



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

“ANÁLISIS EXTREMAL DEL OLEAJE
COMO FENÓMENO
HIDROMETEOROLÓGICO MEDIANTE LA
APLICACIÓN DE TEORÍA DE VALORES
EXTREMOS BIVARIADA ”

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIO

PRESENTA:

MAYELA EDNA AVILÉS RAMÍREZ

DIRECTOR DE TESIS:

DRA. ANA MEDA GUARDIOLA



2017

CIUDAD UNIVERSITARIA, CD.MX.



Universidad Nacional
Autónoma de México



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

1. Datos del alumno

Avilés

Ramírez

Mayela Edna

26087604

Universidad Nacional Autónoma de México

Facultad de Ciencias

Actuaría

306149607

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

Sergio Iván

López

Ortega

6. Datos del sinodal 4

Actuaría

Viviana

Díaz

Magallanes

7. Datos del trabajo escrito

Análisis extremal del oleaje como fenómeno hidrometeorológico mediante la aplicación de Teoría de Valores extremos bivariada

112 p

2017

Agradecimientos

Agradezco a mis padres Abel Avilés Duarte y María Eugenia Ramírez por su apoyo incondicional durante toda mi vida y por su paciencia durante este proceso. A mi hermano Erick y a mi confidente y mejor amiga, mi hermana Mayra, que me ha brindado su apoyo en cada etapa de mi vida así como su compañía y ánimo durante todo este tiempo.

Agradezco a la Universidad Nacional Autónoma de México y a la Facultad de Ciencias por darme la oportunidad de concluir mi carrera universitaria. A la Doctora Ana Meda Guardiola por brindarme su apoyo asesoría en la realización de este trabajo y por su paciencia, esfuerzo y comprensión durante todo el proceso y tiempo involucrado. Al Doctor Rodolfo Silva Casarín y al Doctor Edgar Mendoza Baldwin por todo su apoyo brindado para hacer posible el inicio y la realización de este trabajo y que a través del Instituto de Ingeniería de la UNAM, la Coordinación de Hidráulica y el Laboratorio de Costas y Puertos me proporcionaron recursos, asesoría, un espacio de trabajo y más atenciones que me permitieron llevar a cabo este trabajo y por lo cual les estoy muy agradecida. También agradezco a la Maestra Mireille Del Carmen Escudero Castillo y al Doctor Miguel Angel Delgadillo Calzadilla, por su asesoría. A mis sinodales, por su tiempo dedicado en la revisión de este trabajo.

A mis profesores y amigos, en particular a la Maestra Giselle Ochoa Hofmann y Jorge Ramírez por su inspiración que me han brindado, su amistad y apoyo. Al Doctor Javier Páez y al Doctor Alejandro Darío Rojas Sánchez de quienes he recibido apoyo, guía, consejo y amistad en todo momento. A mis compañeros y amigos, en especial a mi amigo Isaac por su apoyo y amistad en los buenos y malos momentos; por último pero no menos importante a mi novio Ángel por su amor, ánimo y apoyo brindado que ha sido muy importante para mí.

Índice general

Índice de figuras	7
Índice de tablas	9
Introducción	11
1. Teoría de Valores Extremos (TVE)	15
1.1. Modelación de valores extremos en \mathbb{R}	17
1.1.1. Máximos por Bloque	18
1.1.2. Dominios de Atracción del Máximo de las DVE	24
1.1.3. Distribución Generalizada de Valores Extremos (DGVE)	25
1.2. Excedencias sobre un umbral (POT)	28
1.2.1. Distribución Generalizada de Pareto	29
2. Teoría de Valores Extremos Multivariada	33
2.1. Modelación de valores extremos en \mathbb{R}^d	34
2.1.1. Máximos por bloque bivariados	38
2.1.2. Modelo Bivariado de Excedencias sobre un umbral	38
2.2. Cópulas de valores extremos	39
2.2.1. Representación de cópulas de valores extremos	41
2.2.2. Modelos paramétricos	44
3. Herramientas estadísticas para el análisis de los datos	47
3.1. Ajuste de valores extremos univariado	47
3.1.1. Estimación de parámetros	48
3.1.2. Diagnóstico y evaluación del modelo	49

3.2. Ajuste bivariado de valores extremos	52
3.2.1. Estimación de parámetros	52
3.2.2. Diagnóstico y evaluación del modelo	53
4. Análisis de valores extremos del Oleaje mediante altura de Ola y velocidad de viento	55
4.1. Análisis exploratorio de los datos	57
4.2. Análisis de eventos extremos en el oleaje.	58
4.2.1. Máximos por bloque	59
4.2.2. Modelación bivariada de máximos por bloque	62
4.2.3. Excedencias sobre un umbral	66
4.2.4. Modelación bivariada de excedencias sobre un umbral	72
Conclusiones	75
A. Proceso de Ajuste	79
B. Proceso de Ajuste	81
C. Resultados varios de Probabilidad	83
D. Herramientas Estadísticas	93
E. Código R	97

Índice de figuras

1. Encontrado en facebook	14
1.1. Métodos de Valores Extremos	17
2.1. Función de dependencia de Pickands.	43
4.1. Localización geográfica de la estación 42002 de la NDBC.	56
4.2. Histogramas de altura de Ola significativa y velocidad de Viento y Box Plot.	58
4.3. Diagrama de dispersión de Altura de ola y Velocidad de viento	59
4.4. ACF y PACF de la submuestra de máximos semestrales de altura de Ola significativa y velocidad de Viento.	60
4.5. Gráficas de diagnóstico del ajuste de Máximas alturas de ola significativa.	62
4.6. Gráficas de diagnóstico del ajuste de Máximos de velocidad de viento.	63
4.7. Gráfico Chi-plot de la altura de Ola significativa y Velocidad de viento.	64
4.8. Gráfico de líneas de contorno: líneas punteadas azules para el ajuste de la Cópula Gumbel y líneas punteadas amarillas para el ajuste de la Cópula Tawn. 65	
4.9. Ajuste Bivariado de Máximos por Bloque.	66
4.10. Gráfico de Función de medias de excesos de altura de ola significativa y velo- cidad de viento.	67
4.11. Gráfico de estimaciones de los parámetros de la DGP de escala modificada y forma para distintos umbrales de altura de ola.	68
4.12. Gráfico de estimaciones de los parámetros de la DGP de escala modificada y forma para distintos umbrales de velocidad de viento.	69
4.13. Gráficas de diagnóstico del ajuste de alturas de ola por encima de un umbral: $u = 4.3$	71

4.14. Gráficas de diagnóstico del ajuste de velocidades de viento por encima de un umbral: $u = 16.7$	72
4.15. Gráfico de líneas de contorno: Ajuste de la Cópula Tawn estimada por el método de excedencias sobre un umbral.	74
A.1. Diagrama de flujo del proceso de ajuste MB	79
B.1. Diagrama de flujo del proceso de ajuste POT	81

Índice de tablas

4.1. Resumen estadístico de los datos.	57
4.2. Estimadores e Intervalos de Confianza al 95 % de los parámetros de la DGVE.	61
4.3. Resultados del ajuste.	61
4.4. Resultados del ajuste.	64
4.5. Resultados del ajuste.	65
4.6. Estimadores e Intervalos de Confianza al 95 % de los parámetros de la DGP.	70
4.7. Estadísticos y p-values de las pruebas de bondad de Ajuste Kolmogorov-Smirnov y Anderson-Darling de la estimación de los parámetros de la DGP.	70
4.8. Resultados del ajuste.	73

Introducción

Las zonas costeras son afectadas por diferentes fenómenos naturales, en su mayoría de origen hidrometeorológico, como las mareas, las precipitaciones, tormentas, olas altas, etc, que pueden alcanzar magnitudes extraordinarias durante la ocurrencia de eventos como huracanes y tsunamis. Las consecuencias directas de estos eventos extremos son inundaciones (derivados del aumento medio del nivel del mar), daños a las playas, daños en infraestructura y hábitats naturales, etc. A su vez, el hecho de que exista un latente cambio climático representa un aumento de los riesgos que se presentan en las zonas costeras.

En [37] se explica que uno de los factores de gran importancia que modifican las costas es el Oleaje, que es un fenómeno que está determinado por la acción de las fuerzas de la naturaleza en cualquier superficie libre de agua, donde la más obvia de estas fuerzas es la acción del viento sobre la superficie del mar y que la altura de ola representa uno de los parámetros más relevantes de estudio. Además es el principal agente en la formación de las playas, que junto con los efectos de las mareas, tormentas, tsunamis y vientos, las moldean de manera activa. Silva resalta que debe considerarse como un fenómeno aleatorio: *"Dado que este fenómeno presenta gran variabilidad pues no se repite en el espacio ni en el tiempo y cuando se observa una altura de ola en un punto dado del mar, no se puede precisar cuál será la altura de ola siguiente en ese punto, la forma más razonable de caracterizarlo es a través de los métodos estadísticos, tratando al oleaje como un fenómeno aleatorio"*.

Como se afirma en ([37], [38] y [1]), en México los trabajos y estudios disponibles sobre la evaluación de riesgos para nuestras zonas costeras son todavía muy poco desarrollados y se carece de una fuente de información generalizada ya que los datos de oleaje disponibles corresponden a trabajos muy locales. Silva *et al.* en [37] comenta que *"Ante tal deficiencia, la ingeniería costera moderna se ha dado a la tarea de realizar predicciones numéricas de*

las condiciones marinas en aguas profundas a partir de datos de viento, de los cuales sí se cuenta con información confiable".

El análisis de eventos extremos está presente en diversas áreas de gestión de riesgos por lo tanto la teoría de valores extremos tiene un gran número de aplicaciones en diversos campos como las finanzas, ciencias atmosféricas y oceanográficas así como en la ingeniería y sector asegurador, entre otras. Los eventos extremos son poco frecuentes, caracterizados generalmente con una severidad grande y fuera de la experiencia normal, lo que implica un alto grado de incertidumbre; éstos ocurren cuando un riesgo toma valores en la cola de su distribución, así se desarrolla un modelo mediante la selección de una distribución de probabilidad particular de un análisis estadístico mediante datos empíricos del fenómeno en cuestión. En este caso la teoría de valores extremos es una herramienta que trata de proporcionar la mejor estimación posible del comportamiento de la cola de la distribución.

