escala anual, digamos que buscamos el nivel de retorno que se espera se exceda una vez cada T años. Si hay n_y observaciones al año entonces corresponde al m-observado nivel de retorno con $m = Tn_y$. Observemos que dada una probabilidad p de no excedencia sobre u, el periodo de retorno en años asociado a esa probabilidad es igual a $T(p) = 1/(n_y(1-p))$. Entonces, el nivel de retorno asociado a un periodo de retorno de T años está dado por

$$x_T = u + \frac{\beta}{\xi}\left((Tn_y)^{\xi} - 1\right) \tag{2.15}$$

para $\xi \neq 0$, y

$$x_T = u + \beta \ln(Tn_y) \tag{2.16}$$

si $\xi = 0$.

Para estimar los niveles de retorno se requiere que se sustituyan los parámetros estimados $\hat{\xi}$ y $\hat{\beta}$. Como no se tienen la misma cantidad de observaciones de excedencias cada año y se supone que se tiene una serie estacionaria, se suele estimar n_y con k/N, donde k es el número total de excedencias observadas y N el número de años registrados.

Al igual que en el modelo de Máximos por Bloque se puede realizar la gráfica de niveles de retorno, la cual consiste en graficar los periodos de retorno contra sus niveles de retorno asociados, con escala logarítimica en el eje de las abscisas. Esta gráfica también representa una herramienta útil para diagnósticar el ajuste de las estimaciones, para ello se ubican los puntos de la forma

$$\left(T(p_i), x_{(i)}\right),$$

donde las p_i asociadas a cada observación se determinan de la misma forma ya mencionada, $T(p_i) = 1/(n_y p_i)$ y $x_{(i)}$ es la i-ésima excedencia de las observaciones ordenadas $x_{(1)} \leq \ldots \leq x_{(k)}$. Esta gráfica puede interpretarse del mismo modo que la gráfica de niveles de retorno de la DGVE, si los puntos graficados se asemejan a la curva de nivel de retorno se tiene que el modelo estimado representa un buen ajuste a las observaciones.

Estimación de la cola de F

Habiendo estimado la distribución condicional de los excesos, es posible dar una estimación de la cola de la distribución desconocida F. Para ello nos fijamos en la expresión para F_u en términos de la función de distribución F de X

$$F_u(x-u) = \frac{F(x) - F(u)}{1 - F(u)},$$

para $x > u$, de donde

$$\overline{F}(x) = \overline{F}(u)\overline{F}_u(x-u). \tag{2.17}$$

Para obtener una estimación reemplazamos en (2.17) $\overline{F}_u$ por $\overline{H}_{\xi,\beta}$ con los parámetros estimados. Por otro lado se puede estimar a $\overline{F}(u)$ con $1 - \tilde{F}_n(u) = k/n$, la cual corresponde a la

cola de la primera función de distribución empírica definida en 2.1 evaluada en u, donde k es el número de excedencias sobre u y n el total de observaciones, lo cual nos da la expresión siguiente.

$$\widehat{\overline{F}(x)} = \frac{k}{n}\left(1+\hat{\xi}\left(\frac{x-u}{\hat{\beta}}\right)\right)^{-1/\hat{\xi}},$$

si $\xi \neq 0$, y

$$\widehat{\overline{F}(x)} = e^{-(x-u)/\hat{\beta}},$$

si $\xi = 0$.

Diagnóstico del modelo DGP

Para verificar que el modelo estimado explique adecuadamente la información utilizamos las herramientas de diagnóstico antes descritas las cuales son las gráficas de probabilidades, de cuantiles, de densidad y de nivel de retorno. Supongamos que tenemos una muestra ordenada de excesos $y_{(1)}, \ldots, y_{(k)}$ sobre un umbral u y $\hat{H}$ es el correspondiente modelo DGP estimado, dado por

$$\hat{H}(y) = 1 - \left(1+\frac{\hat{\xi}y}{\hat{\beta}}\right)^{-1/\hat{\xi}},$$

para $\xi \neq 0$, para $\xi = 0$ se usa la expresión correspondiente. Entonces la gráfica de probabilidades consiste en los puntos

$$\left\{\left(p_i, \hat{H}(y_{(i)})\right); i = 1, \ldots, k\right\}.$$

Por otro lado, la gráfica de cuantiles está conformada por los puntos

$$\left\{\left(\hat{H}^{\leftarrow}(p_i), y_{(i)}\right); i = 1, \ldots, k\right\},$$

donde

$$\hat{H}^{\leftarrow}(y) = u + \frac{\hat{\beta}}{\hat{\xi}}\left(y^{-\hat{\xi}} - 1\right).$$

Y finalmente, se puede comparar la función de densidad ajustada $\hat{f}$ al histograma de los datos. Las gráficas de probabilidades y de cuantiles tienen la misma interpretación que en el modelo DGVE, donde ambas deben asemejarse a una recta diagonal si el modelo representa una buena aproximación.

Recapitulación

Recapitulando lo visto a lo largo de este capítulo tenemos que el proceso de ajustes de valores extremos, en síntesis, puede visualizarse mediante los siguientes diagramas de flujo.

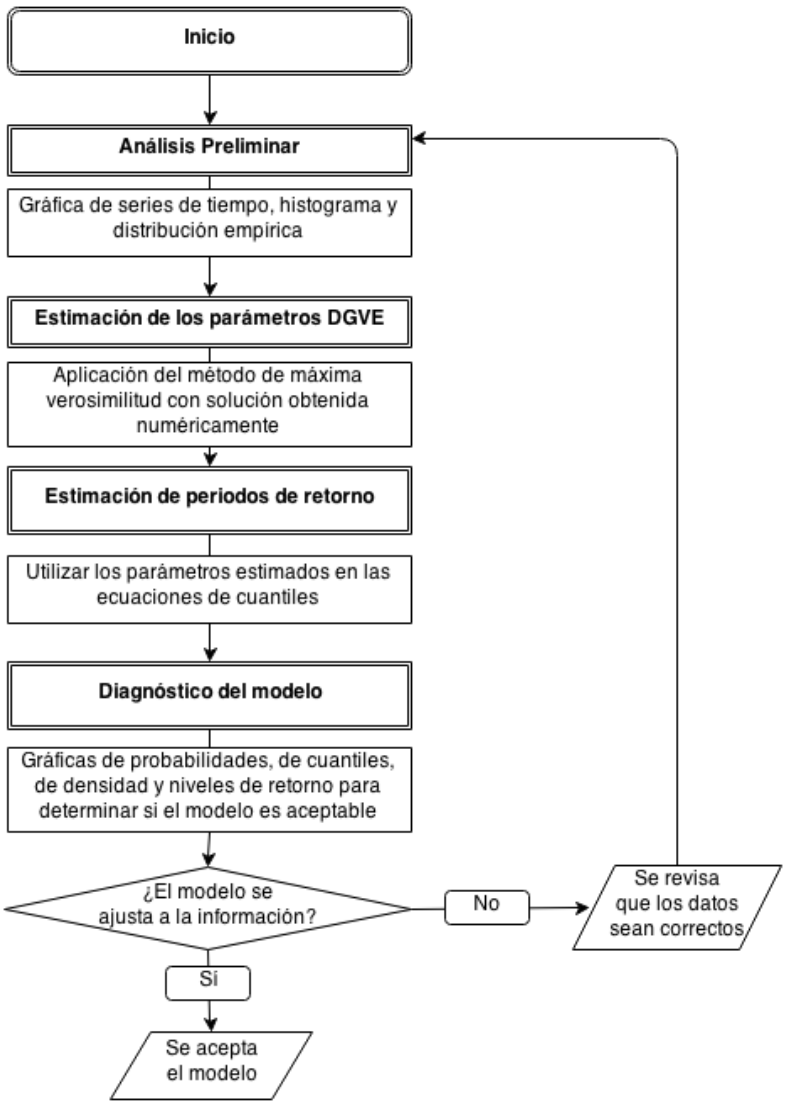

Figura 2.6: Diagrama de flujo del procedimiento seguido para estimar un modelo de máximos anuales a DGVE.

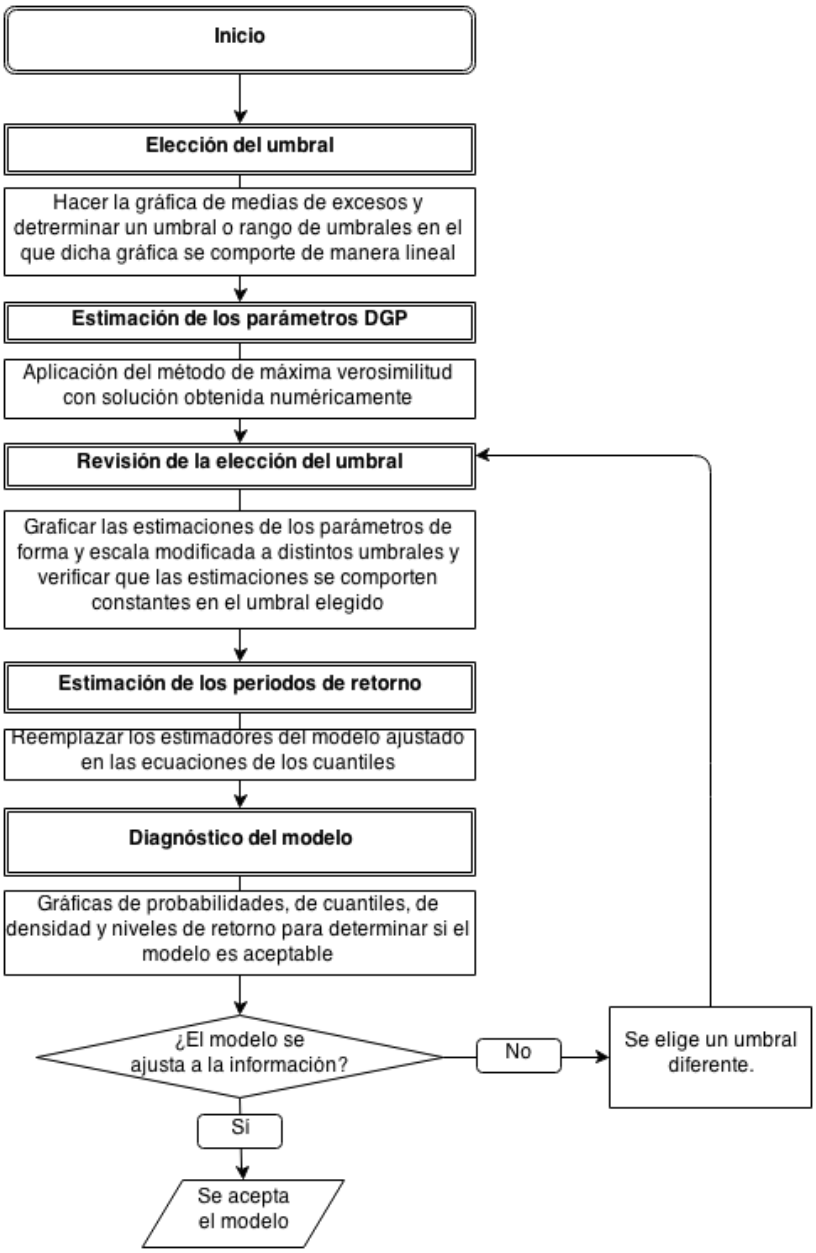

Figura 2.7: Diagrama de flujo del procedimiento seguido para estimar un modelo excesos sobre un umbral a DGP.

Capítulo 3

Aplicación de la TVE en el análisis de eventos hidrometeorológicos.

En el presente capítulo se aplicarán los resultados y herramientas revisados en los capítulos anteriores para el análisis del comportamiento de los valores extremos de los datos, los cuales provienen del Atlas de Clima Marítimo de la Vertiente Atlántica Mexicana [24], proporcionados por el Instituto de Ingeniería de la UNAM, que consisten en una base de datos horarios de la altura de ola en metros y velocidad de viento en metros sobre segundo, que corresponden a los resultados del modelo numérico híbrido WAM-HURAC, presentado por Ruiz-Martínez *et al.* en [22] para el periodo del 1 de enero de 1948 al 31 de diciembre de 2010, en la celda ubicada en las coordenadas 19°25' Latitud Norte, 91°45' Longitud Oeste cuya ubicación se muestra en el mapa de la Figura 3.1.

Se realizó un análsis de los dos fenómenos antes mencionados bajo los dos modelos principales de la Teoría de Valores Extremos en el caso univariado, los cuales son el modelo de Máximos por Bloque y el de Picos sobre un Umbral, sustentados por los Teoremas de Fisher-Tippet-Gnedenko y el de Pickans-Balkema-de Haan respectivamente.

Las estimaciones realizadas, así como las gráficas fueron realizadas en *R*, el cual es un

entorno de programación para análisis estadístico y gráfico. Es un proyecto de software libre que puede descargarse en http://cran.r-project.org/. Los paquetes estadísticos utilizados fueron *evd, evir, ismev, POT, stats y ADGofTest*, así como funciones propias programadas en el mismo entorno. Puede encontrarse en el Apéndice B el código del programa utilizado en este proceso con algunos comentarios como guía para el mismo. Para mayor información sobre este software y su uso pueden consultarse algunas guías y manuales, por ejemplo en [16] o [28].

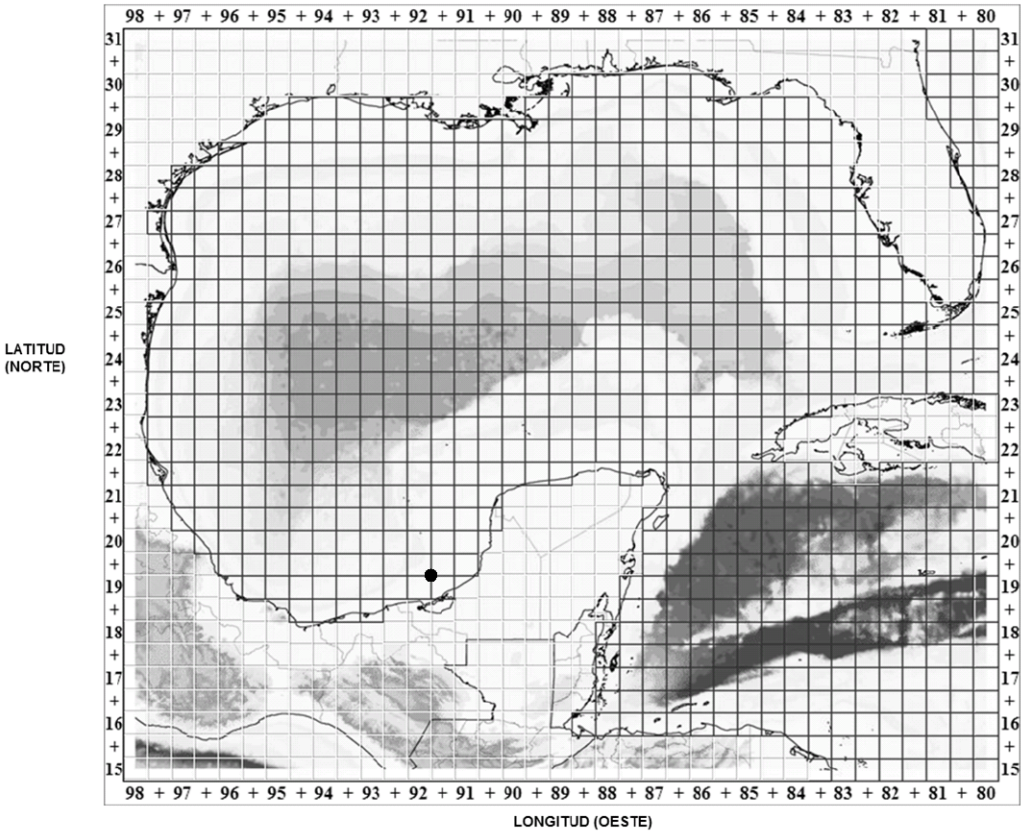

Figura 3.1: Localización geográfica de los datos.