El estudio de la dependencia presente en los eventos extremos ha tenido un creciente interés, alimentado por acontecimientos como la crisis financieras o los catástrofes naturales que en esencia son propiciados por diversos factores que actúan de manera conjunta. Las cópulas de valores extremos combinan la teoría de valores extremos con teoría de cópulas y surgen como distribuciones límite de cópulas de vectores de máximos de muestras independientes e idénticamente distribuidas y, por lo tanto, se consideran como modelos apropiados para la estructura de dependencia entre los fenómenos de valores extremos.

El objetivo de esta tesis es presentar un modelo estadístico de eventos extremos del fenómeno oceánico del oleaje para modelar un conjunto de datos bivariados con mediciones de altura de ola y velocidad de viento. Se escogieron estos dos, pues, al ser parámetros con una estrecha relación entre ellos, esperamos una gran correlación entre sí para así poder realizar un análisis de su comportamiento conjunto. Se emplearán los principales resultados de la Teoría de Valores Extremos con Teoría de Cópulas para el comportamiento conjunto.

El Instituto de Ingeniería de la UNAM (IINGEN) brindó apoyo en la realización de esta tesis. Los datos se obtuvieron del banco de datos histórico de la *National Oceanic and Atmospheric Administration* (NOAA) y la *National Data Buoy Center* (NDBC) en un punto geográfico localizado en el Océano Atlántico dentro del Golfo de México, cercano a la costa de

Tamaulipas y Texas. La NDBC cuenta con mediciones provenientes de boyas oceanográficas de varios parámetros oceánicos, entre ellos, los que se ocuparán en esta tesis.

En el IINGEN se han realizado análisis previos de valores extremos en algunas zonas costeras de México, pero no se cuenta con estudios previos de valores extremos bivariados, por tanto, el análisis que se realizará en esta tesis podría servir para la simulación de escenarios de extremos bivariados que complemente y mejore la calibración de modelos existentes. Se enfatiza que no se discutirá toda la teoría existente de cópulas ni la referente a todos los diferentes enfoques en la Teoría de Valores extremos multivariada, sino que se pretende ejemplificar una forma de realizar un análisis estadístico para aplicar un estudio bivariado de valores extremos.

La presente tesis se desarrolló en cuatro capítulos, en el primer capítulo se presentan los principales resultados de la teoría de valores extremos univariada para los métodos de máximos por bloque y excedencias sobre un umbral. En el segundo capítulo se desarrolla la caracterización de la Teoría de valores extremos multivariada y se describen los métodos de máximos por bloque y excedencias sobre un umbral desde un enfoque bivariado abordando resultados básicos de cópulas y cópulas de valores extremos. En el tercer capítulo se describen las herramientas estadísticas para el análisis, estimación y evaluación del ajuste univariado y bivariado de los modelos descritos en el primer y segundo capítulo. En el capítulo cuatro se muestra la aplicación de los tres capítulos anteriores mediante el análisis conjunto de eventos extremos de las variables altura de ola y veocidad de viento. Finalmente se presentan las conclusiones del trabajo y un apéndice con algunos resultados empleados en el desarrollo de la tesis.

El análisis estadístico así como los gráficos se realizaron con el software libre para análisis estadístico y gráfico llamado **R**. Los cálculos se llevaron a cabo empleando las paqueterías *evd*, *evir*, *ismev*, *POT*, *copula*, *fCopulae*, *stats* y *ADGofTest* y se incluye el código de éstos en el apéndice.

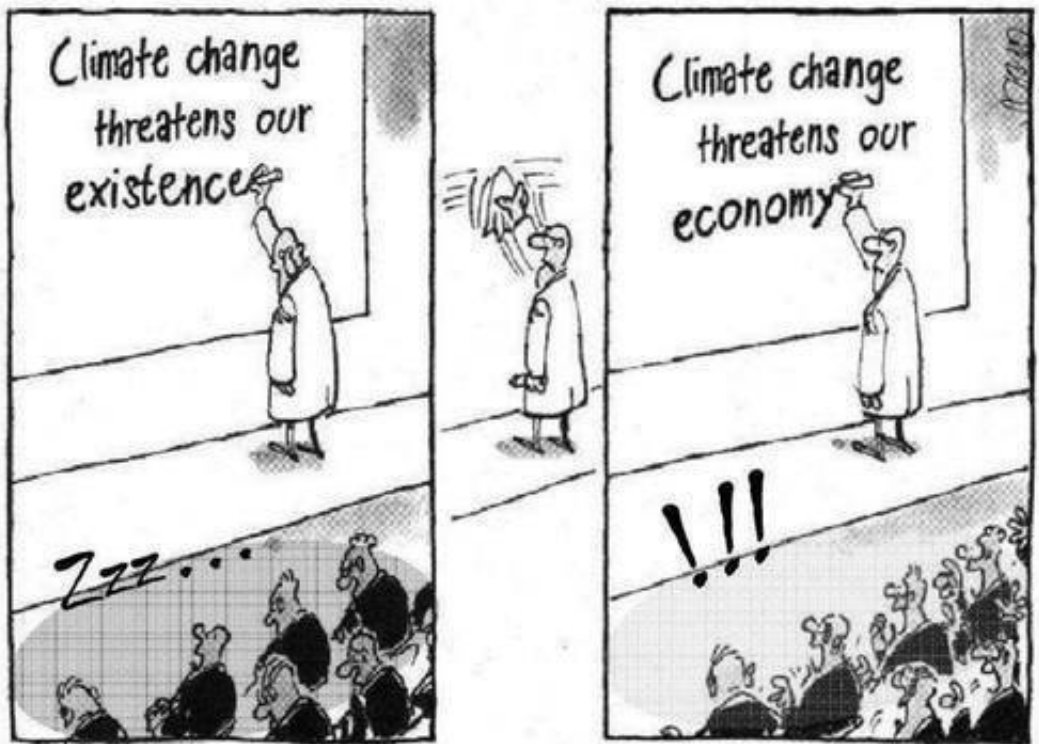


Figura 1: Encontrado en facebook

Capítulo 1

Teoría de Valores Extremos (TVE)

La teoría de valores extremos comprende el estudio del comportamiento estocástico de los valores extremos en un proceso. Las bases que asientan el desarrollo de la TVE parten de las características y conducta del máximo de una secuencia de n variables aleatorias independientes e idénticamente distribuidas (v.a.i.i.d.) X_1, X_2, \dots, X_n con función de distribución común F . Éste representa la observación más extrema dentro del conjunto de observaciones y lo denotaremos como:

$$M_n = \max(X_1, \dots, X_n).$$

con función de distribución dada por:

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

Al hablar de eventos extremos nos referimos a aquellos valores máximos (mínimos) que alcanza una variable aleatoria y ocurren con poca probabilidad situándose en la cola de la función de distribución de las mismas.

Denotamos al extremo derecho de F como

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

Tenemos entonces que, para todo $x < x_F$,

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

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

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

De aquí se puede ver que $M_n \xrightarrow{P} x_F$ cuando $n \rightarrow \infty$.

Proposición 1.1. $M_n \xrightarrow{P} x_F$ cuando $n \rightarrow \infty$ para $x_F \leq \infty$.

Demostración. P.D. Para todo $\epsilon > 0$, $\lim_{n \rightarrow \infty} \mathbb{P}(|M_n - x_F| < \epsilon) = 1$

Sea $\epsilon > 0$, entonces

$$\begin{aligned} \mathbb{P}(|M_n - x_F| < \epsilon) &= \mathbb{P}(-\epsilon < M_n - x_F < \epsilon) \\ &= \mathbb{P}(x_F - \epsilon < M_n < x_F + \epsilon) \\ &= F^n(x_F + \epsilon) - F^n(x_F - \epsilon) \end{aligned}$$

Tomando el límite,

$$\lim_{n \rightarrow \infty} \mathbb{P}(|M_n - x_F| < \epsilon) = \lim_{n \rightarrow \infty} F^n(x_F + \epsilon) - F^n(x_F - \epsilon) = 1 - 0 = 1.$$

□

Además como la sucesión (M_n) es no decreciente en n se puede demostrar que converge casi seguramente a x_F . Sin embargo, esto no nos proporciona suficiente información pues en la práctica F es desconocida, esto nos lleva a estudiar el comportamiento asintótico de M_n y buscar familias aproximadas de modelos para F^n .

Para un sólo proceso, el comportamiento de los máximos puede ser descrito mediante alguna de las familias de distribuciones que surgen formalmente como distribuciones límite para el máximo (o mínimo) normalizado de una secuencia de variables aleatorias. Éstas son llamadas Distribuciones de Valores Extremos (DVE) y son las siguientes:

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

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

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

Estas distribuciones tienen distinta forma pues presentan un comportamiento distinto en su cola. Para la distribución Weibull su extremo derecho x_F es finito, considerada entonces como de cola corta. Para las distribuciones Gumbel y Fréchet $x_F = \infty$ sin embargo, la cola de la distribución Gumbel decae exponencialmente y la de la distribución Fréchet polinomialmente siendo entonces ésta última de cola más "pesada".

Para continuar describiendo los temas centrales de la TVE, enunciaremos a continuación algunos fundamentos importantes en la modelación de valores extremos.

1.1. Modelación de valores extremos en \mathbb{R}

En términos generales, existen dos tipos de modelos para valores extremos recogidos de muestras de observaciones independientes e idénticamente distribuidas, estos son: *modelos de máximos por bloque* y *modelos de excedencias sobre un umbral*. El primero es el grupo más antiguo de modelos para las observaciones máximas de muestras de observaciones idénticamente distribuidas, mientras que los de excedencias sobre un umbral resultan de mayor utilidad para aplicaciones prácticas, principalmente por el uso eficiente de los datos al tomar todas las observaciones que exceden un indicado umbral suficientemente alto. A continuación se desarrollarán los principales resultados de ambos métodos.