3.1. Análisis Exploratorio

Como primer acercamiento exploratorio a los datos tenemos las gráficas de las series de tiempo de la altura de ola (en metros) y de la velocidad de viento (en metros sobre segundo). Esto nos permite tener una perspectiva inicial del comportamiento de las observaciones a

través del tiempo.

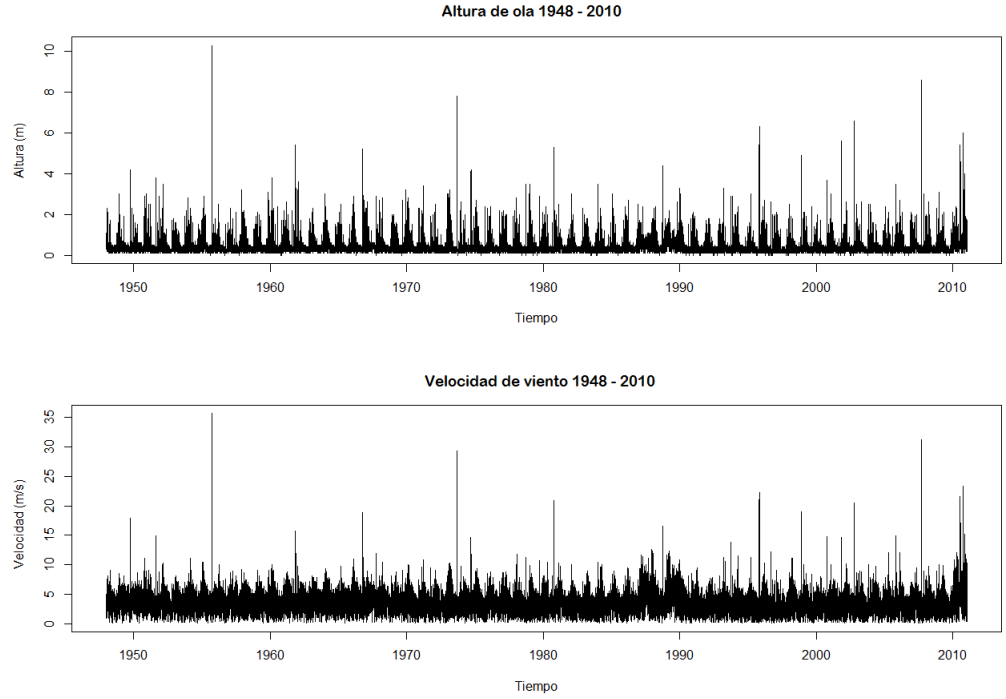

Figura 3.2: Altura de Ola y Velocidad de viento de 1948 a 2010.

En la figura 3.2 podemos observar la existencia de valores altos en comparación a aquellos que representan la mayoría o la media. Otra observación que puede hacerse a partir de esta gráfica es que valores altos de altura de ola corresponden a valores altos de velocidad de viento. Puede realizarse un análisis de dependencia y de distribución conjunta de estas variables pero son temas que se encuentran más allá del propósito de este trabajo, en el que se analizarán las variables extremas por separado de forma univariada.

El siguiente acercamiento a la información que podemos hacer es sobre la frecuencia de ocurrencia de las diferentes intensidades observadas así como su distribución aproximada, esto mediante un histograma y la función de distribución empírica respectivamente. Los histogramas se realizaron eligiendo celdas de longitud 0.1 para la altura de ola y 0.5 para la velocidad de viento. En la figura 3.3 se puede observar que la mayoría de las observaciones se encuentran conglomeradas alrededor de una media, pero también se refleja la existencia

de valores muy grandes en las colas de su distribución.

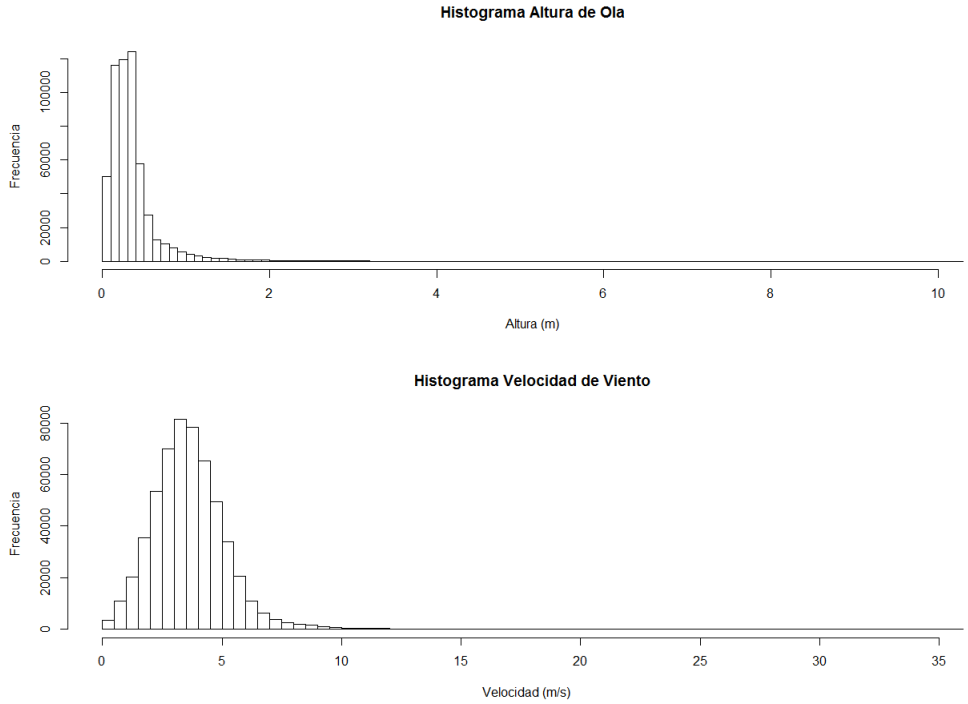

Figura 3.3: Histograma de frecuencias Altura de Ola y Velocidad de Viento.

Posteriormente, otro acercamiento a la forma de la distribución de los datos se puede hacer mediante las funciones de distribución empíricas. En las gráficas de la Figura 3.4 podemos observar que parece haber suficientes datos en la cola de la distribución para poder hacer un análisis de la distribución de aquellos valores que se encuentran alejados de la media.

3.2. Análisis de los datos bajo modelos de Valores Extremos

Procedemos ahora a analizar ambas muestras bajo los dos modelos descritos en este trabajo que consisten en el modelo de máximos por Bloque, relacionado con la Distribución Generalizada de Valores Extremos por el Teorema de Fisher-Tippet-Gnedenko, y el modelo de Picos sobre un Umbral, relacionado con la Distribución Generalizada de Pareto por el Teorema de Pickands-Balkema-de Haan. Para estos dos métodos utilizados en este trabajo

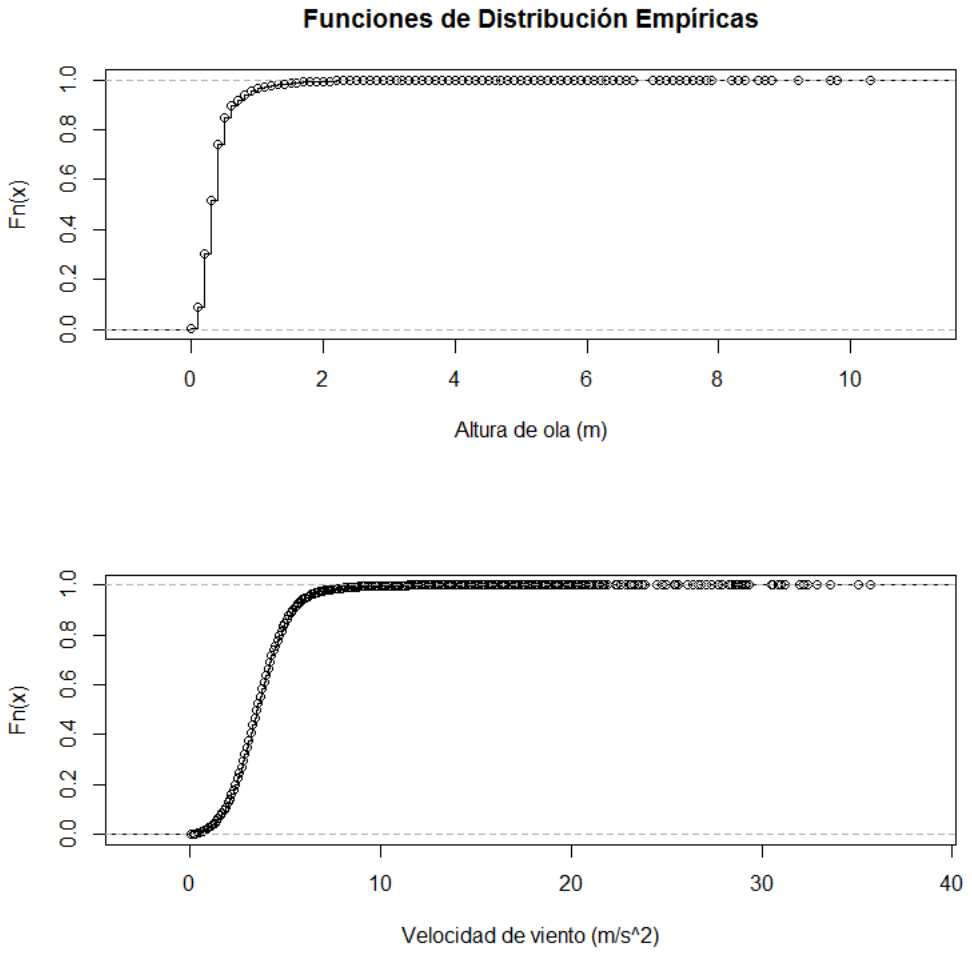

Figura 3.4: Distribuciones empíricas de la altura de ola y velocidad de viento.

se consideraron algunos criterios tanto para la elección del tamaño del bloque como para el umbral sobre el cual se tomarían las observaciones como máximos, los cuales se describen más adelante en sus respectivas secciones. Primero se realizarán las consideraciones para hacer las estimaciones de los parámetros por el método de máxima verosimilitud y posteriormente se harán diagnósticos de los modelos ajustados utilizando las herramientas descritas en el capítulo anterior.

3.2.1. Máximos por Bloque

Como se mencionó anteriormente, el Teorema de Fisher-Tippet-Gnedenko provee un modelo límite para la distribución del máximo valor obtenido por el método de Máximos por Bloque, tomando las intensidades máximas de bloques de igual tamaño. Al aplicar este modelo, la elección del tamaño de los bloques resulta de gran importancia. Si se eligen bloques muy pequeños, se corre el riesgo de no tener suficientes observaciones por bloque, haciendo pobre la aproximación por el Teorema 1.1; si se eligen bloques muy grandes, se obtienen pocas observaciones máximas y se obtienen varianzas estimadas grandes. En el caso de la altura de ola y la velocidad de viento no se puede suponer independencia de los datos, pero sí que se trata de un proceso estacionario, en cuyo caso los resultados límite siguen representando aproximaciones razonables. Puede consultarse más información al respecto en [5], capítulo 5.

Usualmente se utilizan bloques anuales, ya que se considera más robusto que un análisis donde se usen bloques de menor longitud temporal. En el análisis de fenómentos hidrometeorológicos es conveniente analizar los máximos anuales, pues si se toman bloques de menor tamaño éstos se ven afectados por la variabilidad de las condiciones que afectan la intensidad de dichos fenómenos a lo largo del año con el cambio de las estaciones. Con estas consideraciones tenemos que las observaciones de máximos anuales de nuestras observaciones constituyen una muestra aleatoria de variables con distribución generalizada de valores extremos, y se procede a estimar los parámetros de dicha distribución por máxima verosimilitud. Los estimadores de cada parámetro y los intervalos de confianza al 95 % de los mismos se muestran en la Tabla 3.1.

		Estimadores	**Intervalos**
Altura	$\hat{\xi}$	0.33256	(0.12465 , 0.54033)
	$\hat{\mu}$	2.78268	(2.57314 , 2.99243)
	$\hat{\sigma}$	0.75476	(0.57444 , 0.93522)
Velocidad	$\hat{\xi}$	0.60683	(0.35321 , 0.86024)
	$\hat{\mu}$	9.99688	(9.48751 , 10.50709)
	$\hat{\sigma}$	1.82127	(1.29228 , 2.35093)

Tabla 3.1: Estimadores e Intervalos de Confianza al 95 % de los parámetros de la DGVE.

En la práctica se busca ajustar el modelo más sencillo posible. Es por ello que en ocasiones se busca ajustar una distribución generalizada de valores extremos con $\xi = 0$ a los máximos anuales, es decir, se busca que dichos máximos sigan una distribución Gumbel. En las estimaciones mostradas en la Tabla 3.1 podemos ver que en ambas variables el parámetro de forma ξ es positivo, lo cual nos indica que la distribución de los máximos anuales de altura de ola y velocidad de viento siguen una distribución del tipo Fréchet. Y analizando los intervalos de confianza podemos notar que ninguno de los estimados para ξ contiene al cero, por lo que podemos descartar el ajuste de máximos anuales a Gumbel.

Una vez hechas las estimaciones, realizamos un diagnóstico del ajuste del modelo mediante las gráficas de Probabilidad, de Cuantiles y sobre el historial de los máximos para comprobar que las estimaciones corresponden a un modelo adecuado. También se incluyen las gráficas de nieveles de retorno en este diagnóstico. En las Figuras 3.5 y 3.6 se puede observar que se trata de modelos adecuados para describir a los máximos anuales de altura de ola y velocidad de viento, ya que en las gráficas de probabilidad y de cuantiles los datos graficados asemejan una línea recta con su contraparte ajustada. También observamos que la curvas de densidades de la DGVE estimada se ajustan bien a los histogramas de frecuencias de las observaciones máximas. Por último, en las gráficas del niveles de retorno, vemos que los datos se apegan a las curvas del modelo estimado.

Además de las herramientas gráficas de diagnóstico del modelo, como recurso adicional se realizaron pruebas de bondad de ajuste para probar si las estimaciones realizadas explican correctamente el comportamiento de los máximos de los datos. Se realizaron para cada uno de los ajustes las pruebas de Kolmogorov-Smirnov y Anderson-Darling (Ver Apéndice A), obteniendo los estadísticos (D y A respectivamente) y p-values de la Tabla 3.2, en la que podemos ver a un nivel de significancia $\alpha = 0.05$, o 95 % de confianza, que ambas pruebas no rechazan el modelo estimado. Por útlimo incluimos las funciones de distribución de los máximos anuales de altura G_a y velocidad de viento G_v en las ecuaciones (3.1) y (3.2) respectivamente, así como la expresión de sus cuantiles $x_a(p)$ y $x_v(p)$ asociados al periodo de retorno $T(p) = 1/p$ en (3.3) y (3.4).

$$G_a(x) = \exp\left\{-\left(1 + \left(\frac{x - 2.78268}{0.75476}\right)\right)^{-1/0.33256}\right\} \tag{3.1}$$

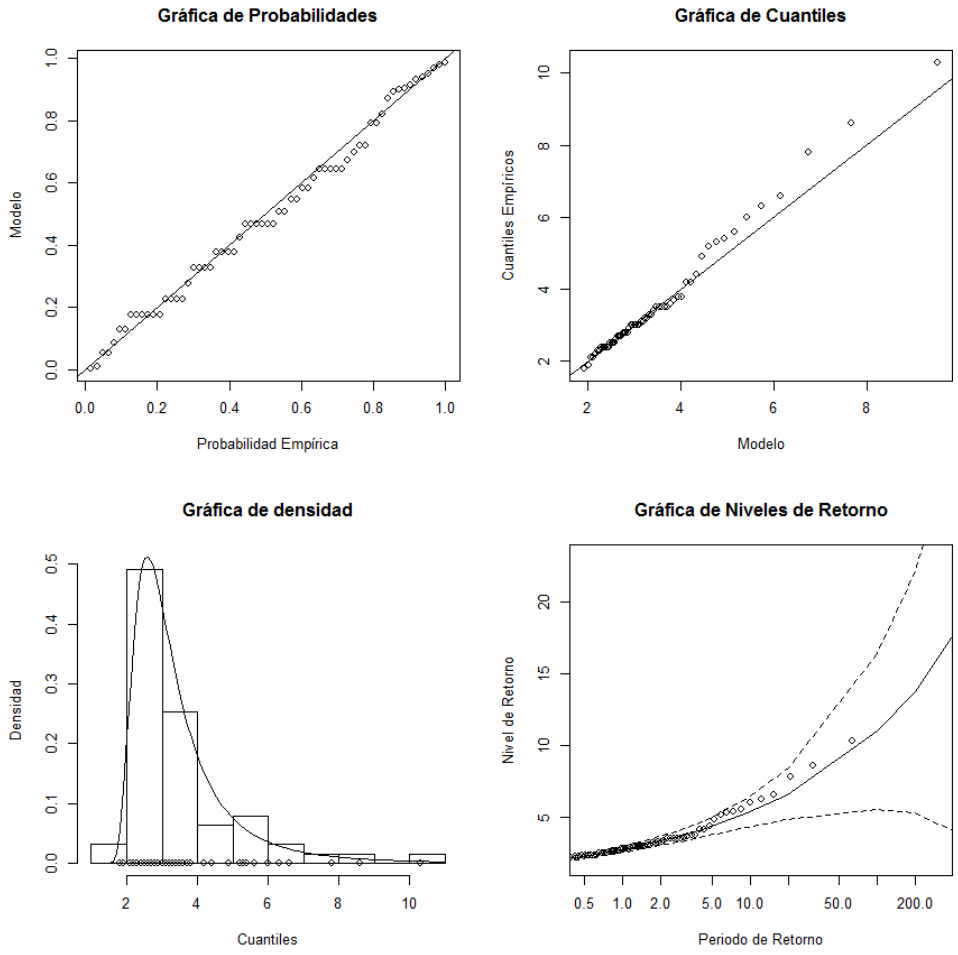

Figura 3.5: Gráficas de diagnóstico del ajuste de Máximas Alturas de Ola Anuales.

$$G_v(x) = \exp\left\{-\left(1+\left(\frac{x-9.99688}{1.82127}\right)\right)^{-1/0.60683}\right\} \tag{3.2}$$

$$x_a(p) = 2.78268 + \frac{0.75476}{0.33256}\left((-\ln(1-p))^{-0.33256} - 1\right) \tag{3.3}$$

$$x_v(p) = 9.99688 + \frac{1.82127}{0.60683}\left((-\ln(1-p))^{-0.60683} - 1\right) \tag{3.4}$$

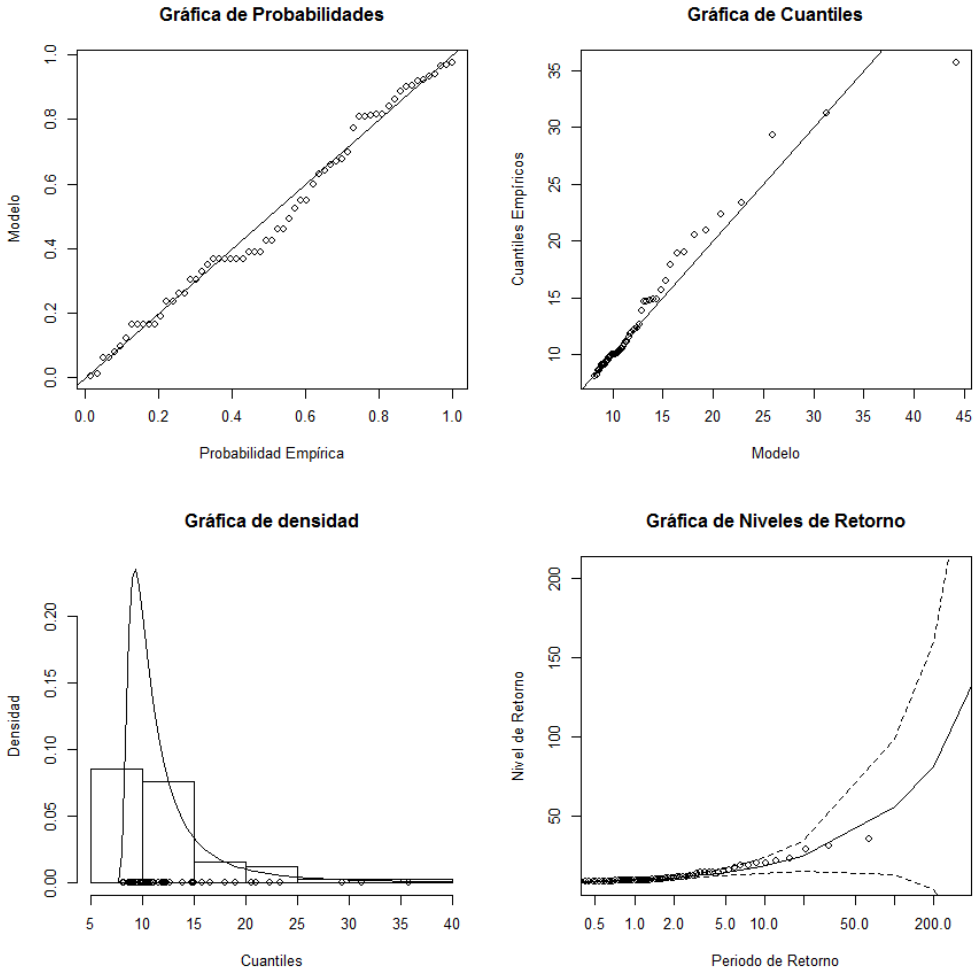

Figura 3.6: Gráficas de diagnóstico del ajuste de Máximas Velocidades de Viento Anuales.

3.2.2. Excedentes sobre un umbral

El primer paso para iniciar el análisis con el modelo de Picos sobre un Umbral es seleccionar el umbral u sobre el cual se considerarán las excedencias. Como se mencionó anteriormente, la herramienta con la que podemos comenzar este análisis es la gráfica de media de excesos, la cual debe mostrar un comportamiento lineal cerca del umbral adecuado para realizar el ajuste. Se grafica entonces la esperanza o media de los excesos para un rango de valores de u como

$$\left\{ \left(u, \frac{1}{k} \sum_{i=1}^{k} (x_{\langle i \rangle} - u) \right) : u < x_{(n)} \right\}.$$

	Estadístico	p-value
Altura	D = 0.0689	0.9261
	A = 0.3004	0.9377
Velocidad	D = 0.0878	0.7159
	A = 0.41	0.8382

Tabla 3.2: Estadísticos y p-values de las pruebas de bondad de Ajuste Kolmogorov-Smirnov (D) y Anderson-Darling (A) de la estimación de los parámetros de la DGVE.

A primera vista podemos observar en la figura 3.7, en la gráfica de medias de excesos de la altura de ola que la función se comienza a comportar de manera aproximadamente lineal alrededor del nivel de umbral $u = 2$. En un ajuste anteriormente realizado se eligió un umbral igual a 1 con el que se compararon los resultados obtenidos. Para el caso de la Velocidad de viento, en su respectiva gráfica se observa que el nivele que podría ser adecuado es $u = 9$, donde la gráfica parece empezar a ser lineal aproximadamente.

Sabemos que en series de datos que provienten de eventos hidrometeorológicos representan series de tiempo con alta correlación. Para evadir hasta cierto punto dicha correlación al momento de "extraer" los extremos de la base, se detectaron y redujeron los grupos de excedencias cercanas en tiempo con valores similares. A estos grupos se les conoce como *clústers*, y la manera de detectarlos y eliminarlos es la siguiente. Cuando un valor excede el umbral se considera el inicio de un clúster, y a todos los valores siguientes que excedan el umbral y se encuentren a un tiempo específico de distancia de la primera excedencia se les considera dentro del mismo. El clúster se considera terminado en el momento en el que se encuentre un valor que no excede el umbral, o la condición del tiempo termine. Este tiempo se especifica para asegurar un cierto grado de independencia entre las observaciones que serán consideradas como extremos, y para efectos de este trabajo, en la detección de clústers se consideró una condición temporal de cinco días.

Hechas estas consideraciones, lo siguiente es analizar las estimaciones de los parámetros de forma ξ y escala modificada β^*, descrito en el capítulo anterior, a diferentes valores de umbrales. Como se mencionó anteriormente, un umbral adecuado para el ajuste será aquel valor para el cual las gráficas de estimación de los parámetros de forma y escala modificada

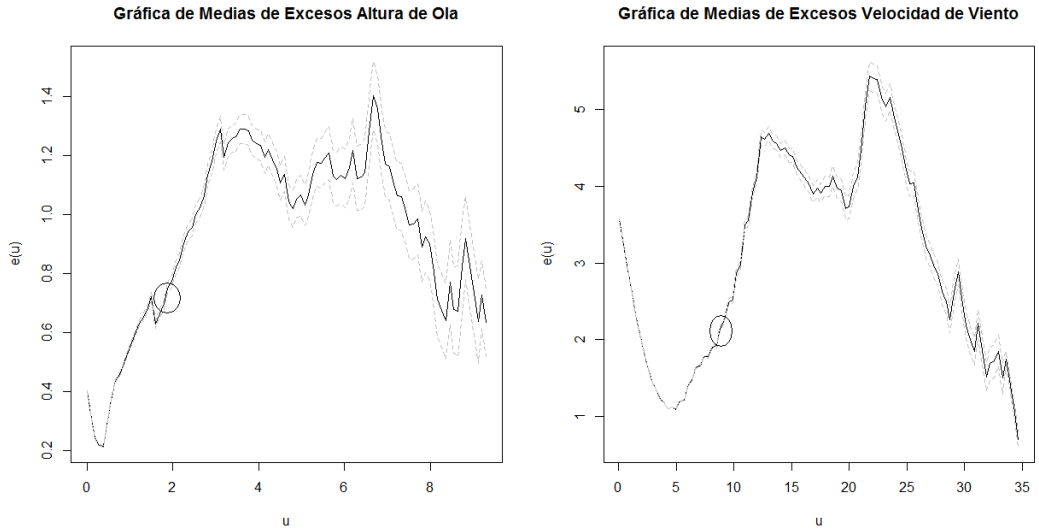

Figura 3.7: Funciones de Medias de Excesos de Altura de Ola (derecha) y Velocidad de Viento (izquierda).

se comporten aproximadamente constantes. Si hay un rango de umbrales para los que hay estabilidad, se elige preferentemente el menor de ellos.

Podemos observar en la Figura 3.8 que las estimaciones de los parámetros para la DGP de altura de ola se comportan aproximadamente constantes cerca de $u = 2.5$, tanto la gráfica de los estimadores de β^* como los de ξ, por lo que tomaremos dicho umbral. Notemos que no hay evidencia de que en $u = 1$ se tenga un comportamiento constante.

En el caso de la velocidad de viento, podemos observar en la Figura 3.9 que las estimaciones se muestran constantes a partir de $u = 9$. Estos valores de umbrales se asemejan a los considerados mediante la gráfica de medias de excesos y son los que utilizaremos para obtener los estimadores por el método de máxima verosimilitud de la DGP, los cuales se muestran en la Tabla 3.3, así como los intervalos de confianza de los mismos y el número de excedencias N.

Se incluyeron también los resultados de la estimación considerando el umbral $u = 1$ con fines comparativos, a pesar de que este umbral no cumple con las características que podrían apreciarse en la gráfica de medias de exceso ni en la gráfica de estimaciones bajo diferentes niveles de umbral. Este umbral fue elegido como 1.5 veces la altura de ola cuadrática signifi-

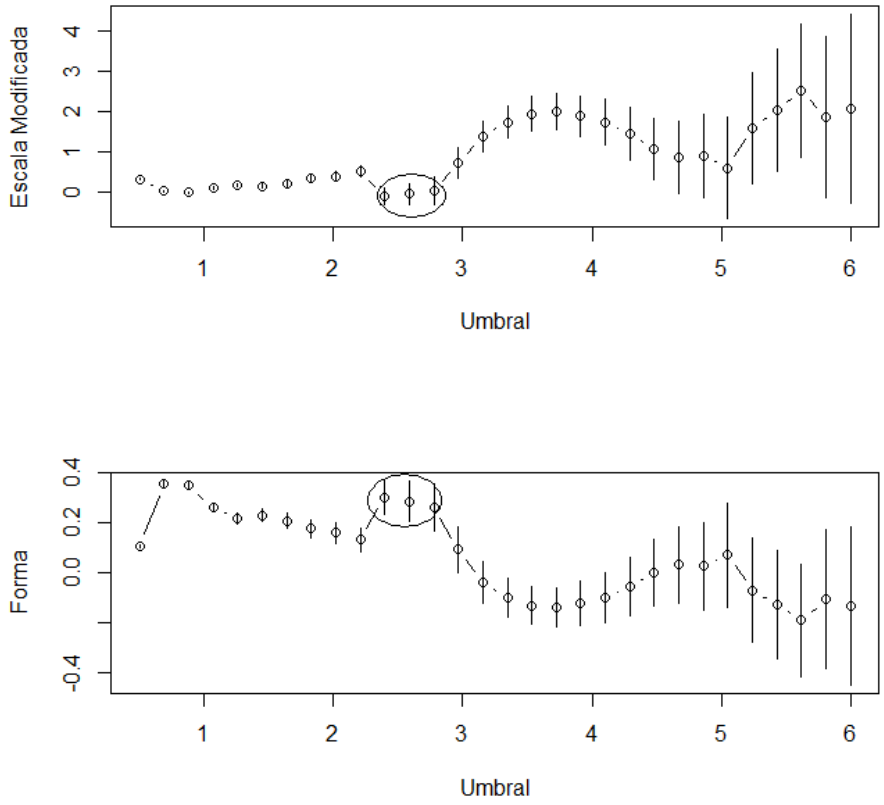

Figura 3.8: Estimaciones de los parámetros de la DGP de escala modificada (arriba) y de forma (abajo) para distintos umbrales de altura de ola.

cante media (Hrms) que es considerada como condiciones de tormenta de alta energía y, por tanto, es tomada como el umbral que define una tormenta (Silva R, comunicación personal).

Una vez hechas las estimaciones, procedemos a hacer el diagnóstico del ajuste de Picos sobre Umbrales de altura de ola y velocidad de viento mediante las herramientas gráficas anteriormente descritas.