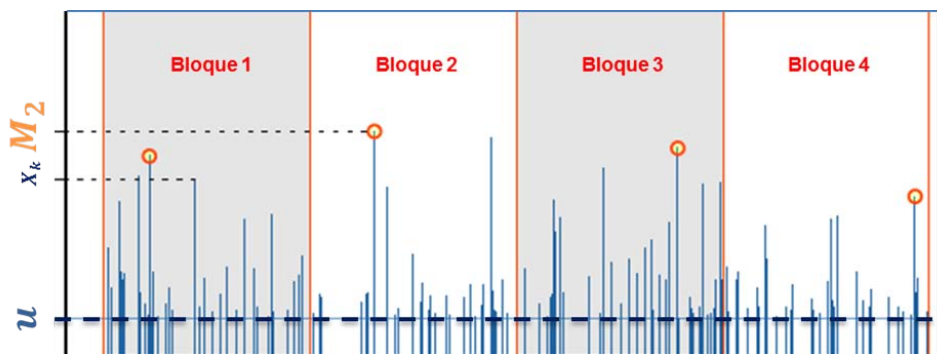


Figura 1.1: Métodos de Valores Extremos

1.1.1. Máximos por Bloque

Una posibilidad de extraer valores máximos de un conjunto de datos es mediante el método conocido como *Máximos por bloque* (*Block Maxima models*). Como su nombre lo indica, se refiere a la selección de máximos mediante la división en bloques de los valores de una variable aleatoria. Los bloques se eligen del mismo tamaño, el cual frecuentemente corresponde a períodos de tiempo, por ejemplo, horas, días, meses, años, etc. De tal forma, si el tamaño de los bloques se elige de un año, cada bloque tendrá el número de observaciones en un año, por consiguiente se seleccionará el máximo de cada bloque obteniendo así una muestra de máximos anuales.

Como se explicó anteriormente se requieren modelos que describan el comportamiento asintótico de M_n , sin embargo se requiere cierta cautela con las distribuciones límite, pues para no dar lugar a formas degeneradas se busca tener una representación análoga del Teorema del Límite Central para suma de variables aleatorias, esto es, se necesita una normalización adecuada mediante la existencia de dos secuencias de constantes $\{a_n > 0\}$ y $\{b_n\} \in \mathbb{R}$ tal que la función

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

sea una distribución no degenerada.

Todo lo anterior se resume en el siguiente resultado base en la TVE:

Teorema 1.1. (Fisher-Tippet, Gnedenko): ¹. Sean X_1, X_2, \dots, X_n variables aleatorias independientes e idénticamente distribuidas, y sea $M_n = \max\{X_1, X_2, \dots, X_n\}$. Si existen constantes $a_n > 0$, $b_n \in \mathbb{R}$, y alguna función de distribución no degenerada G , tales que:

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

donde \mathcal{C} denota al conjunto de puntos de continuidad de G ; entonces G tiene la forma de una de las tres funciones de distribución de valores extremos.

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

$$F^{[nt]}(a_{[nt]}x + b_{[nt]}) \rightarrow G(x) \quad \forall x \in \mathcal{C}(G),$$

¹Este teorema fue originalmente propuesto por Fisher y Tippett en 1928 y demostrado rigurosamente por Gnedenko en 1943.

donde $[x]$ representa la parte entera de x . Por otro lado, también tenemos que

$$F^{[nt]}(a_n x + b_n) = (F^n(a_n x + b_n))^{\frac{[nt]}{n}} \rightarrow G^t(x) \quad \forall x \in \mathcal{C}(G^t)$$

Entonces, por el Teorema de Convergencia a Tipos (Teorema C.1), G y G^t son del mismo tipo, es decir, existen $\alpha(t) > 0$ y $\beta(t) \in \mathbb{R}$ tal que para todo $t > 0$,

$$\lim_{n \rightarrow \infty} \frac{a_n}{a_{[nt]}} = \alpha(t), \quad \lim_{n \rightarrow \infty} \frac{b_n - b_{[nt]}}{a_{[nt]}} = \beta(t).$$

se tiene que

$$G^t(x) = G(\alpha(t)x + \beta(t)), \quad (1.2)$$

Primero veremos que las funciones $\alpha(t)$ y $\beta(t)$ son Lebesgue medibles. Para ver que $\alpha(t)$ es Lebesgue medible, como los límites y cociente de funciones medibles son medibles, basta mostrar que, para $n \in \mathbb{N}$, la función $t \mapsto \frac{a_n}{a_{[nt]}}$ es medible. Como a_n no depende de t , lo anterior es cierto si $t \mapsto a_{[nt]}$ es medible. Notemos que el rango de esta función es $\{a_j : j \geq 1\}$, que es numerable y así basta probar que $\{t > 0 : a_{[nt]} = a_j\}$ es medible. Como este conjunto corresponde a $\bigcup_{k: a_k = a_j} [\frac{k}{n}, \frac{k+1}{n})$ el cuál es medible por ser unión a lo más numerable de intervalos. Análogamente se demuestra para $\beta(t)$.

Utilizando (1.2) se tiene que para todo $t, s > 0$

$$G^{ts}(x) = G(\alpha(ts)x + \beta(ts))$$

que además se puede escribir como

$$\begin{aligned} G^{ts}(x) &= (G^s(x))^t = G(\alpha(s)x + \beta(s))^t \\ &= G(\alpha(t)(\alpha(s)x + \beta(s)) + \beta(t)) \\ &= G(\alpha(t)\alpha(s)x + \alpha(t)\beta(s) + \beta(t)), \end{aligned}$$

y como suponemos que G es no degenerada se tiene: ²

$$\begin{aligned} \alpha(ts) &= \alpha(t)\alpha(s) \\ \beta(ts) &= \alpha(t)\beta(s) + \beta(t) \\ &= \alpha(s)\beta(t) + \beta(s) \\ &= \beta(st). \end{aligned} \quad (1.3)$$

²Además se usa el resultado: Si una función de distribución F no degenerada satisface $F(ax + b) = F(cx + d) \forall x \in \mathbb{R}$ y $a, c > 0$ entonces $a = c$ y $b = d$.

Notemos que (1.3) es una variante de la ecuación de Cauchy llamada ecuación de Hamel (C.15, definida en el Apéndice C), por tanto se tiene:

$$\alpha(t) = t^{-\theta}$$

. por tanto se tienen los casos $\theta = 0$, $\theta > 0$ y $\theta < 0$:

1. $\theta = 0$

Si $\theta = 0$ de (1.3) tenemos:

$$\beta(ts) = \beta(s) + \beta(t)$$

. Entonces $\exp \beta(\cdot)$ satisface la ecuación de Hamel (C.15) y por lo tanto

$$\exp \beta(t) = t^{-c},$$

por lo que para $c \in \mathbb{R}$

$$\beta(t) = -c \log t.$$

Además en (1.2) tenemos que

$$G^t(x) = G(x - c \log t) \tag{1.4}$$

Ahora hacemos las siguientes observaciones de (1.4):

- $c \neq 0$ pues de lo contrario se contradice que G sea no degenerada.
- Para x fija, $G^t(x)$ es no decreciente en t , lo que nos indica que $c > 0$.
- Si para $x_0 \in \mathbb{R}$, $G^t(x_0) = 1$ entonces pra toda $t > 0$

$$1 = G(x_0 - c \log t),$$

y haciendo un cambio de variable $u = x_0 - c \log t$ tenemos que $1 = G(u)$ para toda $u \in \mathbb{R}$ lo que contradice que G sea no degenerada. De la misma forma si $G(x_0) = 0$, $0 = G^t(x_0) = G(x_0 - c \log t)$ para toda t y por tanto $G(x) = 0$ para toda $x \in \mathbb{R}$.

Por lo tanto $0 < G(y) < 1$ pra toda $y \in \mathbb{R}$.

Sustituyendo $x = 0$ en (1.4) tenemos que

$$G^t(0) = G(-c \log t). \tag{1.5}$$

Tomando $G(0) = \exp\{-e^{-x}\} \in (0, 1)$ y $u = -c \log t$, aplicando el cambio de variable en (1.5) obtenemos:

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

2. $\theta > 0$

Fijando $s_0 \neq 1$ (pues de lo contrario $\alpha(s_0) = s_0^{-\theta} = 1$),

$$\alpha(t)\beta(s_0) + \beta(t) = \alpha(s_0)\beta(t) + \beta(s_0),$$

entonces para $s_0 \neq 1$ y $t \neq 1$,

$$\begin{aligned} \beta(t) &= \left(\frac{\beta(s_0)}{1 - \alpha(s_0)}\right)(1 - \alpha(t)) \\ &= c(1 - \alpha(t)) \\ &= c(1 - t^{-\theta}). \end{aligned}$$

De aquí tenemos que (1.2) se reescribe como

$$G^t(x) = G(t^{-\theta}x + c(1 - t^{-\theta})) = G(t^{-\theta}(x - c) + c).$$

Sea $y = x - c$ entonces

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

Y sea $H(y) = G(y + c)$, entonces H y G son del mismo tipo. Como G es no degenerada y satisface

$$H^t(y) = H(t^{-\theta}y), \tag{1.6}$$

observemos que:

- Si hacemos $y = 0$ y tomando logaritmo $t \log G(0) = \log H(0)$, por lo tanto $\log H(0) = 0$ o $\log H(0) = -\infty$ lo que implica que $H(0) = 0$ o $H(0) = 1$. Pero $H(0) \neq 1$ pues de lo contrario existe $x < 0$ para el cual el lado izquierdo de (1.6) es decreciente en t , mientras que el lado derecho es creciente en t .

Por lo tanto $H(0) = 0$.

- De (1.6) haciendo $y = 1$ tenemos $H^t(1) = H(t^{-\theta})$.

Ahora si $H(1) = 0$ entonces $H \equiv 0$ y si $H(1) = 1$ entonces $H \equiv 1$ pero en ambos casos se contradice que H sea no degenerada.

por tanto $H(1) \in (0, 1)$

Sea $\theta^{-1} = \alpha$, $H(1) = \exp\{-p^{-\alpha}\}$, $u = t^{-\theta}$.

Sustituyendo con $y = 1$ en (1.6) tenemos para $u > 0$:

$$H(u) = (\exp\{-p^{-\alpha}\})^t = \exp\{-p^{-\alpha}t\} = \exp\{-(pu)^{-\alpha}\} = \Phi_\alpha(pu).$$

a) $\theta < 0$

De la expresión (1.6) tenemos lo siguiente:

- Haciendo $y = 0$, $H(0) = 0$ o $H(0) = 1$, pero $H(0) \neq 0$ pues de lo contrario existiría $x < 0$ tal que el lado izquierdo de la expresión (1.6) es decreciente en t y el lado derecho es creciente en t .

Por lo tanto $H(0) = 1$

- Haciendo $y = -1$ tenemos que $H^t(-1) = H(-t^{-\theta})$, pero si $H(-1) = 0$ entonces $H(-t^{-\theta}) = 0^t \equiv 0$ para todo t real y si $H(-1) = 1$ entonces $H(-t^{-\theta}) = 1^t \equiv 1$ para todo t real. En ambos casos se contradice que H sea no degenerada.

por lo tanto $H(-1)$ está en $(0, 1)$.

Sea $\theta^{-1} = -\alpha$, $G(-1) = \exp\{-(-p)^\alpha\}$, $u = t^{-\theta}$. Sustituyendo con $y = -1$ en (1.6), para $u > 0$:

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

□

Es importante tener presente que el Teorema (1.1) no garantiza la existencia de un límite no degenerado para M_n , ni nos dice cuál es el límite cuando existe, lo que sí nos dice es que teniendo máximos con la normalización adecuada y que convergen en distribución a un límite no degenerado, entonces este límite debe de ser una distribución de valores extremos;

más aún, no es necesario conocer la distribución F de la muestra original.

El siguiente concepto define qué clases de distribuciones caracterizan las posibles distribuciones límite para el máximo M_n de una secuencia de v.a.i.i.d. bajo transformaciones afines.

Definición. (Distribución Max-estable) Una distribución G se dice que es max-estable si para todo $n \in \mathbb{N}$ existen constantes $a_n > 0$ y $b_n \in \mathbb{R}$ tales que

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

Como G^n es la función de distribución de $M_n = \max(X_1, \dots, X_n)$ de v.a.i.i.d. con función de distribución G , se tiene que la max-estabilidad es una propiedad satisfecha por aquellas distribuciones cuyos máximos conducen a una distribución del mismo tipo que la original. A continuación se presenta un teorema que describe la relación de las funciones max-estables con las DVE.

Teorema 1.2. Propiedad límite de las distribuciones max-estables La clase de las distribuciones max-estables coincide con la clase de las posibles distribuciones límite no degeneradas para máximos apropiadamente normalizados de v.a.i.i.d.

Demostración. Por definición sabemos que toda función max-estable es una distribución límite para máximos de v.a.i.i.d. Resta demostrar que aquellas distribuciones límite para máximos apropiadamente normalizados de v.a.i.i.d. son max-estables. Supongamos entonces que para constantes apropiadas tenemos que para alguna función G no degenerada

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

y por el Teorema (1.1) sabemos que las posibles funciones de distribución límite son continuas, por lo que para toda $k \in \mathbb{N}$,

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

además

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

Por el Teorema de Convergencia a Tipos (C.1) tenemos que existen constantes $\tilde{a}_k > 0$ y $\tilde{b}_k \in \mathbb{R}$ tales que

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

y

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

para cada k . Por tanto, G es una distribución max-estable. \square

Se concluye entonces que la clase de las DVE representan las únicas distribuciones max-estables posibles. Basta simple manipulación algebraica para demostrar la max-estabilidad de cada una de las tres DVE. Nótese que con (1.2) y (1.1) se puede concluir que las distribuciones de valores extremos coinciden precisamente con la clase de las distribuciones max-estables.

1.1.2. Dominios de Atracción del Máximo de las DVE

Para conocer cuales son las condiciones necesarias para obtener distribuciones límite o más precisamente las distribuciones de valores extremos G y así reunir en una misma clase a todas las distribuciones F cuyos máximos normalizados tienen el mismo tipo de distribución límite se requiere definir algunos conceptos.

Definición. (Dominio de Atracción del máximo) Se dice que la función de distribución F pertenece al dominio de atracción de la distribución de valor extremo G ($F \in \mathcal{D}(G)$), si existen constantes $a_n > 0$, $b_n \in \mathbb{R}$ tales que

$$F^n(a_n x + b_n) = \mathbb{P}(M_n \leq a_n x + b_n) \rightarrow G(x), \quad x \in \mathbb{R}. \quad (1.7)$$

Es importante hacer notar que no es necesario conocer la forma exacta de F para determinar a qué dominio de atracción pertenece, sus máximos sólo pueden pertenecer al dominio de atracción de alguna de las tres distribuciones de valores extremos y el comportamiento que tenga en su cola \bar{F} determinará su dominio de atracción. Mediante la siguiente proposición se tiene otro criterio para precisar a los dominios de atracción.

Proposición 1.2. Sea F una función de distribución, entonces $F \in \mathcal{D}(G)$ con constantes normalizantes $a_n > 0$ y $b_n \in \mathbb{R}$ si y sólo si

$$\lim_{n \rightarrow \infty} n \bar{F}(a_n x + b_n) = -\ln G(x). \quad (1.8)$$

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

Demostración. Supongamos que $G(x) > 0$, sacando logaritmos en (1.7) tenemos

$$-n \ln(1 - \bar{F}(a_n x + b_n)) \rightarrow -\ln G(x).$$

Usando que $\lim_{z \rightarrow 0} -\ln(1-z)/z = 1$ tenemos que $-\ln(1-z) \sim z$ cuando z tiende a 0, deducimos que

$$n\bar{F}(a_n x + b_n) \rightarrow -\ln G(x).$$

Basta mostrar que $\bar{F}(a_n x + b_n) \rightarrow 0$ cuando $n \rightarrow \infty$. Esto se da pues de otro modo existiría una subsucesión con índices $\{n_k\}$ tal que $\bar{F}(a_{n_k} x + b_{n_k})$ estaría acotada por algo positivo, de modo que $(1 - \bar{F}(a_{n_k} x + b_{n_k}))^{n_k} < 1$, lo que implicaría que $\bar{F}^{n_k}(a_{n_k} x + b_{n_k}) \rightarrow 0$ llegando a una contradicción, pues supusimos que $\bar{F}^n(a_n x + b_n) \rightarrow G(x) > 0$. Para el recíproco, supongamos (1.8). Tenemos entonces que

$$\begin{aligned} F^n(a_n x + b) &= (1 - \bar{F}(a_n x + b_n))^n \\ &= \left(1 - \frac{n\bar{F}(a_n x + b_n)}{n}\right)^n \rightarrow \exp\{-(-\ln G(x))\} = G(x). \end{aligned}$$

En caso de que $G(x) = 0$, supongamos (1.8) interpretando el límite como ∞ , pero que $F^n(a_n x + b_n)$ no converge a 0 cuando n tiende a infinito, esto implica que existe una subsucesión para la cual $F^{n_k}(a_{n_k} x + b_{n_k}) \rightarrow H(x) > 0$ para alguna función H se sigue que $n_k \bar{F}(a_{n_k} x + b_{n_k}) \rightarrow -\ln H(x) < \infty$, lo cual es una contradicción. De manera similar se prueba el recíproco. \square

1.1.3. Distribución Generalizada de Valores Extremos (DGVE)

De acuerdo con Von Mises y Jenkinson (puede consultarse en [39]), es posible obtener una representación de las tres distribuciones de valores extremos involucrando un solo parámetro. Esta representación se conoce como Distribución Generalizada de Valores Extremos (DGVE, GEV por sus siglas en inglés) y resulta de gran importancia en las aplicaciones estadísticas.

Definición. (Distribución de valores extremos generalizada (DGVE)):³ *La distribución generalizada de valores extremos (DGEV o GEV por sus siglas en inglés) se tiene como una distribución con una parametrización común de las tres distribuciones de valores extremos y está dada por*

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

donde $(1 + \xi x) > 0$

³Representación propuesta por Von Mises (1954) y Jenkinson (1955).

El parámetro ξ es conocido como el parámetro de forma de la DGVE, cuyo valor determina la correspondiente distribución de valores extremos. Tenemos entonces que el soporte de G_ξ corresponde a

$$\begin{aligned} x &> -\xi^{-1} && \text{para } \xi > 0 && \text{DVE Fréchet,} \\ x &< -\xi^{-1} && \text{para } \xi < 0 && \text{DVE Weibull,} \\ x &\in \mathbb{R} && \text{para } \xi = 0 && \text{DVE Gumbel.} \end{aligned}$$

Introduciendo parámetros de localización μ y de escala $\sigma > 0$ en la DGVE:

$$\begin{aligned} G_{\xi,\mu,\sigma}(x) &:= G_\xi\left(\frac{x-\mu}{\sigma}\right) \\ G_{\xi,\mu,\sigma}(x) &= \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \end{aligned} \quad (1.9)$$

La DGVE proporciona una representación de gran utilidad en la aplicación estadística pues unifica las tres representaciones de las DVE. 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 $\alpha = -\xi^{-1}$

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

Las estimaciones de cuantiles extremos de la distribución de valores extremos se obtienen entonces invirtiendo (1.9):

$$x_p = \begin{cases} \mu - \frac{\sigma}{\xi}(1 - (-\ln(1-p)))^{-\xi} & , \xi \neq 0 \\ \mu - \sigma \ln(-\ln(1-p)) & , \xi = 0 \end{cases} \quad (1.10)$$

donde $G(x_p) = 1 - p$. A x_p se le conoce como nivel de retorno asociado al periodo de retorno $1/p$, se espera que el nivel de retorno x_p sea superado en promedio una vez cada $1/p$ años.

Un resultado propuesto por Pickands, Balkema y de Haan [30] brinda una caracterización del dominio de atracción de la Distribución Generalizada de Valores Extremos y se enuncia a continuación.