Observamos que en las gráficas de diagnóstico del ajuste con umbral $u = 1$ en la Figura 3.10 que el modelo estimado para la altura de ola parece no dar una muy buena aproximación, ya que a pesar de tener cierta semejanza con las rectas diagonales en el caso de las gráficas

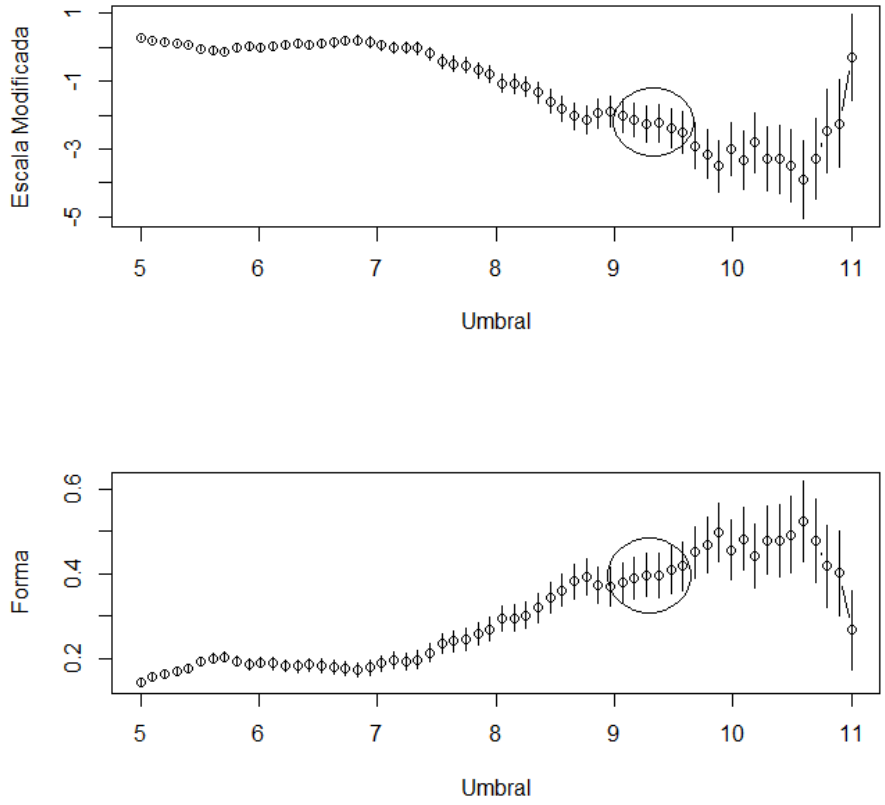

Figura 3.9: Estimaciones de los parámetros de la DGP de escala modificada (arriba) y de forma (abajo) para distintos umbrales de velocidad de viento.

de probabilidades y de cuantiles, y con la curva de niveles de retorno, podemos encontrar que hay valores de la cola de la distribución que se alejan demasiado, al punto de quedar incluso fuera de los intervalos de confianza. A este ajuste, al igual que a los otros, se le realizaron las pruebas de bondad de ajuste de Kolmovorov-Smirnov y Anderson-Darling para comprobar si explica el comportamiento de los excesos sobre el umbral establecido.

Por otro lado, podemos ver en las gráficas de diagnóstico del ajuste para un umbral $u = 2.5$ en la Figura 3.11 que los datos se asemejan más a la línea recta diagonal tanto de las gráficas de probabilidades y de cuantiles, y que los puntos de niveles de retorno se asemejan a su vez a la curva del ajuste. En cuanto a las gráficas de diagnóstico de la Figura 3.12

		N	$\hat{\theta}$	**Estimadores**	**Intervalos**
Altura	$u = 1$	724	$\hat{\beta}$	0.74942	(0.67997 , 0.82419)
			$\hat{\xi}$	0.05732	(0.00340 , 0.12328)
	$u = 2.5$	80	$\hat{\beta}$	0.8603	(0.60025 , 1.21998)
			$\hat{\xi}$	0.2547	(0.01716, 0.59393)
Velocidad	$u = 9$	147	$\hat{\beta}$	1.6470	(1.25685 , 2.13528)
			$\hat{\xi}$	0.3989	(0.20885 , 0.64763)

Tabla 3.3: Estimadores e Intervalos de Confianza al 95 % de los parámetros de la DGP.

	Umbral	Estadístico	p-value
Altura	$u = 1$	D = 0.1245	3.614e-10
		A = 10.1741	5.688e-06
	$u = 2.5$	D = 0.1082	0.3058
		A = 1.1557	0.2849
Velocidad	$u = 9$	D = 0.0729	0.4155
		A = 1.0004	0.3569

Tabla 3.4: Estadísticos y p-values de las pruebas de bondad de Ajuste Kolmogorov-Smirnov (D) y Anderson-Darling (A) de la estimación de los parámetros de la DGP.

pertenecientes al ajuste realizado a las excedencias sobre $u = 9$, podemos apreciar también la similitud que tienen los datos con las estimaciones del ajuste realizado. Adicionalmente se realizaron las pruebas de bondad de ajuste antes mencionadas, de las cuales se obtuvieron los estadísticos y p-values que se muestran en la Tabla 3.4, en la que podemos notar que el modelo ajustado para $u = 1$ se rechaza en ambas pruebas a un nivel de confianza del 95 %, por lo que lo descartamos. Por otro lado tenemos que para el ajuste con un umbral de $u = 2.5$ ninguna de las pruebas rechaza la hipótesis de que las excedencias se distribuyen como una DGP al mismo nivel de confianza. En cuanto al modelo de velocidad de viento, también observamos que ambas pruebas dan evidencia de no rechazar la hipótesis de la distribución de los datos con los parámetros estimados.

Por último, tenemos las expresiones de las distribuciones generalizadas de pareto ajustadas para los picos sobre umbrales para altura de ola H_a y velocidad de viento H_v en las

ecuaciones (3.5) y (3.6) respectivamente, y en las ecuaciones (3.7) y (3.8) la expresión de sus respectivos cuantiles en años. Como la base comprende de 1948 a 2010 se consideraron 63 años para la estimación de $n_y = N/63$, donde N es el número de excedencias de cada variable.

$$H_a(x) = 1 - \left(1 + 0.2547\left(\frac{x - 2.5}{0.8603}\right)\right)^{-1/0.2547}, \quad x > 2.5 \tag{3.5}$$

$$H_v(x) = 1 - \left(1 + 0.3989\left(\frac{x - 9}{1.6470}\right)\right)^{-1/0.3989}, \quad x > 9 \tag{3.6}$$

$$x_a(T) = 2.5 + \frac{0.8603}{0.2547}\left(\left(\frac{80}{63}T\right)^{0.2547} - 1\right) \tag{3.7}$$

$$x_v(T) = 9 + \frac{1.6470}{0.3989}\left(\left(\frac{147}{63}T\right)^{0.3989} - 1\right) \tag{3.8}$$

En el presente capítulo se aplicaron los resultados y herramientas revisados en los capítulos anteriores para el análisis del comportamiento de los valores extremos de los datos bajo el método de Máximos por Bloque aproximando la distribución de los máximos anuales con la Distribución Generalizada de Valores Extremos; y el método de Picos sobre un Umbral, aproximando la distribución de los excesos sobre un umbral u con la Distribución Generalizada de Pareto. Notamos la importancia de hacer un análisis cuidadoso en particular en el modelo de Picos sobre un Umbral, ya que una elección arbitraria pordía implicar la obtención de un modelo poco adecuado a la información, por lo que no se recomienda que este tipo de procesos se realice de manera automatizada ([5]).

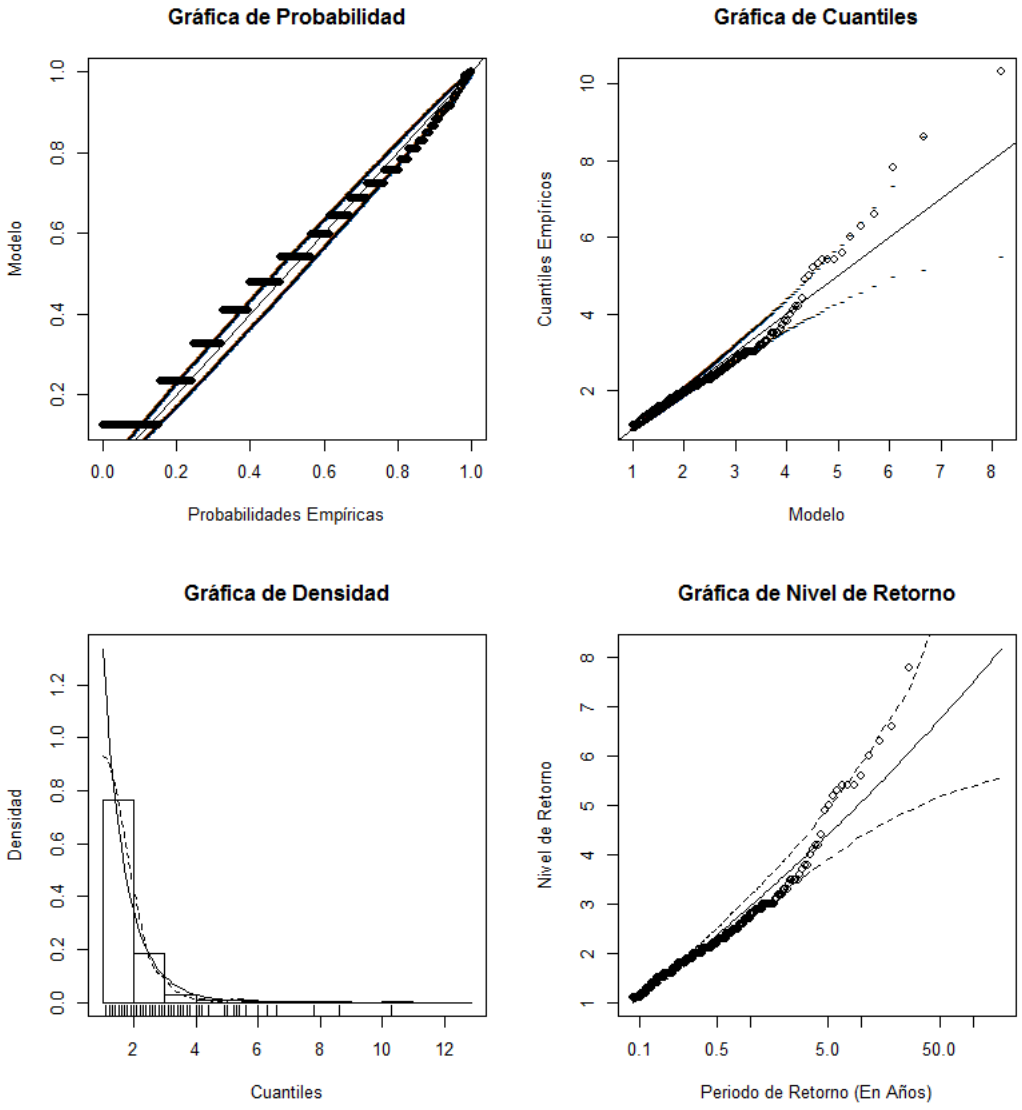

Figura 3.10: Gráficas de diagnóstico del ajuste de alturas de ola por encima de un umbral $u = 1$.

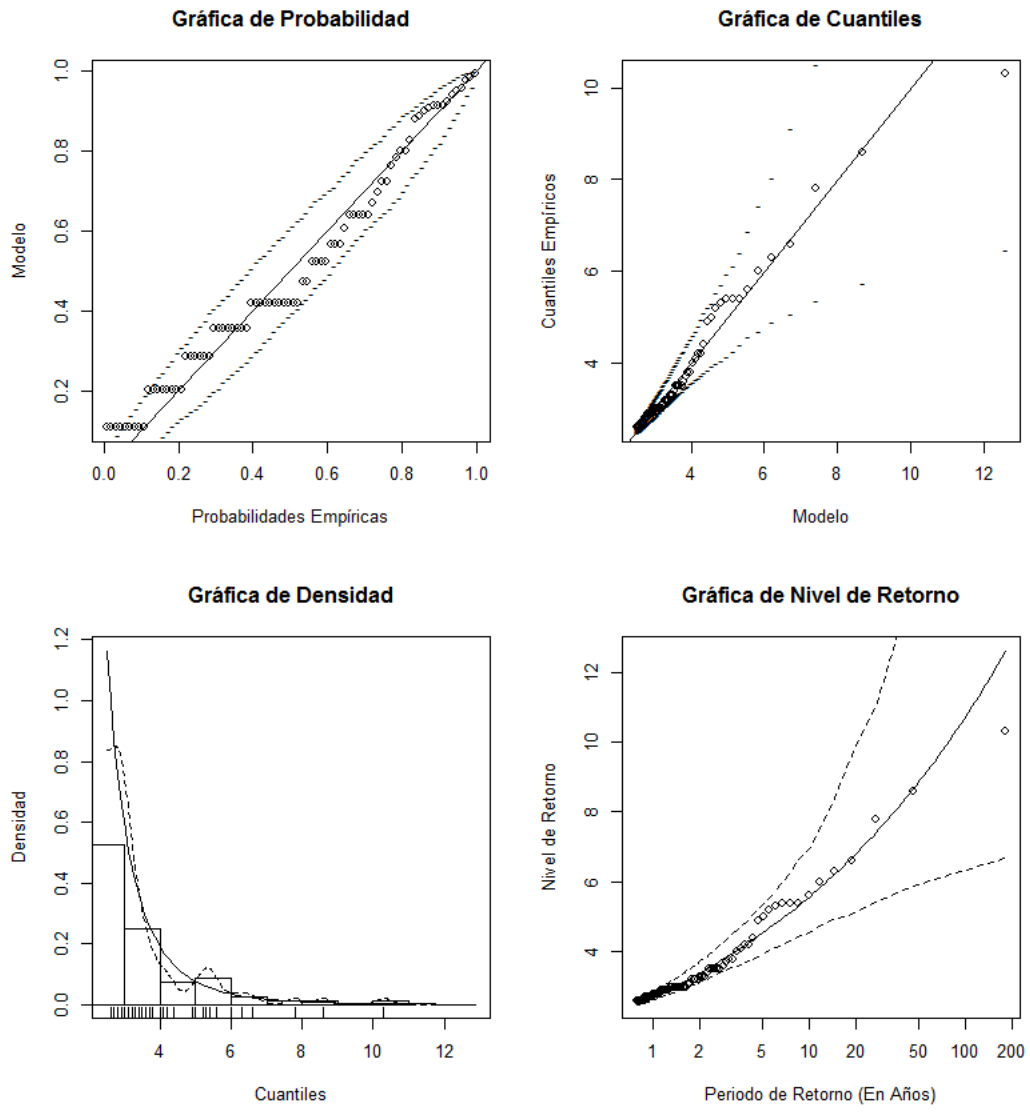

Figura 3.11: Gráficas de diagnóstico del ajuste de alturas de ola por encima de un umbral $u = 2.5$.