Proposición 1.3. *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 $1 + \xi x > 0$,*

$$\lim_{u \uparrow \omega_F} \frac{\bar{F}(u + xa(u))}{\bar{F}(u)} = \begin{cases} (1 + \xi x)^{-1/\xi} & , \xi \neq 0 \\ e^{-x} & , \xi = 0 \end{cases} \quad (1.11)$$

Demostración. Para el caso $\xi = 0$, se trata de la Proposición (1.2).

Para $\xi > 0$ tenemos entonces que (1.11) es equivalente a $F \in \mathcal{D}(\Phi)$, lo cuál implica que $\bar{F} \in VR_{-\alpha}$, con $\alpha = \xi^{-1}$. Entonces, por la Representación de Karamata para funciones de variación regular (ver apéndice C), \bar{F} tiene la siguiente representación

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

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

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

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

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

cuando $u \rightarrow \infty$. Haciendo el cambio de variable $t = u + va(u)$ tenemos 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 concluimos (1.11). Recíprocamente, si se cumple b), tomamos $a_n = (1/\bar{F})^{\leftarrow}(n)$ como en la demostración de la Proposición (1.2). Tenemos que $\bar{F}(a_n) \sim n^{-1}$, y que $a_n \rightarrow \omega_F$

cuando $n \rightarrow \infty$. Entonces

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

lo cual implica, por la Proposición (1.2), que $F \in \mathcal{D}(G_\xi)$. El caso para $\xi < 0$ puede demostrarse de manera similar.

Es importante resaltar que el método de máximos por bloque es el enfoque central en el desarrollo de la TVE y se necesita tomar en cuenta un aspecto crítico al implementarlo pues se debe tener cuidado con la elección del tamaño de bloque, por un lado, bloques demasiados pequeños llevarían a que la aproximación límite dada en el teorema (1.1), será con mucha probabilidad pobre, causando un sesgo considerable en la estimación y por el otro, grandes bloques ocasionarían que se tuvieran pocos máximos lo que llevaría a una gran varianza en la estimación.

Teniendo en cuenta las desventajas mencionadas, existe otro método que es comúnmente más utilizado en la práctica ya que permite el aprovechamiento de más datos en el análisis y el cuál es conocido como método de excedencias sobre un umbral (POT, *Peaks Over threshold*). Gracias al resultado (1.3) se tiene una interpretación muy útil que relaciona a la Distribución Generalizada de Valores Extremos con la Distribución Generalizada de Pareto, la cuál se verá a continuación.

□

1.2. Excedencias sobre un umbral (POT)

A continuación se describe el método que considera valores extremos como aquellos valores que exceden un umbral específico y es llamado modelo de Picos o excedencias sobre un umbral.

Partimos de X una variable aleatoria con función de distribución F y extremo derecho x_F . Para $u < x_F$ fijo, decimos que ha ocurrido una excedencia de u si $X > u$ y llamamos excedencia al valor de X y exceso sobre el umbral u a la cantidad $X - u$ que fue excedida por X por encima del umbral.

Se describe este método partiendo de una secuencia de variables aleatorias i.i.d. X_1, X_2, \dots con función de distribución F y considerando a los eventos extremos o datos excedentes como aquellas observaciones x_i que excedan un umbral fijo u lo suficientemente alto; por lo que resulta importante conocer el comportamiento de éstos valores excedentes que está dado por una función de distribución condicional usualmente llamada *función de distribución de Exceso* y que se define a continuación .

Definición. (Función de distribución de excesos): Sea X una variable aleatoria con función y extremo derecho x_F . Para $u < x_F$ fijo,

$$F_u(y) = \mathbb{P}(X - u \leq y | X > u), \quad 0 \leq y \leq x_F - u \quad (1.12)$$

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

Notemos que podemos escribir a F_u en términos de F como se muestra:

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

1.2.1. Distribución Generalizada de Pareto

La distribución empleada para modelar las excedencias sobre un umbral es la distribución Generalizada de Pareto y que definimos a continuación. Para profundizar en sus características y resultados principales puede consultarse [13] y [30].

Definición. (Distribución Generalizada de Pareto (DGP)): La distribución Generalizada de Pareto H_ξ es definida como:

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

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

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

si $\xi < 0$ en el último caso.

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 cuál se le pueden añadir parámetros de localización ν y escala β , para representar a la familia de distribuciones $H_{\xi,\nu,\beta}$, reemplazando al argumento x por $(x - \nu)/\beta$ para $\nu \in \mathbb{R}$ y $\beta > 0$ y ajustando el soporte adecuadamente.

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 cuál es la esperanza del valor del exceso de una variable aleatoria X sobre un umbral u , condicionada a que dicha variable excedió el umbral. Servirá como una herramienta gráfica para la selección de un umbral adecuado para realizar una aproximación con la DGP.

Definición. (Función Media de Excesos). Sea X una variable aleatoria con función de distribución F y extremo derecho x_F . Para $u < x_F$ fijo,

$$e(u) = E(X - u | X > u) \quad (1.13)$$

es llamada la *Función Media de Exceso de X* .

La DGP y la función de excesos se relacionan con el dominio de atracción de la Distribución Generalizada de Valores Extremos mediante la Proposición 1.3. 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

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

De aquí tenemos 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 distribución de los excesos escalados sobre un umbral alto u .

Tal relación junto con la preposición (1.3) nos lleva a uno de los resultados centrales en la Teoría de Valores Extremos. El enunciado de este teorema fue consultado en [30].

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*

$$F_u(y) \sim H_{\xi,\beta}(y), \quad u \rightarrow \infty,$$

donde

$$H_{\xi,\beta}(y) = \begin{cases} 1 - \left(1 + \xi \frac{y}{\beta}\right)^{-1/\xi} & , \xi \neq 0, \\ 1 - e^{-y/\beta} & , \xi = 0, \end{cases} \quad (1.14)$$

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 (1.14) 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 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

Teoría de Valores Extremos Multivariada

El concepto de dependencia es un aspecto fundamental que resume la relación existente entre dos o más variables aleatorias. En la teoría de valores extremos multivariada (TVEM) se trata de entender cómo los eventos extremos de una variable se relacionan con los eventos extremos de otra, así como cuáles son las posibles estructuras de dependencia que describen esa relación y cómo estimarlas.

Así como en el caso univariado, la TVEM también se aborda mediante los métodos de máximos por bloques y excedencias sobre un umbral y su estudio se forma por dos componentes, que son las distribuciones marginales y la estructura de dependencia que presenta el comportamiento conjunto entre las variables en cuestión. En este capítulo se desarrollará la caracterización de los extremos multivariados mediante los métodos de máximos por bloque y excedencias sobre un umbral para modelar las distribuciones marginales, y usando teoría de cópulas para la estructura de dependencia. Además se describirán algunos resultados de cópulas y cópulas de valores extremos bivariados.

2.1. Modelación de valores extremos en \mathbb{R}^d

En forma similar al caso univariado, una forma de modelar eventos extremos multivariados está basada en el método de máximos por bloque, es decir, dada una serie de vectores aleatorios i.i.d. en \mathbb{R}^d se forma una secuencia de bloques del mismo tamaño, donde cada bloque se reduce en su componente máximo y después la colección de bloques de máximos es modelada mediante una distribución multivariada de valores extremos.

Sea $\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$. Si $\mathbf{X}_i = (X_{i,1}, \dots, X_{i,d})$, $i = 1, \dots, n$ una muestra de n vectores aleatorios independientes e idénticamente distribuidos (i.i.d) de dimensión d con función de distribución conjunta \mathbf{F} y funciones de distribución marginales F_j para $j = 1, \dots, d$, se define el vector de componentes máximos (\mathbf{M}_n) como,

$$\mathbf{M}_n = (M_{n,1}, \dots, M_{n,d}) = \left(\max_{1 \leq i \leq n} X_{i,1}, \dots, \max_{1 \leq i \leq n} X_{i,d} \right) \quad (2.1)$$

La función de distribución de \mathbf{M}_n evaluada en \mathbf{x} está dada por,

$$\begin{aligned} \mathbb{P}(\mathbf{M}_n \leq \mathbf{x}) &= \mathbb{P}(M_{n,1} \leq x_1, \dots, M_{n,d} \leq x_d) \\ &= \mathbb{P}(X_{1,1} \leq x_1, \dots, X_{n,1} \leq x_1, \dots, X_{1,d} \leq x_d, \dots, X_{n,d} \leq x_d) \\ &= \mathbb{P}(X_{1,1} \leq x_1, \dots, X_{1,d} \leq x_d, \dots, X_{n,1} \leq x_1, \dots, X_{n,d} \leq x_d) \\ &= \mathbb{P}(X_{1,1} \leq x_1, \dots, X_{1,d} \leq x_d)^n \\ &= \mathbf{F}^n(\mathbf{x}) \end{aligned} \quad (2.2)$$

La penúltima identidad es válida porque los vectores son independientes e idénticamente distribuidos.

Es importante hacer notar que \mathbf{M}_n no necesariamente será una realización observada en la muestra original, sin embargo, según fuentes como ([10], [30], [6]), entre otras, éste es el vector que se considera para el estudio de extremos multivariados.

Del mismo modo que en el caso univariado, para evitar degeneración de la distribución límite de \mathbf{M}_n , se buscan secuencias de vectores normalizantes, \mathbf{a}_n y \mathbf{b}_n con componentes $a_{n,j} > 0$ y $b_{n,j} \in \mathbb{R}$, $j = 1, \dots, d$ tales que cuando $n \rightarrow \infty$,

$$\mathbb{P}\left(\frac{\mathbf{M}_n - \mathbf{b}_n}{\mathbf{a}_n} \leq \mathbf{x}\right) = \mathbf{F}^n(\mathbf{a}_n \mathbf{x} + \mathbf{b}_n) \rightarrow \mathbf{H}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{C}(\mathbf{H}) \quad (2.3)$$

para una función de distribución d -dimensional \mathbf{H} con marginales F_j , $j = 1, \dots, d$, no degeneradas. El teorema de convergencia para máximos multivariados queda parcialmente resuelto por la teoría univariada pues observemos que para cada marginal, si $j = 1, \dots, d$

$$\mathbb{P}\left(\frac{M_{n,j} - b_{n,j}}{a_{n,j}} \leq x_j\right) \rightarrow H_j(x_j), \quad \text{cuando } n \rightarrow \infty \quad (2.4)$$

En caso de existir constantes normalizantes $a_{n,j} > 0$ y $b_{n,j}$ que hagan que el límite H_j sea no degenerado, para cada j , por lo visto en el capítulo anterior, por (1.1) las marginales son necesariamente *DGVE*. Nos resta ver el manejo de la estructura de dependencia y la forma de \mathbf{H} .

Por la siguiente definición podemos agrupar a \mathbf{F} en una clase de distribuciones caracterizadas por \mathbf{H} .

Definición. (Dominio de atracción de DMVE) Si se tiene (2.3) para alguna \mathbf{F} y \mathbf{H} , se dice que \mathbf{F} pertenece al dominio de atracción del máximo de \mathbf{H} , $\mathbf{F} \in \mathcal{D}(\mathbf{H})$, y \mathbf{H} es llamada una distribución multivariada de valores extremos (*DMVE*).

Para el desarrollo de los siguientes capítulos es necesario definir el concepto de cópula y sus resultados principales. La palabra cópula fue empleada por primera vez en un sentido matemático o estadístico por Abe Sklar (1959) en el teorema que lleva su nombre y en donde describe las funciones que unen conjuntamente distribuciones univariadas para formar distribuciones multivariadas, éstas se definen a continuación.

Definición. (Cópula) Una cópula bivariada es una función $C : [0, 1]^2 \rightarrow [0, 1]$ con las siguientes propiedades:

1. Para cada u, v en $[0, 1]$

$$C(u, 0) = 0 = C(0, v),$$

$$C(u, 1) = u \quad \text{y} \quad C(1, v) = v.$$

2. C es 2-creciente, es decir, para cada u_1, u_2, v_1, v_2 en $[0, 1]$ tales que $u_1 \leq u_2$ y $v_1 \leq v_2$,

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0.$$

Estas propiedades definen a una cópula bivariada como una función de distribución en $[0, 1]^2$ con marginales uniformemente distribuidas. Mayor detalle de los resultados de cópulas presentados en esta tesis se puede encontrar en [31], [12].

Uno de los teoremas centrales de la teoría de cópulas es el teorema de Sklar (1959). Gracias a éste, cualquier función de distribución bivariada (multivariada) F se puede escribir como la composición de una cópula C y sus distribuciones marginales F_i como sigue:

Teorema 2.1. Sklar, 1959 *Sea H una función de distribución conjunta bivariada con marginales F y G . Entonces existe una cópula C tal que para cualesquiera x, y en $\bar{\mathbb{R}}$*

$$H(x, y) = C(F(x), G(y)). \quad (2.5)$$

Si F y G son continuas, entonces C es única, en cualquier otro caso, C sólo está determinada de forma única sobre el conjunto $\text{Ran } F \times \text{Ran } G$. donde $\text{Ran } F = \{z : z = F(x) \text{ para alg } x\}$ Además si F, G son funciones de distribución dadas, y C es una cópula, entonces la función H definida mediante (2.5) es una función de distribución conjunta bivariada con marginales F y G .

Nótese que las marginales de una cópula tienen que ser uniformes. Esto se logra mediante la transformación integral de probabilidad (Teorema C.2 del Apéndice C). La demostración del teorema se puede consultar en [31].

Existen diferentes familias de cópulas que se han estudiado ampliamente en la literatura y que tienen características particulares; ejemplos de estas familias son las arquimedianas, elípticas y las de valores extremos. Detalles de estas familias se puede encontrar en [31]; a continuación definiremos la familia de cópulas arquimedianas.

Definición. (Cópulas Arquimedianas) *Una cópula C es llamada Arquimediana si existe una función continua, estrictamente decreciente y convexa $\phi : [0, 1] \rightarrow [0, 1]$ con $\phi(1) = 0$ tal que C puede escribirse de la siguiente manera*

$$C(u, v) = \phi^{-1}(\phi(u) + \phi(v)).$$

donde $\phi^{[-1]}$ es la pseudo inversa de ϕ definida por:

$$\phi^{[-1]}(t) = \begin{cases} \phi^{-1}(t) & , \quad 0 \leq t \leq \phi(0), \\ 0 & , \quad \phi(0) \leq t \leq +\infty. \end{cases}$$

La función ϕ es llamado el generador arquimediano de la cópula C .

EJEMPLO 2.1. *Cópula Gumbel:*

Sea $\phi(t) = (-\log(t))^\theta$ con $\theta > 1$, entonces:

- $\phi(1) = 0, \phi(0) = \infty$, es estrictamente decreciente en $[0, 1]$ pues $\phi'(t) = \frac{-\theta(-\ln t)^{\theta-1}}{t} < 0$.
- $\phi''(t) \geq 0$ en $[0, 1]$ por lo que ϕ es convexa.

Entonces la cópula C está dada por:

$$\begin{aligned} C(u, v) &= \phi^{-1}(\phi(u) + \phi(v)) \\ &= \exp\left(-\left[(-\ln u)^\theta + (-\ln v)^\theta\right]^{\frac{1}{\theta}}\right). \end{aligned}$$

Notemos que las marginales de una distribución bivariada (multivariada) de valores extremos \mathbf{H} son necesariamente DGVE y como todas las DGVE son continuas, por el teorema de Sklar tenemos que \mathbf{H} tiene una cópula única. A dicha cópula se le conoce como cópula de valores extremos. Las cópulas de valores extremos presentan características y propiedades particulares. A continuación se presenta un concepto que permitirá caracterizarlas más adelante.

Definición. (Cópulas Max-estables) Una cópula C es max-estable si para cada número real positivo r y para cualesquiera u, v en $(0, 1)$,

$$C(u, v) = C^r(u^{1/r}, v^{1/r}). \quad (2.6)$$

EJEMPLO 2.2. *La cópula Gumbel es max estable:*

$$\begin{aligned} C(u^t, v^t) &= \exp\left[-\left[(-\ln u^t)^\alpha + (-\ln v^t)^\alpha\right]\right] \\ &= \exp\left[-\left[t^\alpha \left[(-\ln u)^\alpha + (-\ln v)^\alpha\right]\right]^{1/\alpha}\right] \\ &= \left[\exp\left[-\left[(-\ln u)^\alpha + (-\ln v)^\alpha\right]^{1/\alpha}\right]\right]^t \\ &= C^t(u, v) \end{aligned}$$

A continuación se presentará la versión bivariada del método de máximos por bloque. Posteriormente se desarrollarán la caracterización de \mathbf{H} mediante cópulas de valores extremos.

2.1.1. Máximos por bloque bivariados

Sea una sucesión de vectores aleatorios i.i.d (X_i, Y_i) , $i = 1, \dots$ y para cada n sea $M_{(x,n)} = \text{máx}(X_1, \dots, X_n)$ y $M_{(y,n)} = \text{máx}(Y_1, \dots, Y_n)$. Es posible encontrar transformaciones lineales $M'_{(x,n)} = a_n M_{(x,n)} + b_n$ ($a_n > 0$) y $M'_{(y,n)} = c_n M_{(y,n)} + d_n$ ($c_n > 0$) tales que, en el límite cuando $n \rightarrow \infty$, $M'_{(x,n)}$ y $M'_{(y,n)}$ siguen alguno de los tres tipos de DVE, teniendo como distribución conjunta límite \mathbf{H} una distribución bivariada de valores extremos.

Una definición general de una distribución bivariada de valores extremos puede ser dada mediante cópulas. Para entender como se distribuye el vector de componentes máximos necesitamos encontrar una función de distribución \mathbf{H} y lo haremos mediante una cópula $C_{(n)}$.

2.1.2. Modelo Bivariado de Excedencias sobre un umbral

Sea la sucesión de vectores aleatorios i.i.d $(X_1, Y_1), \dots, (X_n, Y_n)$ con distribución conjunta $\mathbf{H}(x, y) = C(F(x), G(y))$ para alguna cópula C , marginales F, G , y \mathbf{H} que está en el dominio de atracción de una distribución multivariada de valores extremos. Buscamos aproximar la cola derecha de la distribución de las excedencias \mathbf{H} que sobrepasen cierto vector de umbrales altos $\mathbf{u} = (u_x, u_y)$. La teoría univariada descrita en el primer capítulo, Sección (1.2), nos dice que, para $x \geq u_x$ y u_x suficientemente alto, la cola de la función de distribución marginal F debe de ser aproximada por una forma funcional a base de DPG, al igual que G ,

$$F(x) = 1 - \lambda_x \left(1 + \xi_x \frac{x - u_x}{\beta_x} \right)^{-1/\xi_x} \quad (2.7)$$

donde $\lambda_x = \bar{F}(u_x)$.

Pero C es desconocida y se debe en sí aproximar en la cola. Supondremos entonces que podemos sustituir la cópula C con un tipo de cópula llamada de valores extremos C_* , que definiremos más adelante, y así tener

$$\mathbf{H}(x, y) \approx C_*(F(x), G(y)), \quad x \geq u_x, y \geq u_y. \quad (2.8)$$

Se completa la especificación del modelo escogiendo una cópula paramétrica de valores extremos tratable y flexible para C_* . Según [30] la cópula Gumbel es una de las cópulas de valores extremos particularmente conveniente.

El argumento para (2.8) será dado más adelante al describir resultados principales de cópulas de valores extremos.

2.2. Cópulas de valores extremos

Sean $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ vectores aleatorios con una función de distribución conjunta H , marginales F y G para X_i y Y_i respectivamente; sea C una cópula cualquiera C y sea $C_{(n)}$ la cópula del vector de componentes máximos $M_{(x,n)} = \max\{X_i\}$ y $M_{(y,n)} = \max\{Y_i\}$.