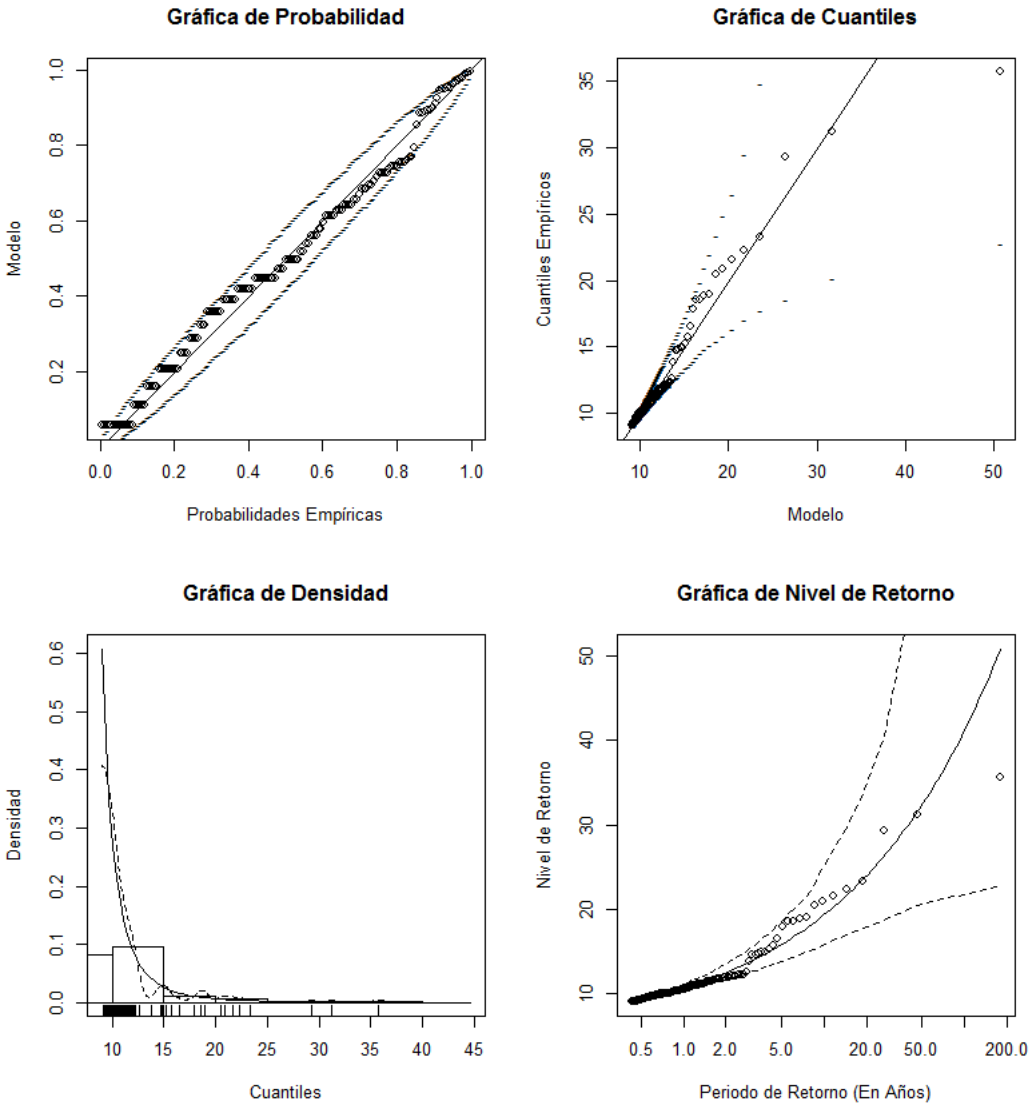

Figura 3.12: Gráficas de diagnóstico del ajuste de velocidades de viento por encima de un umbral $u = 9$.

Conclusiones

El propósito del presente trabajo fue saber si se podían mejorar los ajustes que se habían realizado previamente de los datos que fueron proporcionados por el Instituto de Ingeniería de la UNAM. Para ello se presentaron los principales resultados de la Teoría de Valores Extremos así como algunas de las herramientas gráficas y estadísticas que se utilizan en el ajuste de modelos bajo este enfoque probabilístico, con el fin de hacer inferencia sobre aquellos eventos que se consideran como extremos. Se incluyeron también algunos resultados adicionales como la caracterización del dominio de atracción de la DGVE así como las propiedades principales de la DGP, las cuales respaldaron el uso de dichas herramientas.

Entre los principales resultados que se vieron están el Teorema de Fisher-Tippet-Gnedenko, el cual brinda como distribución límite la Distribución Generalizada de Valores Extremos para el modelo de Máximos por Bloque; y el Teorema de Pickands-Balkema-de Hann el cual sustenta que la Distribución Generalizada de Pareto surge como distribución límite para los excesos sobre umbrales altos.

Todos estos resultados se utilizaron para dar una aproximación a la distribución de los valores extremos registrados de los eventos hidrometeorológicos de altura de ola y velocidad de viento de una zona cercana a la costa de Campeche bajo los criterios de los dos principales modelos de la teoría de valores extremos los cuales son Máximos por Bloque y Picos sobre un Umbral. Se espera que este trabajo sea de utilidad en el análisis extremal en otras zonas.

A lo largo del análisis realizado podemos destacar que debido al cuidado que debe ponerse al momento de realizar estimaciones en cuantiles extremos, este tipo de procesos no deberían automatizarse, ya que a pesar de ser factible el realizar un buen ajuste, se corre el riesgo de obtener resultados poco adecuados por no haber realizado un análisis cuidadoso. En la

inferencia para la DGVE, la elección de una familia límite es de suma importancia por la diferencia que representa en el comportamiento de las colas de las distribuciones determinado por el parámetro de forma ξ, el cual hace agresivamente diferentes las estimaciones de cuantiles como niveles de retorno. Puede evitarse realizar esta elección antes de realizar la inferencia mediante la expresión de la DGVE, la cual unifica las tres distribuciones límite posibles. Por otro lado, para la aproximación de la distribución de los excesos por una DGP la elección de un umbral adecuado representa un proceso que debe realizarse cuidadosamente. Esto se reflejó en la comparación que se hizo de los ajustes del modelo de Picos sobre un Umbral para el caso de altura de ola, realizados con un análisis utilizando las herramientas descritas en esta tesis, contra el análisis que se tenía previamente hecho. En este último se eligió un umbral que si bien no era arbitrario, no fue elegido bajo los criterios propuestos.

Para la comparación se utilizaron herramientas de diagnóstico como las gráficas de probabilidades y de cuantiles, la gráfica de densidad y la gráfica de niveles de retorno, en las cuales se pudo observar que el modelo realizado usando los resultados de la TVE representaba un mejor ajuste a la información, al dar una estimación mejor apegada a los datos que el análisis previo. Al revisar la elección de los umbrales, también vimos en la gráfica de estimaciones de los parámetros que el umbral anteriormente utilizado hacía que las estimaciones no contaran con las propiedades que se esperarían de una distribución generalizada de Pareto. Por último, se realizaron pruebas de bondad de ajuste que constataron que el ajuste anterior no representaba un modelo adecuado a un 95 % de confianza, y en cambio el ajuste nuevo representaba un modelo aceptable al mismo nivel de confianza.

Como continuación de este trabajo, puede realizarse un análisis multivariado, que permita realizar estimaciones de distribuciones de probabilidad conjunta de extremos, pero esos temas se encuentran más allá de los propósitos de este trabajo, donde se buscó hacer una introducción simple pero formal a los procesos de estimación de distribuciones extremas.

Se espera que los resultados arrojados de este análisis coadyuven a una mejor cuantificación de los riesgos hidrometeorológicos para la optimización de la toma de decisiones y gestión de áreas susceptibles este tipo de riesgos con el fin de evitar pérdidas humanas, minimizar las pérdidas económicas y el daño, y mejorar las medidas preventivas ante desastres naturales. Para ello se incluye en el Anexo B el código que se utilizó para realizar el análisis,

con el fin de que sea útil en el estudio de otras zonas.

México cuenta con un extenso litoral y por ende presenta alta vulnerabilidad ante los ciclones tropicales tanto de la vertiente pacífica como atlántica. Una de las instituciones que realiza más contribuciones en evaluación de riesgos hidrometeorológicos es el Centro Nacional de Desastres (CENAPRED), que ha publicado guías y directrices para la construcción de atlas de riesgos, para el cual un factor importante es el conocimiento, o en este caso estimación de las probabilidades de que se presenten eventos potencialmente dañinos. Se espera que con trabajos como el presente se impulse el desarrollo y el uso de herramientas teóricas recientes y sofisticadas para obtener mejores análisis y evaluaciones de modelos basados en valores extremos para la prevención y toma de decisiones en materia de riesgo hidrometeorológico.

Apéndice A

Herramientas adicionales

En esta sección se detallan algunas definiciones y resultados usados en algunas secciones de esta tesis.

A.1. Convergencia

En el presente trabajo se hace mención a diferentes conceptos referentes a convergencia, que a continuación se definen.

DEFINICIÓN A.1. **(Convergencia Débil o en Distribución)**. *Una sucesión de variables aleatorias $X_1, X_2, \ldots$, con funciones de distribución $F_1, F_2, \ldots$ respectivamente, se dice que converge en distribución a la variable aleatoria X, con función de distribución F si*

$$F_n(x) \to F(x)$$

cuando $n \to \infty$ para todos los puntos de continuidad x de F. Y se denota por $X_n \xrightarrow{\mathrm{d}} X$. Si F es una función continua, entonces la sucesión de funciones F_n converge uniformemente a F, es decir

$$\lim_{n \to \infty} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = 0.$$

DEFINICIÓN A.2. **(Convergencia Fuerte o Casi Segura)**. *Se dice que una sucesión de variables aleatorias $X_1, X_2, \ldots$ converge casi seguramente a una variable aleatoria X (Puede ser degenerada a una constante) si*

$$\mathbb{P}(\lim_{n \to \infty} X_n = X) = 1$$

y se denota por $X_n \xrightarrow{c.s.} X$.

DEFINICIÓN A.3. **(liminf, limsup).** *Sea x_n una sucesión de reales extendidos, es decir $x_n \in \overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. Se define*

$$\begin{aligned} \liminf_{n\to\infty} x_n &= \sup_n \inf_{k\geq n} x_k \\ \limsup_{n\to\infty} x_n &= \inf_n \sup_{k\geq n} x_k. \end{aligned}$$

Se pueden interpretar al límite inferior y superior como el menor y mayor límite convergente de la sucesión x_n respectivamente.

Teorema A.1. (Propiedades de liminf y limsup). *Sea x_n una sucesión de reales, entonces*

a) $\limsup_{n\to\infty} x_n = \lim_{n\to\infty} \left(\sup_{k\geq n} x_k\right)$

b) $\liminf_{n\to\infty} x_n = \lim_{n\to\infty} \left(\inf_{k\geq n} x_k\right)$

c) $\liminf_{n\to\infty} x_n \leq \limsup_{n\to\infty} x_n$

d) $x_n \to x$ *si y sólo si* $\liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n = x$

Un Teorema importante del que se hace mención y uso en este trabajo es el Teorema de Gilvenko-Cantelli, el cual justifica la aproximación a la función de distribución real de una sucesión de variables aleatorias independientes e idénticamente distribidas mediante la función de distribución empírica.

DEFINICIÓN A.4. **(Función de Distribución Empírica)** *Sea $X_1, X_2, \ldots$ una sucesión de variables aleatorias independientes e idénticamente distribuidas con función de distribución F. Se define a la Función de Distribución Empírica $\tilde{F}_n$ para $X_1, X_2, \ldots, X_n$ como*

$$\tilde{F}_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{(-\infty,x]}(X_i).$$

Teorema A.2. (Glivenko-Cantelli) *Sea $X_1, X_2, \ldots, X_n$ una colección de variables aleatorias independientes con distribución común F y sea $\tilde{F}_n$ la función de distribución empírica correspondiente. Entonces*

$$\sup_x |\tilde{F}_n(x) - F(x)| \to 0,$$

casi seguramente cuando $n \to \infty$.

A.2. Funciones Inversas Generalizadas

Las funciones inversas generalizadas, también llamadas funciones cuantil, son una herramienta importante utilizada a lo largo de este trabajo. A continuación se definen y se enuncian algunas propiedades de éstas.

DEFINICIÓN A.5. *Supongamos que F es una función no decreciente en $\mathbb{R}$. Con la convención de que el ínfimo de un conjunto vacío es $+\infty$ se define a la inversa generalizada de F o función cuantil de F como*

$$F^{\leftarrow}(y) := \text{ínf}\{s : F(s) \geq y\}.$$

Proposición A.1. (Propiedades de $F^{\leftarrow}$). *Sea F una función no decreciente continua por la derecha, y sea $F^{\leftarrow}$ su función inversa generalizada. Entonces $F^{\leftarrow}$ tiene las siguientes propiedades:*

a) Es monótona no decreciente.

b) Es continua por la izquierda.

c) $A(y) := \{s : F(s) \geq y\}$ es cerrado.

d) $F\left(F^{\leftarrow}(y)\right) \geq y$.

e) $\begin{cases} F^{\leftarrow}(y) \leq t & sii \quad y \leq F(t) \\ t < F^{\leftarrow}(y) & sii \quad y > F(t) \end{cases}$

f) Si $Y = \frac{X-b}{a}$, entonces $F_Y^{\leftarrow}(y) = \frac{F_X^{\leftarrow}(y)-b}{a}$

Demostración. a) Para ver que $F^{\leftarrow}$ es monótona no decreciente, tomamos $y_1, y_2 \in \mathbb{R}$ tal que $y_1 < y_2$. Entonces tenemos que

$$F^{\leftarrow}(y_1) = \text{ínf}\{s : F(s) \geq y_1\} \leq \text{ínf}\{s : F(s) \geq y_2\} = F^{\leftarrow}(y_2),$$

lo cual se da por la monotonía de F y por contención de conjuntos.

b) Para mostrar que $F^{\leftarrow}$ es continua por la izquierda, tomamos a $y_n \uparrow y$, pero suponemos que $\lim_{n\to\infty} F^{\leftarrow}(y_n) < F^{\leftarrow}(y)$. Entonces existe $\delta > 0$ y t tales que para toda n,

$$F^{\leftarrow}(y_n) < t < F^{\leftarrow}(y) - \delta. \tag{A.1}$$

Por la desigualdad de la izquierda, $t > \text{ínf}\{s : F(s) \geq y_n\}$ por lo que $F(t) \geq y_n$ para toda n y cuando $n \to \infty$, $F(t) \geq y$, que por la definición de $F^{\leftarrow}$ tenemos que $t \geq \text{ínf}\{s : F(s) \geq y\} = F^{\leftarrow}(y)$, lo cual contradice la desigualdad derecha en (A.1).

c) Para demostrar que $A(y)$ es cerrado, tomamos una sucesión $s_n \in A(y)$ para toda n, tal que $s_n \to s_0$. Supongamos que $s_0 \notin A(y)$, entonces por la definición de $A(y)$, $F(s_0) < y$. Sea $\epsilon = y - F(s_0) > 0$, y por la continuidad por la derecha de F, existe $\delta_0 > 0$ tal que si $s > s_0$ y $s - s_0 < \delta_0$ entonces $F(s) - F(s_0) < \epsilon = y - F(s_0)$, es decir

$$F(s) < y \quad \forall\, s \in (s_0, s_0 + \delta_0). \tag{A.2}$$

Por otro lado, tenemos que como $s_n \to s_0$, entonces existe n_0 tal que para toda $n \geq n_0$, $|s_n - s_0| < \delta_0$, es decir

$$s_0 - \delta_0 < s_n < s_0 + \delta_0, \quad \forall\, n \geq n_0. \tag{A.3}$$

Supongamos ahora que $s_n \leq s_0$ para alguna $n \geq n_0$, entonces $F(s_n) \leq F(s_0) < y$, lo cual implica que $s_n \notin A(y)$, que es una contradicción. Por lo tanto, $s_n > s_0$ para toda $n \geq n_0$, entonces, por las desigualdades en (A.3), $s_n \in (s_0, s_0 + \delta_0)$, que por (A.2) implica que $F(s_n) < y$ lo cual es de nuevo una contradicción. Por lo tanto $s_0 \in A(y)$ y éste es un conjunto cerrado. Además podemos hacer la observación que este conjunto es el intervalo $[F^{\leftarrow}(y), \infty)$, pues F es no decreciente.

d) Como $A(y)$ es un conjunto cerrado, entonces tenemos que $F^{\leftarrow}(y) = \text{ínf}\, A(y) \in A(y)$, lo cual implica que $F(F^{\leftarrow}(y)) \geq y$.

e) Podemos verificar que por la definición de $F^{\leftarrow}$ y del conjunto $A(y)$ que $F^{\leftarrow}(y) \leq t$ si y sólo si $t \in A(y)$ y esto ocurre si y sólo si $F(t) \geq y$. De manera similar observamos que $t < F^{\leftarrow}(y)$ si y sólo si $t \notin A(y)$ lo cual ocurre si y sólo si $F(t) < y$.

f) Como definimos a $Y = \frac{X-b}{a}$, entonces

$$F_Y(x) = \mathbb{P}(Y \leq x) = \mathbb{P}(X \leq ax + b) = F_X(ax + b)$$

por lo que

$$\begin{aligned} F_Y^{\leftarrow}(y) &= \text{ínf}\{s : F_Y(s) \geq y\} \\ &= \text{ínf}\{s : F_X(as + b) \geq y\} \\ &= \text{ínf}\left\{\frac{s - b}{a} : F_X(s) \geq y\right\} \\ &= \frac{\text{ínf}\{s : F_X(s) \geq y\} - b}{a} \\ &= \frac{F_X^{\leftarrow}(y) - b}{a} \end{aligned}$$

□

Usando la notación $\mathcal{C}(F)$ para el conjunto de puntos de continuidad de cualquier función monótona F tenemos la siguiente proposición.