Primero encontramos las funciones de distribución $F_{(n)}$ y $G_{(n)}$ de $M_{(x,n)}$ y $M_{(y,n)}$, respectivamente. Como para todo X_i , $i = 1, \dots, n$ tenemos que

$$\mathbb{P}(M_{(x,n)} \leq x) = \mathbb{P}(X_i \leq x) = [\mathbb{P}(X_1 \leq x)]^n \quad (2.9)$$

por tanto $F_{(n)}(x) = [F(x)]^n$ y $G_{(n)}(y) = [G(y)]^n$. Entonces

$$\begin{aligned} H_{(n)}(x, y) &= \mathbb{P}(M_{(x,n)} \leq x, M_{(y,n)} \leq y) \\ &= \mathbb{P}(X_i \leq x, Y_i \leq y) \quad , \quad i \in \{1, \dots, n\} \\ &= [H(x, y)]^n \\ &= [C(F(x), G(y))]^n \\ &= \left[C \left[(F_{(n)}(x))^{1/n}, (G_{(n)}(x))^{1/n} \right] \right]^n \end{aligned}$$

Por lo tanto tenemos

$$C_{(n)}(u, v) = C^n(u^{1/n}, v^{1/n}) \quad \text{para } u, v \in [0, 1]. \quad (2.10)$$

Esto es, $C_{(n)}$ es max estable.

Teorema 2.2. Cópula de vector de máximos *Si C es una cópula y n un entero positivo, entonces la función $C_{(n)}$ dada por (2.10) es una cópula. Además, si (X_i, Y_i) , $i \in \{1, \dots, n\}$ son pares de vectores aleatorios independientes e idénticamente distribuidas con cópula C , entonces $C_{(n)}$ es la cópula de $\mathbf{M}_n = (M_{(x,n)}, M_{(y,n)})$.*

El teorema anterior habla sobre cópulas $C_{(n)}$ definidas para cada n . Si consideramos la secuencia $\{C_{(n)}\}$, podemos preguntarnos sobre el límite de ésta, lo cual nos lleva a la noción de cópulas de valores extremos (CVE), que definiremos a continuación.

Definición. (Cópula de valores extremos) Una cópula C_* es una cópula de valores extremos si existe una cópula C tal que

$$C_*(u, v) = \lim_{n \rightarrow \infty} C^n(u^{1/n}, v^{1/n}) \quad (2.11)$$

para u, v en $(0, 1)$. En ese caso, se dice que C pertenece al dominio de atracción de C_* , ($C \in \mathcal{D}(C_*)$).

De manera análoga a la teoría univariada, el concepto de Max-estabilidad caracteriza a las clases de distribuciones límite para el máximo \mathbf{M}_n . El siguiente teorema describe la relación de las cópulas max-estables con las CVE.

Teorema 2.3. CVE-Cópula max-estable Una cópula C es una cópula de valores extremos si y solo si es max-estable.

Demostración. Por la definición (2.6) cada cópula max-estable es una cópula de valores extremos. De forma inversa, si $C_{(*)}$ es una CVE, entonces $C_{(*)}$ satisface (2.11) para alguna cópula C , entonces para cualquier real positivo r ,

$$C_*^r(u)(u^{1/r}, v^{1/r}) = \lim_{n \rightarrow \infty} C^{rn}(u^{1/rn}, v^{1/rn}) = C_*(u, v)$$

por lo tanto C_* es max-estable. □

Recapitulando se tiene que en el caso de dimensiones mayores o iguales a dos, si (2.3) se cumple para alguna función de distribución H con marginales DGVE, entonces la cópula C asociada de manera natural con H como en (2.1) satisface (2.10) y es la cópula de una distribución multivariada de valores extremos.

Una vez caracterizada la clase de distribuciones bivariadas de valores extremos por medio de cópulas de valores extremos, se busca tener una representación más general de ellas. El siguiente teorema nos da una representación de las CVE mediante una medida exponente, el teorema se puede encontrar en [22] y [14].

Teorema 2.4. Una cópula d -variada C es una cópula de valores extremos si y solo si existe una medida finita de Borel H en $\Delta_{d-1} = \{(u_1, \dots, u_d) \in [0, \infty)^d : \sum_j u_j = 1\}$, llamada medida espectral, tal que

$$C(u_1, \dots, u_d) = \exp\{-\ell^*(-\log u_1, \dots, -\log u_d)\}, \quad (u_1, \dots, u_d)^\top \in (0, 1]^d, \quad (2.12)$$

donde la función de dependencia de cola estable $\ell^* : [0, \infty)^d \rightarrow [0, \infty)$ está dada por

$$\ell^*(x_1, \dots, x_d) = \int_{\Delta_{d-1}} \max_{j=1, \dots, d} \{w_j x_j\} dH(w_1, \dots, w_d), \quad (x_1, \dots, x_d)^\top \in [0, \infty)^d. \quad (2.13)$$

La medida espectral H es arbitraria y cumple

$$\int_{\Delta_{d-1}} w_j dH(w_1, \dots, w_d) = 1, \quad j \in \{1, \dots, d\}$$

La función de dependencia de cola estable ℓ^* es convexa, homogénea de grado uno, esto es que para $c > 0$

$$\ell^*(cx_1, \dots, cx_d) = c \ell^*(x_1, \dots, x_d) \quad (2.14)$$

y además, para $(x_1, \dots, x_d) \in [0, \infty)^d$, $\max(x_1, \dots, x_d) \geq \ell^*(x_1, \dots, x_d) \geq x_1 + \dots + x_d$. Para más detalles sobre la función de cola estable y sus propiedades se puede consultar [22] y [14].

A continuación presentamos una definición muy importante para la representación de CVE; ésta definición se puede encontrar en [14].

Definición. (Función de dependencia de Pickands) La función de dependencia de Pickands $A : \Delta_{d-1} \rightarrow [1/d, 1]$ está definida como

$$A(w_1, \dots, w_d) = \frac{\ell^*(x_1, \dots, x_d)}{\sum_{i=1}^d x_i}, \quad (2.15)$$

donde $w_j = \frac{x_j}{\sum_{i=1}^d x_i}$, $j = 1, \dots, d$ para $(x_1, \dots, x_d)^\top \in [0, \infty)^d \setminus \{0\}$, además ocurre,

1. A es convexa
2. $\max\{w_1, \dots, w_{d-1}\} \leq A(w_1, \dots, w_{d-1}) \leq 1$ para $(w_1, \dots, w_{d-1})^\top \in \Delta_{d-1}$

2.2.1. Representación de cópulas de valores extremos

La definición (2.15) junto con (2.12) da lugar a una representación para cópulas de valores extremos expresada en términos de la función de dependencia de Pickands A y que según [14] es la más usada para las CVE. La representación es la siguiente,

$$C(u_1, \dots, u_d) = \exp \left\{ \left(\sum_{i=1}^d \log u_i \right) A \left(\frac{\log u_1}{\sum_{i=1}^d \log u_i}, \dots, \frac{\log u_{d-1}}{\sum_{i=1}^d \log u_i} \right) \right\}. \quad (2.16)$$

De acuerdo a [22] en \mathbb{R}^2 se restringe a la función de dependencia de cola estable ℓ^* en $\Delta_1 = \{(1-t, t) : t \in [0, 1]\}$ y entonces siguiendo (2.12), sabemos que podemos representar a una cópula bivariada de valores extremos como

$$C(u_1, u_2) = \exp\{-\ell^*(-\log(u_1), -\log(u_2))\} \quad (2.17)$$

Tomando $w_1 = 1 - t$ y $w_2 = t$ como se hace en [14], tenemos por (2.15) que la función de dependencia de Pickands, para el caso bivariado, queda definida como

$$\begin{aligned} A(1-t, t) &= \frac{\ell^*(-\log(u_1), -\log(u_2))}{-\log(u_1) - \log(u_2)} \\ &= \frac{-\ell^*(\log(u_1), \log(u_2))}{-(\log(u_1) + \log(u_2))} \\ &= \ell^*\left(\frac{\log(u_1)}{\log(u_1 u_2)}, \frac{\log(u_2)}{\log(u_1 u_2)}\right). \end{aligned}$$

Sea

$$t = \frac{\log(u_2)}{\log(u_1 u_2)}. \quad (2.18)$$

Por comodidad suele escribirse $A(t)$ en vez de $A(1-t)$, como se puede ver en [14], de esta forma tenemos

$$A(t) = \ell^*\left(\frac{\log(u_1)}{\log(u_1 u_2)}, \frac{\log(u_2)}{\log(u_1 u_2)}\right) = \ell^*(1-t, t), \quad (2.19)$$

En la figura (2.1) se muestra un gráfico de función de Pickands en \mathbb{R}^2 . Por la propiedad 2) de la definición, se encuentra definida en el área triangular que se muestra en la gráfica, la línea amarilla punteada corresponde al caso de la cópula producto o independencia y la curva respresenta un ejemplo de una función de dependencia.

Los límites de la función de dependencia Pickands A tienen significados particulares. El borde superior $A(t) = 1$ que se observa en la Figura (2.1) corresponde al caso de independencia, correspondiendo a la cópula producto $C(u_1, u_2) = u_1 u_2 = \prod^2(u_1, u_2)$. La frontera inferior (los otros dos lados del triángulo) $A(t) = \max\{1-t, t\}$ corresponde al caso de dependencia perfecta, $C(u_1, u_2) = \max\{u_1, u_2\} = M^2(u_1, u_2)$.