Proposición A.2. *Si $\{F_n, n \geq 0\}$ es una sucesión de funciones no decrecientes y $F_n(x) \to F_0(x)$ para todo $x \in \mathcal{C}(F_0)$, entonces $F_n^{\leftarrow}(y) \to F_0^{\leftarrow}(y)$ para toda $y \in \mathcal{C}(F_0^{\leftarrow})$*

Demostración. Sea $\epsilon > 0$ y sea $y \in \mathcal{C}(F_0^{\leftarrow})$ (usaremos el hecho de haber tomado y como punto de continuidad de $F_0^{\leftarrow}$ en la última parte de la demostración). Como las discontinuidades de F_0 son a lo más numerables, entonces existe $x_1 \in (F_0^{\leftarrow}(y) - \epsilon, F_0^{\leftarrow}(y))$ tal que $x_1 \in \mathcal{C}(F_0)$, pues de otro modo este intervalo sería un conjunto de puntos de discontinuidad de F_0, el cual es no numerable. Como $x_1 < F_0^{\leftarrow}(y)$ entonces $F_0(x_1) < y$, por la definición de $F_0^{\leftarrow}$. Como $x_1 \in \mathcal{C}(F_0)$, entonces $F_n(x_1) \to F_0(x_1)$, así que para n grande, $F_n(x_1) < y$ que por la definición de $F_n^{\leftarrow}$, implica que $x_1 \leq F_n^{\leftarrow}(y)$, llegando entonces a que

$$F_0^{\leftarrow}(y) - \epsilon < x_1 \leq F_n^{\leftarrow}(y)$$

para n grande, y como ϵ fue elegido arbitrariamente,

$$F_0^{\leftarrow}(y) \leq \liminf_{n\to\infty} F_n^{\leftarrow}(y). \tag{A.4}$$

Tomemos ahora $\delta > 0$ y el hecho de que para $y' = y + \delta > y$ podemos tomar $x_2 \in \mathcal{C}(F_0)$ tal que $x_2 \in (F_0^{\leftarrow}(y'), F_0^{\leftarrow}(y') + \epsilon)$. Ya que $x_2 > F_0^{\leftarrow}(y')$ y por la definición de la inversa, tenemos que $F_0(x_2) \geq y' > y$. Como $x_2 \in \mathcal{C}(F_0)$ entonces $F_n(x_2) \to F_0(x_2)$, así que para n grande, $y \leq F_n(x_2)$ y entonces $F_n^{\leftarrow}(y) \leq x_2$, con lo cual llegamos a que

$$F_n^{\leftarrow}(y) \leq x_2 < F_0^{\leftarrow}(y') + \epsilon$$

para n grande, así que

$$\limsup_{n\to\infty} F_n^{\leftarrow}(y) \leq F_0^{\leftarrow}(y')$$

pues ϵ fue elegido arbitrariamente, y como δ también es arbitrario, y por la continuidad de $F_0^{\leftarrow}$ en y, legamos a que

$$\limsup_{n\to\infty} F_n^{\leftarrow}(y) \leq F_0^{\leftarrow}(y). \tag{A.5}$$

Ahora, por las desigualdades (A.4) y (A.5) tenemos que

$$\limsup_{n\to\infty} F_n^{\leftarrow}(y) \leq F_0^{\leftarrow}(y) \leq \liminf_{n\to\infty} F_n^{\leftarrow}(y)$$

lo cual demuestra que $F_n^{\leftarrow} \to F_0^{\leftarrow}$ para todo $y \in \mathcal{C}(F_0^{\leftarrow})$. □

A.3. Teorema de Convergencia a Tipos

El siguiente Teorema, propuesto por Khintchine, es un resultado importante que es de utilidad para caracterizar a las distribuciones max-estables, así como para la demostración del Teorema de Fisher, Tippet y Gnedenko.

DEFINICIÓN A.6. *Dos distribuciones son del mismo tipo o pertenecen a la misma familia si para algunas constantes* $a > 0$, $b \in \mathbb{R}$,

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

Teorema A.3. (Convergencia a tipos). *Sean* $X_1, X_2, \ldots$ *variables aleatorias con función de distribución* F_n *para cada* $n \geq 1$, *y sean* U, V *variables aleatorias con funciones de distribución* F_U *y* F_V *respectivamente (no degeneradas). Sean* $a_n > 0, \alpha_n > 0, b_n \in \mathbb{R}, \beta_n \in \mathbb{R}$ *constantes.*

a) Si

$$\begin{array}{rcl} F_n(a_n x + b_n) & \to & F_U(x) \ \forall\ x \in \mathcal{C}(F_U) \\ & y & \\ F_n(\alpha_n x + \beta_n) & \to & F_V(x) \ \forall\ x \in \mathcal{C}(F_V) \end{array} \tag{A.6}$$

o equivalentemente

$$\frac{X_n - b_n}{a_n} \xrightarrow{\mathrm{d}} U \quad y \quad \frac{X_n - \beta_n}{\alpha_n} \xrightarrow{\mathrm{d}} V \tag{A.7}$$

entonces existen constantes $A > 0$, $B \in \mathbb{R}$ *tales que:*

$$\lim_{n\to\infty} \frac{\alpha_n}{a_n} = A > 0 \quad y \quad \lim_{n\to\infty} \frac{\beta_n - b_n}{\alpha_n} = B \in \mathbb{R} \tag{A.8}$$

y

$$F_V(x) = F_U(Ax + B), \quad V \overset{d}{=} \frac{U - B}{A}. \tag{A.9}$$

b) Si se cumple (A.8)*, entonces cualquiera de las relaciones en* (A.6) *implica la otra, y* (A.9) *se cumple.*

Demostración. Comenzaremos demostrando el inciso *b*). Suponemos entonces (A.8) y que $G_n := F_n(a_n x + b_n) \to F_U(x)$ para todo $x \in \mathcal{C}(F_U)$. Entonces

$$\begin{aligned} G_n\left(\frac{\alpha_n}{a_n}x + \frac{\beta_n - b_n}{a_n}\right) &= F_n\left(a_n\left(\frac{\alpha_n}{a_n}x + \frac{\beta_n - b_n}{a_n}\right) + b_n\right) \\ &= F_n(\alpha_n x + \beta_n - b_n + b_n) \\ &= F_n(\alpha_n x + \beta_n) \end{aligned}$$

Sea $x \in \mathcal{C}(F_U(A \cdot + B))$ y supongamos que $x > 0$ (análogamente puede demostrarse para $x \leq 0$). Entonces, si tomamos $\epsilon > 0$, por (A.8) $\exists\, n_0$ tal que $\forall\, n \geq n_0$,

$$\left| \frac{\alpha_n}{a_n} - A \right| \leq \epsilon \quad , \quad \left| \frac{\beta_n - b_n}{a_n} - B \right| \leq \epsilon$$

entonces

$$A - \epsilon \leq \frac{\alpha_n}{a_n} \leq A + \epsilon \quad , \quad B - \epsilon \leq \frac{\beta_n - b_n}{a_n} \leq B + \epsilon$$

de donde

$$(A - \epsilon)x + (B - \epsilon) \leq \frac{\alpha_n}{a_n} x + \frac{\beta_n - b_n}{a_n} \leq (A + \epsilon)x + (B + \epsilon).$$

Tenemos entonces que

$$\begin{aligned} \limsup_{n \to \infty} F_n(\alpha_n x + \beta_n) &= \limsup_{n \to \infty} G_n \left(\frac{\alpha_n}{a_n} x + \frac{\beta_n - b_n}{a_n} \right) \\ &\leq \limsup_{n \to \infty} G_n((A + \epsilon)x + (B + \epsilon)) \end{aligned}$$

Por tanto, para todo $z \in \mathcal{C}(F_U)$ tal que $z > (A + \epsilon)x + (B + \epsilon)$ tenemos que

$$\limsup_{n \to \infty} F_n(\alpha_n x + \beta_n) \leq \limsup_{n \to \infty} G_n(z) = F_U(z),$$

dándose esta última igualdad por (A.6). En consecuencia,

$$\limsup_{n \to \infty} F_n(\alpha_n x + \beta_n) \leq \text{ínf} \{F_U(z) : z > (A + \epsilon)x + (B + \epsilon), z \in \mathcal{C}(F_U)\}$$

y como $\epsilon > 0$ fue tomado arbitrariamente, y por la continuidad por la derecha de F_U,

$$\limsup_{n \to \infty} F_n(\alpha_n x + \beta_n) \leq \text{ínf} \{F_U(z) : z > Ax + B, z \in \mathcal{C}(F_U)\} = F_U(Ax + B).$$

Por otro lado, tenemos que

$$\begin{aligned} \liminf_{n \to \infty} F_n(\alpha_n x + \beta_n) &= \liminf_{n \to \infty} G_n \left(\frac{\alpha_n}{a_n} x + \frac{\beta_n - b_n}{a_n} \right) \\ &\geq \liminf_{n \to \infty} G_n((A - \epsilon)x + (B - \epsilon)) \end{aligned}$$

Entonces, para todo $z \in \mathcal{C}(F_U)$ tal que $z < (A - \epsilon)x + (B - \epsilon)$ tenemos que

$$\liminf_{n \to \infty} F_n(\alpha_n x + \beta_n) \geq \liminf_{n \to \infty} G_n(z) = F_U(z),$$

y al ser $\epsilon > 0$ arbitrario, y $x \in \mathcal{C}(F_U(A \cdot + B))$,

$$\liminf_{n \to \infty} F_n(\alpha_n x + \beta_n) \geq \sup \{F_U(z) : z < Ax + B, z \in \mathcal{C}(F_U)\} = F_U(Ax + B).$$

Entonces tenemos que

$$\limsup_{n \to \infty} F_n(\alpha_n x + \beta_n) \leq F_U(Ax + B) \leq \liminf_{n \to \infty} F_n(\alpha_n x + \beta_n).$$

Por lo tanto,

$$F_n(\alpha_n x + \beta_n) \to F_U(Ax + B) = F_V(x)$$

para todo $x \in \mathcal{C}(F_U(A \cdot + B))$, o bien, para todo $x \in \mathcal{C}(F_V)$.

Para la demostración del inciso a) supongamos que $F_n(a_n x + b_n) \to F_U(x)$ y $F_n(\alpha_n x + \beta_n) \to F_V(x)$. Entonces, por la Proposición A.2 tenemos que

$$\begin{aligned} \tfrac{F_n^{\leftarrow}(y) - b_n}{a_n} &\to F_U^{\leftarrow}(y) \quad , \quad \forall\, y \in \mathcal{C}(F_U^{\leftarrow}) \\ \tfrac{F_n^{\leftarrow}(y) - \beta_n}{\alpha_n} &\to F_V^{\leftarrow}(y) \quad , \quad \forall\, y \in \mathcal{C}(F_V^{\leftarrow}) \end{aligned} \tag{A.10}$$

Como F_U y F_V son no degeneradas, y monótonas, podemos tomar dos puntos y_1, y_2 tales que

$$\begin{gathered} y_1, y_2 \in \mathcal{C}(F_U^{\leftarrow}) \cap \mathcal{C}(F_V^{\leftarrow}), \quad y_1 < y_2 \\ F_U^{\leftarrow}(y_1) < F_U^{\leftarrow}(y_2) \quad y \quad F_V^{\leftarrow}(y_1) < F_V^{\leftarrow}(y_2) \end{gathered}$$

(De otro modo, tendríamos que $\mathcal{C}(F_U^{\leftarrow})^C$ o $\mathcal{C}(F_U^{\leftarrow})^C$ sería no numerable). Por lo tanto, por (A.10) para $i = 1, 2$ tenemos que

$$\frac{F_n^{\leftarrow}(y_i) - b_n}{a_n} \to F_U^{\leftarrow}(y_i) \quad , \quad \frac{F_n^{\leftarrow}(y_i) - \beta_n}{\alpha_n} \to F_V^{\leftarrow}(y_i) \tag{A.11}$$

entonces, restando las expresiones para $i = 1$ de las de $i = 2$, obtenemos

$$\frac{F_n^{\leftarrow}(y_2) - F_n^{\leftarrow}(y_1)}{a_n} \to F_U^{\leftarrow}(y_2) - F_U^{\leftarrow}(y_1) > 0 \tag{A.12}$$

$$\frac{F_n^{\leftarrow}(y_2) - F_n^{\leftarrow}(y_1)}{\alpha_n} \to F_V^{\leftarrow}(y_2) - F_V^{\leftarrow}(y_1) > 0 \tag{A.13}$$

Dividiendo (A.12) entre (A.13) llegamos a que

$$\frac{\alpha_n}{a_n} \to \frac{F_U^{\leftarrow}(y_2) - F_U^{\leftarrow}(y_1)}{F_V^{\leftarrow}(y_2) - F_V^{\leftarrow}(y_1)} =: A > 0.$$

Además, de (A.11) obtenemos

$$\frac{F_n^{\leftarrow}(y_1) - b_n}{a_n} \to F_U^{\leftarrow}(y_1) \quad y \quad \frac{F_n^{\leftarrow}(y_1) - \beta_n}{a_n} = \frac{F_n^{\leftarrow}(y_1) - \beta_n}{\alpha_n} \cdot \frac{\alpha_n}{a_n} \to F_V^{\leftarrow}(y_1) A.$$

Y restando estas dos últimas expresiones tenemos que

$$\frac{\beta_n - b_n}{a_n} \to F_U^{\leftarrow}(y_1) - F_V^{\leftarrow}(y_1) A =: B.$$

De modo que (A.8) se cumple, y por el inciso b), como se cumplen (A.6) y (A.8), entonces también (A.9). □

El Teorema de Convergencia a Tipos nos dice que si una sucesión de funciones de distribución converge normalizada a un cierto tipo de distribución, entonces el uso de otras constantes de normalización diferentes no cambia el tipo de la distribución límite, y además, nos da una relación entre las constantes utlizadas.

A.4. Funciones Medibles

En esta sección se dan algunas definiciones y resultados sobre Borel medibilidad que se utilizaron en la demostración del Teorema de Fisher-Tippet-Gnedenko. El material fue principalmente consultado en [10] y [11].

DEFINICIÓN A.7. *Sea $\mathcal{S}$ una colección no vacía de subconjuntos de $\mathbb{R}$. Decimos que $\mathcal{S}$ es una sigma-álgebra (σ-álgebra) de subconjuntos de $\mathbb{R}$ si*

i) $\mathbb{R} \in \mathcal{S}$

ii) $A \in \mathcal{S} \Rightarrow A^c \in \mathcal{S}$

iii) $A_n \in \mathcal{S}, n \geq 1 \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{S}$

DEFINICIÓN A.8. *Sea $\mathcal{A}$ una colección no vaciá de subconjuntos de $\mathbb{R}$. La sigma-álgebra más pequeña que contiene a $\mathcal{A}$ es llamada la sigma-álgebra generada por $\mathcal{A}$ y se denota por $\sigma\{\mathcal{A}\}$. Es decir, $\sigma\{\mathcal{A}\}$ cumple*

i) $\mathcal{A} \subset \sigma\{\mathcal{A}\}$

ii) Si $\mathcal{S}$ es una sigma-álgebra tal que $\mathcal{A} \subset \mathcal{S}$ entonces $\sigma\{\mathcal{A}\} \subset \mathcal{S}$.

DEFINICIÓN A.9. ***(Sigma-álgebra de Borel)**. Sea $\mathcal{I}$ la clase de los subconjuntos abiertos de $\mathbb{R}$. La sigma-álgebra de Borel se define como la sigma-álgebra generada por $\mathcal{I}$ y se denota por $\mathcal{B}_{\mathbb{R}}$.*

Puede demostrarse que $\mathcal{B}_{\mathbb{R}}$ coincide con la sigma-álgebra generada por la clase de los intervalos abiertos con extremos en $\mathbb{R}$, la de los intervalos cerrados y por la de los intervalos semicerrados. A los elementos de $\mathcal{B}_{\mathbb{R}}$ se les conoce como borelianos, y algunos de ellos son: los conjuntos abiertos, los conjuntos cerrados, los conjuntos numerables y los intervalos de todo tipo ([10]).

DEFINICIÓN A.10. *Un conjunto Ω junto con una sigma-álgebra asociada $\mathcal{S}$, es decir a la pareja $(\Omega, \mathcal{S})$ se le conoce como espacio medible. Al espacio $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ se le conoce como el espacio de Borel.*

DEFINICIÓN A.11. ***(Borel medible)**. Sea $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ el espacio de Borel. Una función $f : \mathbb{R} \to \mathbb{R}$ se llama medible relativa a $\mathcal{B}_{\mathbb{R}}$ si $f^{-1}(A) \in \mathcal{B}_{\mathbb{R}}$ para todo A boreliano, donde $f^{-1}(A) = \{x \in \mathbb{R} : f(x) \in A\}$. En ese caso se dice que f es Borel medible.*

Proposición A.3. *Sea $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ el espacio de Borel. Entonces*

i) Si f y g son funciones Borel medibles y $\alpha \in \mathbb{R}$, entonces $\alpha f, f + g, fg$ y $1/f$ $(f \neq 0)$ son Borel medibles.

ii) Si f_n es una sucesión convergente de funciones Borel medibles entonces $\lim_{n\to\infty} f_n$ es Borel medible.

Para las demostraciones y más información pueden consultarse las referencias mencionadas, entre otras fuentes sobre Teoría de la Medida.

A.5. Ecuaciones Funcionales

En la prueba del Teorema de Fisher-Tippet-Gnedenko se hacen uso de algunas soluciones a ecuaciones de funciones aditivas, las cuales se presentan a continuación.

DEFINICIÓN A.12. *Una función $f : \mathbb{R} \to \mathbb{R}$ es aditiva si*

$$f(x + y) = f(x) + f(y), \quad x, y \in \mathbb{R}\,.$$

También se dice que f satisface la ecuación funcional de Cauchy.

Teorema A.4. [1] *Si $f : \mathbb{R} \to \mathbb{R}$ es aditiva y Borel medible, entonces $f(x) = cx$, donde $c = f(1) \in \mathbb{R}$.*

Corolario A.1. *Sea $f : \mathbb{R} \to \mathbb{R}$ una función Borel medible.*

a) Si $f(xy) = f(x) + f(y)$, entonces $f(x) = c \ln x$ para alguna $c \in \mathbb{R}$.

b) Si $f(xy) = f(x)f(y)$, entonces $f(x) = x^c$ para alguna $c \in \mathbb{R}$.

[1] El primer volumen de *Fundamenta Mathematicae* contiene dos artículos, uno escrito por Sierpiński y el otro por Banach, que dan pruebas diferentes a este Teorema.

c) Si $f(x+y) = f(x)f(y)$, entonces $f(x) = e^{cx}$ para alguna $c \in \mathbb{R}$.

Demostración. a) Haciendo un cambio de variable tenemos que

$$f(e^{x+y}) = f(e^x \cdot e^y) = f(e^x) + f(e^y),$$

entonces por el Teorema A.4 tenemos que $f(e^x) = cx$, que es lo mismo que $f(x) = f(e^{\ln x}) = c \ln x$.

b) En este caso tenemos con un cambio de variable que

$$\ln f(e^{x+y}) = \ln f(e^x \cdot e^y) = \ln \left(f(e^x) \cdot f(e^y)\right) = \ln f(e^x) + \ln f(e^y)$$

y entonces $\ln f(e^x) = cx$, luego $f(x) = \exp\{\ln f(e^{\ln x})\} = \exp\{c \ln x\} = x^c$.

c) Reducimos al caso *b)* haciendo

$$f(\ln(xy)) = f(\ln x + \ln y) = \ln x \cdot \ln y.$$

Entonces $f(\ln x) = x^c$, y $f(x) = f(\ln e^x) = e^{cx}$

□

A.6. Funciones de Variación Regular

La teoría de Funciones de Variación Regular fue desarrollada por J. Karamata en 1930 y gradualmente se volvió evidente el rol importante que desarrolla en Teoría de la Probabilidad y en particular en la Teoría de Valores Extremos.

A continuación se enuncian algunas herramientas utilizadas para la caracterización de los dominios de atracción de las distribuciones límite para máximos.

DEFINICIÓN A.13. *Una función medible $F : \mathbb{R}_+ \to \mathbb{R}_+$ es de variación regular en ∞ con índice ρ si para $x > 0$*

$$\lim_{t \to \infty} \frac{F(tx)}{F(t)} = x^\rho.$$

Y se denota por $F \in VR_\rho$. A ρ se le conoce como el exponente de variación.

Si $\rho = 0$ decimos que F es de variación lenta y dichas funciones generalmente se denotan por $\mathcal{L}(x)$. Si $F \in VR_\rho$, entonces $F(x)/x^\rho \in VR_0$, y escribiendo $\mathcal{L}(x) = F(x)/x^\rho$, notamos que siempre es posible representar una función de variación regular con índice ρ como $x^\rho \mathcal{L}(x)$.

El siguiente teorema es uno de los resultados centrales de la teoría de las funciones de variación regular.

Teorema A.5. ***(Karamata).***

(a) Si $\rho \geq -1$ *entonces* $F \in VR_\rho$ *implica que* $\int_0^x F(t)dt \in VR_{\rho+1}$ *y*

$$\lim_{x\to\infty} \frac{xF(x)}{\int_0^x F(t)dt} = \rho + 1$$

Si $\rho < 1$ *(o si* $\rho = 1$ *y* $\int_x^\infty F(t)dt < \infty$*) entonces* $F \in VR_\rho$ *implica que* $\int_x^\infty F(t)dt$ *es finita,* $\int_x^\infty F(t)dt \in VR_{\rho+1}$ *y*

$$\lim_{x\to\infty} \frac{xF(x)}{\int_x^\infty F(t)dt} = -\rho - 1$$

(b) Si F *satisface*

$$\lim_{x\to\infty} \frac{xF(x)}{\int_0^x F(t)dt} = \lambda \in (0,\infty)$$

entonces $F \in VR_{\lambda-1}$*. Si* $\int_x^\infty F(t)dt < \infty$ *y*

$$\lim_{x\to\infty} \frac{xF(x)}{\int_x^\infty F(t)dt} = \lambda \in (0,\infty)$$

entonces $F \in VR_{-\lambda-1}$*.*

Corolario A.2. ***(Representación de Karamata).*** $\mathcal{L}$ *es de variación lenta si y sólo si puede representarse como*

$$\mathcal{L}(x) = c(x)\exp\left\{\int_1^x \frac{\epsilon(t)}{t}dt\right\} \tag{A.14}$$

para $x > 0$ *donde* $c : \mathbb{R}_+ \to \mathbb{R}_+$*,* $\epsilon : \mathbb{R}_+ \to \mathbb{R}_+$ *y*

$$\lim_{x\to\infty} c(x) = c_0 \in (0,\infty) \tag{A.15}$$

$$\lim_{x\to\infty} \epsilon(x) = 0. \tag{A.16}$$

La demostración tanto del Teorema A.5 como la del corolario A.2 pueden encontrarse en [19].

Observación A.1. *Si* $F \in VR_\alpha$*, entonces* F *tiene la siguiente representación*

$$F(x) = c(x)\exp\left\{\int_1^x \frac{\delta(t)}{t}dt\right\} \tag{A.17}$$

donde $c(\cdot)$ *satisface* (A.15) *y* $\lim_{t\to\infty}\delta(t) = \alpha$*. Esto se obtiene del Corolario A.2 escribiendo* $F(x) = x^\alpha\mathcal{L}(x)$ *y usando la representación para* $\mathcal{L}$*.*

A.7. Pruebas de Bondad de Ajuste

Las pruebas de bondad de ajuste son técnicas estadísticas que sirven para determinar si un conjunto de datos proviene de una muestra aleatoria de una distribución o familia de distribuciones específica, a continuación se explican dos de las pruebas más utilizadas en esta tarea y que fueron utilizadas en este trabajo, las cuales son la prueba de Kolmogorov-Smirnov y la de Anderson-Darling.

Kolmogorov-Smirnov

La prueba de Kolmogorov-Smirnov está basada en la función de distribución empírica. Una ventaja que tiene esta prueba es que la distribución del estadístico usado no depende de la distribución que será puesta a prueba, aunque sólo puede usarse para distribuciones continuas y muestra mayor sensibilidad cerca del centro de la distribución que en las colas. Otra limitante que tiene es que necesita que todos los parámetros de la distribución a probar estén dados.

Esta prueba se define como sigue:
H_0 Los datos provienen de la distribución especificada
H_1 Los datos no provienten de la distribución especificada.
El estadístico de Kolmogorov-Smirnov está definido como

$$D = \max_{1 \leq i \leq n} |\tilde{F}_n(x_{(i)} - F_0(x_{(i)})|,$$

donde $\tilde{F}_n$ es la función de distribución empírica, F_0 es la distribución puesta a prueba, y $x_{(i)}$ es el i-ésimo valor de los datos ordenados $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.

La hipótesis nula (H_0) es rechazada a un nivel de significancia α (o nivel de confianza $1 - \alpha$) si el estadístico D es mayor que el valor crítico a dicho nivel, el cual puede ser obtenido de tablas. Muchos de los paquetes estadísticos contienen la información suficiente para proveer el $p - value$ de la prueba, el cual es la probabilidad de haber haber obtenido el resultado obtenido suponiendo cierta la hipótesis nula. Basta con comparar este valor con α para determinar la validez del modelo en prueba: si $p - value \leq \alpha$ se rechaza la hipótesis nula y por ende el modelo.

Anderson-Darling

La prueba de Anderson-Darling es una refinación de la prueba de Kolmogorov Smirnov, que es considerada una prueba más fuerte y da mayor peso a las colas de la distrubición. Su desventaja radica en que los valores críticos del estadístico usado dependen de la distribución que es puesta a prueba, y deben ser calculados para cada caso. Actualmente existen tablas de valores críticos para cierto tipo de distribuciones, así como paquetes estadísticos que realizan los cálculos para obtenerlos.

Esta prueba se define como sigue:
H_0 Los datos provienen de la distribución especificada
H_1 Los datos no provienten de la distribución especificada.
El estadístico de Anderson-Darling está definido como

$$A^2 = -n - S,$$

donde

$$S = \sum_{i=1}^{n} \frac{(2i-1)}{n} \left(\ln F_0(x_{(i)}) + \ln(1 - F_0(x_{(n+1-i)})) \right)$$

donde F_0 es la distribución puesta a prueba y $x_{(i)}$ es el i-ésimo valor de los datos ordenados $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.

El criterio para aceptar o rechazar la hipótesis nula es el mismo que en la prueba de Kolmogorov. De igual forma actualmente hay disponibles diversos paquetes estadísticos que no sólo calculan los valores críticos, sino también proveen del $p - value$ de la prueba, con el cual puede determinarse la validez de un modelo.

Apéndice B

Código R

En esta sección se incluye el código en R utilizado para realizar las estimaciones, las gráficas de diagnóstico y demás cálculos. R es un entorno de programación para análisis estadístico y gráfico que puede descargarse en http://cran.r-project.org/ de manera gratuita.

Los símbolos '#' se utilizan para comentar líneas sin afectar la ejecución del programa. Para mayor información sobre este software pueden consultarse algunas guías y manuales, por ejemplo *http://cran.r-project.org/doc/contrib/Owen-TheRGuide.pdf* o *http://cran.r-project.org/doc/manuals/R-intro.pdf*.

```
library(ismev)
library(POT)
library(evd)
library(evir)
library(stats)
library(ADGofTest)

#Se lee una base de datos y se almacena en la variable 'datos'
datos<-read.table("AT2682501925_cvs.txt", header = FALSE, dec = ".",
col.names=c("YEAR","MONTH","DAY","HOUR","V","DIR_W","H","DIR_H","TP",
"TM","TN","TO"))
#Ahora el dataframe "datos" contiene todas las variables y se puede
# acceder a ellas mediante el signo '$'
```

```
#La variables de interés son altura de ola (H) y velocidad de viento (V)
# aunque más variables son utilizadas.

#Gráficas de densidades, distribución y cuantil de la DGP

#dgpd(x, loc=0, scale=1, shape=0, log = FALSE) #Densidad
#pgpd(q, loc=0, scale=1, shape=0, lower.tail = TRUE) #Distribución
#qgpd(p, loc=0, scale=1, shape=0, lower.tail = TRUE) #Cuantil

#Funciones de distribución empírica (paquetería stats)
plot(ecdf(datos$H))
plot(ecdf(datos$V))

##### ANÁLISIS DE MÁXIMOS POR BLOQUE #####

#Extracción de los máximos anuales
velocidad_a<-c()
altura_a<-c()
for(i in 1948:2010)
{
velocidad_a[i-1947]<-max(subset(datos,YEAR==i)$V)
altura_a[i-1947]<-max(subset(datos,YEAR==i)$H)
}

#Estimación de los parámetros de la DGVE (Máxima Verosimilitud)
#Se puede hacer uso de cualquiera de los paquetes evd,evir o ismev,
# para este trabajo se probaron todas las funciones y se utilzaron
# las estimaciones arrojadas por la paquetería ismev

#Funciones para ajustar DGVE por máxima verosimilitud

#Estimación de los parámetros de la DGVE (paquete ismev)
```

```
gev.fit(altura_a)
gev.fit(velocidad_a)

#Intervalos de confianza (evd)
#Para los intervalos, usamos la función confint() de la paquetería evd

confint(fgev(altura_a))
confint(fgev(velocidad_a))

#Diagnóstico (ismev)
#Se realizaron algunas modificaciones en el código de las funciones:
#gev.pp
#gev.qq
#gev.his
#gev.rl
#Las cuales se encuentran dentro de la función 'gev.diag()',
# que grafica las herramientas de diagnóstico

#A continuación se encuentran los códigos de las funciones modificadas,
# las funciones originales fueron escritas por Janet E. Hefferman.
# El encargado de la paquetería ismev es Eric Gilleland
# <ericg@ucar.edu>
#Las funciones pueden compilarse para hacer que las funciones
# modificadas funcionen

### FUNCIONES MODIFICADAS ###
gev_pp<-function (a, dat)
{
    plot((1:length(dat))/length(dat), gevf(a, sort(dat)),
xlab = "Probabilidad Empírica",
        ylab = "Modelo", main = "Gráfica de Probabilidades")
    abline(0, 1)
}
```

```
gev_qq<-function (a, dat)
{
    plot(gevq(a, 1 - (1:length(dat)/(length(dat) + 1))), sort(dat),
        ylab = "Cuantiles Empíricos", xlab = "Modelo",
main = "Gráfica de Cuantiles")
    abline(0, 1)
}

gev_his<-function (a, dat)
{
    h <- hist(dat, plot = FALSE)
    if (a[3] < 0) {
        x <- seq(min(h$breaks), min(max(h$breaks),
(a[1] - a[2]/a[3] - 0.001)), length = 100)
    }
    else {
        x <- seq(max(min(h$breaks), (a[1] - a[2]/a[3] + 0.001)),
            max(h$breaks), length = 100)
    }
    y <- gev.dens(a, x)
    hist(dat, freq = FALSE, ylim = c(0, max(max(h$density), max(y))),
        xlab = "Cuantiles", ylab = "Densidad",
main = "Gráfica de densidad")
    points(dat, rep(0, length(dat)))
    lines(x, y)
}

gev_rl<-function (a, mat, dat)
{
    eps <- 1e-06
    a1 <- a
    a2 <- a
```

```
    a3 <- a
    a1[1] <- a[1] + eps
    a2[2] <- a[2] + eps
    a3[3] <- a[3] + eps
    f <- c(seq(0.01, 0.09, by = 0.01), 0.1, 0.2, 0.3, 0.4, 0.5,
        0.6, 0.7, 0.8, 0.9, 0.95, 0.99, 0.995, 0.999)
    q <- gevq(a, 1 - f)
    d <- t(gev.rl.gradient(a = a, p = 1 - f))
    v <- apply(d, 1, q.form, m = mat)
    plot(-1/log(f), q, log = "x", type = "n", xlim = c(0.5, 300),
        ylim = c(min(dat, q), max(dat, q)),
xlab = "Periodo de Retorno",
        ylab = "Nivel de Retorno")
    title("Gráfica de Niveles de Retorno")
    lines(-1/log(f), q)
    lines(-1/log(f), q + 1.96 * sqrt(v), lty=2)
    lines(-1/log(f), q - 1.96 * sqrt(v), lty=2)
    points(-1/log((1:length(dat))/(length(dat) + 1)), sort(dat))
}

gum_rl<-function (a, mat, dat)
{
    eps <- 1e-06
    a1 <- a
    a2 <- a
    a1[1] <- a[1] + eps
    a2[2] <- a[2] + eps
    f <- c(seq(0.01, 0.09, by = 0.01), 0.1, 0.2, 0.3, 0.4, 0.5,
        0.6, 0.7, 0.8, 0.9, 0.95, 0.99, 0.995, 0.999)
    q <- gevq(a, 1 - f)
    d1 <- (gevq(a1, 1 - f) - q)/eps
    d2 <- (gevq(a2, 1 - f) - q)/eps
    d <- cbind(d1, d2)
```

```
    v <- apply(d, 1, q.form, m = mat)
    plot(-1/log(f), q, log = "x", type = "n", xlim = c(0.1, 1000),
        ylim = c(min(dat, q), max(dat, q)),
xlab = "Periodo de Retorno",
        ylab = "Nivel de Retorno")
    title("Gráfica de Niveles de Retorno")
    lines(-1/log(f), q)
    lines(-1/log(f), q + 1.96 * sqrt(v), lty=2)
    lines(-1/log(f), q - 1.96 * sqrt(v), lty=2)
    points(-1/log((1:length(dat))/(length(dat) + 1)), sort(dat))
}

gev_diag<-function (z)
{
    n <- length(z$data)
    x <- (1:n)/(n + 1)
    if (z$trans) {
        oldpar <- par(mfrow = c(1, 2))
        plot(x, exp(-exp(-sort(z$data))), xlab = "Empírico",
            ylab = "Modelo")
        abline(0, 1)
        title("Gáfica de Probabilidad Residuales")
        plot(-log(-log(x)), sort(z$data), ylab = "Empírico",
            xlab = "Modelo")
        abline(0, 1)
        title("Residual Quantile Plot (Gumbel Scale)")
    }
    else {
        oldpar <- par(mfrow = c(2, 2))
        gev_pp(z$mle, z$data)
        gev_qq(z$mle, z$data)
        gev_his(z$mle, z$data)
  gev_rl(z$mle, z$cov, z$data)
```

```
    }
    par(oldpar)
    invisible()
}

gum_diag<-function (z)
{
    z$mle <- c(z$mle, 0)
    n <- length(z$data)
    x <- (1:n)/(n + 1)
    if (z$trans) {
        oldpar <- par(mfrow = c(1, 2))
        plot(x, exp(-exp(-sort(z$data))), xlab = "empirical",
            ylab = "model")
        abline(0, 1, col = 4)
        title("Gráfica de Probabilidad de Residuales")
        plot(-log(-log(x)), sort(z$data), xlab = "empirical",
            ylab = "model")
        abline(0, 1, col = 4)
        title("Residual Quantile Plot (Gumbel Scale)")
    }
    else {
        oldpar <- par(mfrow = c(2, 2))
        gev_pp(z$mle, z$data)
        gev_qq(z$mle, z$data)
        gev_his(z$mle, z$data)
  gum_rl(z$mle, z$cov, z$data)
    }
    par(oldpar)
    invisible()
}

###
```

```
#Compilado el código, usamos la función modificada gev_diag
# para las gráficas de diagnóstico

gev_diag(gev.fit(altura_a)) #Diagnóstico altura
gev_diag(gev.fit(velocidad_a)) #Diagnóstico velocidad

#Pruebas de bondad de ajuste (paquetes stats y ADGofTest
#Para ahorrar escritura, se programaron las funciones ks_t y ad_t
# las cuales hacen las pruebas de bondad de ajuste de Kolmogorov-
# Smirnov y la prueba de Anderson-Darling

ks_t<-function(datos,fitted)
{
if(class(fitted)[1]=="gev.fit")
ks.test(datos,pgev,xi=fitted$mle[3],mu=fitted$mle[1],
beta=fitted$mle[2])
else
ks.test(datos,pgpd,loc=fitted$threshold,scale=fitted$param[1],
shape=fitted$param[2])
}

ad_t<-function(datos,fitted)
{
if(class(fitted)[1]=="gev.fit")
ad.test(datos,pgev,xi=fitted$mle[3],mu=fitted$mle[1],
beta=fitted$mle[2])
else
ad.test(datos,pgpd,loc=fitted$threshold,scale=fitted$param[1],
shape=fitted$param[2])
}

#Se procede a compilar las puebas
```

```
#Pruebas para la altura
ks_t(altura_a,gev.fit(altura_a))
ad_t(altura_a,gev.fit(altura_a))
#Pruebas para la velocidad
ks_t(velocidad_a,gev.fit(velocidad_a))
ad_t(velocidad_a,gev.fit(velocidad_a))


##### ANÁLISIS DE EXCESOS SOBRE UN UMBRAL ###

#Gráficas de medias de excesos "Mean Residual Life Plot"
# para elección del umbral

mrlplot(datos$H,main="Gráfica de Medias de Excesos Altura de Ola",
xlab="u")
mrlplot(datos$V,main="Gráfica de Medias de Excesos Velocidad de Viento",
xlab="u")

#Detección y eliminación de clústers
#Creamos primero un data.frame con las variables "obs" de las
# observaciones y "time" una variable de tiempo poder aplicar
# la función cluster

ola<-data.frame(obs=datos$H,time=1:length(datos$H))
viento<-data.frame(obs=datos$V,time=1:length(datos$V))

#Con la función clust se detectan y eliminan los clusters,
# se sugiere usar el umbral sugerido por la gráfica de
# medias de excesos

altura_1<-clust(ola,u=1,tim.cond=5*24,clust.max=TRUE)
altura<-clust(ola,u=2,tim.cond=5*24,clust.max=TRUE)
velocidad<-clust(viento,u=9,tim.cond=5*24,clust.max=TRUE)
```

```
#Revisión del umbral por estabilidad de las estimaciones
#Estimaciones de los parámetros para altura de ola
# para valores de umbral diferentes.
par(mfrow=c(2,1))
tcplot(datos$H, u.range=c(0.5,3),cmax = TRUE, r = 5*24,
ask=FALSE,nt=30)

#Estimaciones de los parámetros para velocidad de viento
# para valores de umbral diferentes.
par(mfrow=c(2,1))
tcplot(datos$V, u.range=c(5,11),cmax = TRUE, r = 5*24,
ask=FALSE,nt=60)

#Una vez elegido el umbral y con los datos sin clústers,
# se estiman los parámetros de la DGP por máxima verosimilitud
pot_iingen<-fitgpd(altura_1[,"obs"],1)
pot_altura<-fitgpd(altura[,"obs"],2.5)
pot_velocidad<-fitgpd(velocidad[,"obs"],9)

#Intervalos de confianza POT al 95% de Confianza

#Intervalos de confianza para los parámetros de altura de ola
# con umbral u=1
gpd.fishape(pot_iingen, conf = 0.95)
gpd.fiscale(pot_iingen, conf = 0.95)

#Intervalos de confianza para los parámetros de altura de ola
# con umbral u=2.5
gpd.fishape(pot_altura, conf = 0.95)
gpd.fiscale(pot_altura, conf = 0.95)

#Intervalos de confianza para los parámetros de velocidad de viento
```

```
# con umbral u=9
gpd.fishape(pot_velocidad, conf = 0.95)
gpd.fiscale(pot_velocidad, conf = 0.95)

#Diagnóstico POT
#Se utilizaron las funciones que a continuación se presentan,
# para las herramientas de diagnóstico en el modelo DGP.
# El autor de las funciones originales es Mathieu Ribatet
# <mathieu.ribatet@math.univ-montp2.fr>
#pp.uvpot
#qq.uvpot
#dens.uvpot
#retlev.uvpot

dgp_diag<-function (fitted)
{
    oldpar <- par(mfrow = c(2, 2))
    plot(fitted,npy=fitted$nat/63,ask=FALSE)
    par(oldpar)
    invisible()
}

dgp_diag(pot_iingen)
dgp_diag(pot_altura)
dgp_diag(pot_velocidad)

#Bondad de Ajuste
#Utilizamos la función antes programada para bondad de ajuste
#Excedencias
sobre1<-altura_1[,"obs"]
sobre2_5<-altura[,"obs"]
sobre9<-velocidad[,"obs"]
```

```
#Bondad de ajuste para u=1 de altura de ola
ks_t(sobre1,pot_iingen)
ad_t(sobre1,pot_iingen)
#Bondad de ajuste para u=2.5 de altura de ola
ks_t(sobre2_5,pot_altura)
ad_t(sobre2_5,pot_altura)
#Bondad de ajuste para u=9 de velocidad de viento
ks_t(sobre9,pot_velocidad)
ad_t(sobre9,pot_velocidad)
```

Bibliografía

[1] Beirlant, J., Goegebeur, Y., and Teugels, J. *Statistics of Extremes. Theory and Applications*. Katholieke Universiteit Leuven, Belgium, 2004.

[2] Bingham, N., Goldie, C., and Teugels, J. *Regular Variation*. Cambridge University Press, Cambridge, 1987.

[3] Campa Rojas, M. A. *Teoría de Valores Extremos con Aplicaciones a Medidas de Riesgo*. Tesis de Licenciatura, Universidad Nacional Auónoma de México, Facultad de Ciencias. 244 páginas, 2001.

[4] CENAPRED. *Guía Básica para la Elaboración de Atlas Estatales y Municipales de Peligros y Riesgos (Fenómenos Hidrometeorológicos)*. Sistema Nacional de Protección Civil, México, 2004.

[5] Coles, S. *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics, University of Bristol, UK, 2001.

[6] de Haan, L. *On Regular Variation and Its Aplication to Weak Convergence of Sample Extremes*. Mathematical Centre Tracts 32, Amsterdam, 1970.

[7] Embrechts, P., Klüppelberg, C., and Mikosch, T. *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, 1997.

[8] Gil Bellosta, C. *Package ADGofTest , Anderson-Darling GoF test*, 2011. http://cran.r-project.org/web/packages/ADGofTest/ADGofTest.pdf.

[9] Gilli, M., and Këllezi, E. *An Application of Extreme Value Theory for Measuring Risk*. Department of Econometrics, University of Geneva and FAME, Switzerland, 2003.

[10] Grabinsky, G. *Teoría de la medida*. Ed. Las prensas de ciencias., Universidad Nacional Autónoma de México. Facultad de Ciencias, Noviembre de 2009.

[11] Gut, A. *Probability: A Graduate Course*. Ed. Springer, United States of America, 2005.

[12] Heffernan, J. E., and Stephenson, A. G. *Package ismev (Version 1.39), An Introduction to Statistical Modeling of Extreme Values*, 2009. http://cran.r-project.org/web/packages/ismev/ismev.pdf.

[13] Hosking, J. R., and Wallis, J. A comparison of unbiased and plotting-position estimators of l moments. *Water Resources Research 31*, 8 (August 1995), 2019–2025.

[14] McNeil, A. J., Frey, R., and Embrechts, P. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, Princeton, 2005.

[15] NIST-SEMATECH. *e-Handbook of Statistical Methods*. U.S. Commerce Department. http://www.itl.nist.gov/div898/handbook/.

[16] Owen, W. *The R Guide (Version 2.5)*, 2010. University of Richmond. Department of Mathematics and Computer Science. Online at http://cran.r-project.org/doc/contrib/Owen-TheRGuide.pdf.

[17] Pfaff, B., McNeil, A., and Stephenson, A. *Package evir (Version 1.7)*, 2011. http://cran.r-project.org/web/packages/evir/evir.pdf.

[18] Reiss, R., and Thomas, M. *Statistical Analysis of Extreme Values: with Applications to Insurance, Finance Hydrology and Other Fields*, 2 ed. Birkhäuser Verlag, 2001.

[19] Resnick, S. I. *Extreme Values, Regular Variation and Point Processes*. Springer-Verlag Series in Operations, Research and Financial Engineering, Cornell University, 1987.

[20] Ribatet, M. *The POT Package*, 2007. http://benz.nchu.edu.tw/ finmyc/POT.pdf.

[21] Ribatet, M. *A User's Guide to the POT Package, Version 1.4*, 2011.

[22] Ruiz Martínez, G., Silva Casarín, R., Pérez Romero, D. M., Posada, G., and Bautista, E. Modelo híbrido para la caracterización del oleaje. *Ingeniería hidráulica en México. XXIV*, 3 (julio-septiembre 2009).

[23] Silva Casarín, R., Mendoza Baldwin, E., Escudero Castillo, M., Posada Vanegas, G., and Arganis Juárez, M. *Characterization Of Risks In Coastal Zones:*

A Review. Instituto de Ingeniería, Universidad Nacional Autónoma de México. Instituto EPOMEX, Universidad Autónoma de Campeche, México, 2012.

[24] Silva Casarín, R., Ruiz Martínez, G., Posada Vanegas, G., Pérez Romero, D. M., Rivillas, G., Espinal, J., and Mendoza Baldwin, E. *Atlas de Clima Marítimo de la Vertiente Atlántica Mexicana*. Instituto de Ingeniería, Universidad Nacional Autónoma de México, 2008.

[25] Smith, R. L. *Statistics of Extremes, With Applications in Environment, Insurance and Finance*. University of North Carolina, USA, 2003.

[26] Stephenson, A., and Ferro, C. *Package evd (Version 2.3-0)*, 2012. http://cran.r-project.org/web/packages/evd/evd.pdf.

[27] The R Graphics Package. *Documentation for package 'graphics' (Version 2.15.3)*. http://stat.ethz.ch/R-manual/R-patched/library/graphics/.

[28] Venables, W., Smith, D., and the R Core team. *An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics. Version 3.1.1*, 2014. http://cran.r-project.org/doc/manuals/R-intro.pdf.