Teorema 2.5. (Cópulas bivariadas de valores extremos) Una cópula bivariada C es una cópula de valores extremos si y solo si

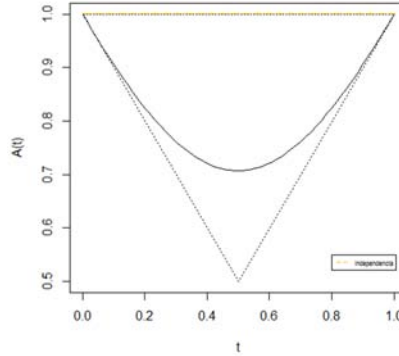


Figura 2.1: Función de dependencia de Pickands.

$$\begin{aligned} C(u_1, u_2) &= \exp \left\{ (\log(u_1) + \log(u_2)) A \left(\frac{\log(u_2)}{\log(u_1 u_2)} \right) \right\} \\ &= (u_1 u_2)^{A(\log(u_2)/\log(u_1 u_2))}, \end{aligned} \quad (2.20)$$

con $(u_1, u_2) \in (0, 1]^2 \setminus \{(1, 1)\}$ y $A(w) = \int_0^1 \max((1-x)w, x(1-w)) dH(x)$ para una medida H en $[0, 1]$ que satisface

$$\max(w, 1-w) \leq A(w) \leq 1, \quad 0 \leq w \leq 1, \quad (2.21)$$

y es convexa. Por el contrario, una función diferenciable y convexa $A(w)$ que satisface (2.21) puede ser usada para construir una cópula de valores extremos usando (2.20)

Demostración. De acuerdo a (2.19) tenemos que

$$A(t) = \frac{\ell^*(-\log(u_1), -\log(u_2))}{-\log(u_1) - \log(u_2)}.$$

si y solo si:

$$\ell^*(-\log(u_1), -\log(u_2)) = -\log(u_1 u_2) A(t). \quad (2.22)$$

Y así sustituyendo (2.22) en la representación bivariada de una CVE presentada en (2.17) tenemos

$$C(u_1, u_2) = \exp\{\log(u_1 u_2) A(t)\},$$

con $t = \log(u_2)/\log(u_1 u_2)$ de acuerdo a (2.18). \square

2.2.2. Modelos paramétricos

A continuación se presentan las cópulas de valores extremos clásicas en la literatura.

Cópula Gumbel o logística

Gumbel observó que a partir de $F_1(x_1)$, $F_2(x_2)$ funciones de valores extremos univariadas, las distribuciones

$$[-\log F_\theta(x_1, x_2)]^\theta = [-\log F_1(x_1)]^\theta + [-\log F_2(x_2)]^\theta, \quad \theta \geq 1 \quad (2.23)$$

definen a una distribución bivariada de valores extremos $F_\theta(x_1, x_2)$. Una prueba de este resultado se puede encontrar en [7]. De (2.23) tenemos

$$F_\theta(x_1, x_2) = \exp\{-[(-\log F_1(x_1))^\theta + (-\log F_2(x_2))^\theta]^{1/\theta}\},$$

que con $u_i = F_i(x_i)$ obtenemos la cópula Gumbel

$$C(u_1, u_2) = \exp\{-[(-\log(u_1))^\theta + (-\log(u_2))^\theta]^{1/\theta}\}, \quad (2.24)$$

$(u_1, u_2)^\top \in [0, 1]^2$, $\theta \geq 1$. Por (2.20) y la restricción usual $t = \log(u_2)/\log(u_2 u_1)$ resulta en la función de dependencia de Pickands

$$A(t) = [t^\theta + (1-t)^\theta]^{1/\theta}, \quad (2.25)$$

para $t \in [0, 1]$ y $\theta \in [1, \infty]$. El parámetro θ de la cópula Gumbel puede ser interpretado como medida de dependencia.

Cópula Tawn o logística asimétrica

La cópula Tawn puede considerarse como una extensión de la cópula Gumbel. La cópula Gumbel es simétrica, es decir, satisface $C(u_1, u_2) = C(u_2, u_1)$ lo cual lleva a las variables X_1, X_2 a ser intercambiables cuando se modelan por C , (ver definición). En algunas aplicaciones no es indicado asumir esta propiedad por lo que Tawn (1988) le dió flexibilidad a la cópula Gumbel incorporando dos parámetros ψ_1, ψ_2 . La cópula Tawn o modelo logístico asimétrico [41] tiene la función de dependencia de Pickands

$$A(t) = (1 - \psi_2)(1 - t) + (1 - \psi_1)t + [(\psi_1(1 - t))^\theta + (\psi_2 t)^\theta]^{1/\theta} \quad (2.26)$$

para $t \in [0, 1]$, $0 \leq \psi_1, \psi_2 \leq 1$ y $\theta \in [1, \infty)$. Los parámetros ψ_1, ψ_2 pueden ser interpretados como el grado de asimetría de la función A . Mientras tanto θ tiene la misma interpretación

que la cópula Gumbel y por lo tanto es el principal determinante de la dependencia. La versión de $A(t)$ de un parámetro se obtiene tomando $\psi_1 = \psi_2 = 1$. Por último la expresión de la cópula Tawn tiene la siguiente forma

$$C(u_1, u_2) = u_1 u_2 \exp\{-[(-\log(u_1))^\theta + (-\log(u_2))^\theta]^{1/\theta}\}, \quad (2.27)$$

Cópula Galambos o negativa logística

Sea C_θ que denote a una cópula arquimediana y \widehat{C}_θ es la correspondiente cópula de supervivencia. Entonces \widehat{C}_θ es la función de distribución del vector $1 - U_1, \dots, 1 - U_d$. Como en el caso de la cópula Gumbel, solo si el límite

$$\lim_{s \downarrow 0} \frac{\phi(s)}{s\phi'(s)} = \delta \in [0, \infty]$$

existe, la adecuada convergencia del vector de máximos de \widehat{C}_θ está garantizada y se tiene que $\widehat{C}_\theta \in \mathcal{D}(C^{Galambos})$. La función de dependencia de Pickands de la cópula Gumbel está dada por

$$A(t) = 1 - [t^{-\delta} + (1-t)^{-\delta}]^{-1/\delta}, \quad (2.28)$$

para $t \in [0, 1]$ y $\delta \in (0, \infty]$. La parametrización es elegida de manera que al incrementarse el parámetro δ nos lleve a un incremento en la dependencia.

Cópula Hüsler-Reiss

Sea C_F una cópula Gaussiana, es decir

$$C_F(u_1, u_2) = C_\rho(u_1, u_2) = \Phi_\rho(\Phi^{-1}(u_1), \Phi^{-1}(u_2))$$

con $\rho \in (-1, 1)$ denota el coeficiente de correlación, Φ_ρ la distribución normal estándar bivarada con coeficiente de correlación ρ y Φ^{-1} la inversa de la función de distribución normal estándar univariada. Se sabe desde Sibuya 1959 [36], que si $\rho < 1$, $C_\rho \in \mathcal{D}(\Pi)$, es decir, la distribución normal bivarada está en el dominio de atracción de la cópula producto. Sin embargo, considerando el caso donde ρ se permite que dependa del tamaño de muestra n y que incrementa con el crecimiento de n de tal forma que $\rho_n \rightarrow 1$ cuando $n \rightarrow \infty$. Entonces si $(1 - \rho_n) \log n \rightarrow \lambda^2 \in [0, \infty]$ cuando $n \rightarrow \infty$, se tiene que $C_{\rho_n} \in \mathcal{D}(C^{HR})$, donde C^{HR} es la cópula Hüsler Reiss [25], que está definida mediante su función de dependencia de Pickands

$$A(t) = (1-t)\Phi(z_{1-t}) + t\Phi(z_t), \quad (2.29)$$

donde $z_t = \left(\frac{1}{\lambda} + \frac{\lambda}{2} \log \frac{t}{1-t}\right)$ para $t \in [0, 1]$, $\lambda \in (0, \infty)$ y Φ es la función de distribución normal estándar.

Cópula t-EV

La cópula t-EV recibe su nombre al ser el límite asintótico del vector de máximos de variables aleatorias independientes distribuidas de acuerdo a la distribución t de Student bivariada, es decir, la cópula t se encuentra en el dominio de atracción de la cópula t-EV. Los dos parámetros de la cópula t-EV son los mismos a los de la cópula t: los grados de libertad $\nu > 0$ y el parámetro de correlación $\rho \in (-1, 1)$. Su función de dependencia de Pickands está dada por

$$A(t) = (1 - t) \cdot T_{\nu+1}(z_{1-t}) + t \cdot T_{\nu+1}(z_t), \quad (2.30)$$

donde $z_t = (1 + \nu)^{1/2} \left(\left[\frac{t}{1-t} \right]^{1/\nu} - \rho \right) (1 - \rho^2)^{-1/2}$ para $t \in [0, 1]$, $\nu > 0$, $\rho \in (-1, 1)$ y T_ν denotando la función de distribución t de Student univariada con ν grados de libertad.

Capítulo 3

Herramientas estadísticas para el análisis de los datos

El análisis de datos requiere del uso de herramientas estadísticas descriptivas e inferenciales para conocer el comportamiento de las variables de interés y hacer inferencia acerca de la población de la que provienen los datos en estudio. En este capítulo se describirán las herramientas estadísticas utilizadas para el análisis de los datos y para el desarrollo de los modelos de eventos extremos univariado y bivariado de máximos por bloque y excedencias sobre un umbral.

3.1. Ajuste de valores extremos univariado

Como primer paso para la realización de los modelos univariado y bivariado de eventos extremos de máximos por bloque y excedencias sobre un umbral, se inicia el estudio de todos los datos mediante un análisis exploratorio individual de cada variable de interés. Las medidas estadísticas descriptivas nos ayudan a identificar las características principales de una muestra de datos de acuerdo a los valores que toman y así analizar sus características como posición o localización, dispersión y forma. Por su parte, las herramientas gráficas permiten tener una visualización de los datos.

Las herramientas gráficas utilizadas en este trabajo para el análisis estadístico de los datos fueron histogramas, QQ plots y Box Plots, así como medidas estadísticas descriptivas

como la media muestral, la desviación estándar y los coeficientes de asimetría y curtosis. En el apéndice D se describen éstas mismas.

3.1.1. Estimación de parámetros

La DGVE y DGP pueden ajustarse usando varios métodos incluyendo el método de momentos y máxima verosimilitud. El método usado en este trabajo es el de máxima verosimilitud y se encuentra disponible en los paquetes estadísticos que utilizan métodos numéricos para obtener los estimadores de máxima verosimilitud ya que no es posible obtener soluciones analíticas explícitas de las ecuaciones máximo verosímiles.

Estimación de los parámetros de la DGVE

Escribiendo $g_{\xi,\mu,\sigma}$ para la función de densidad de la DGVE con $\theta = (\xi, \mu, \sigma)$ su correspondiente vector de parámetros, M_{n1}, \dots, M_{nm} una muestra de m máximos de tamaño de bloque n y para $\xi \neq 0$ se tiene la función de log-verosimilitud de la siguiente forma:

$$\begin{aligned} l(\theta) &= \sum_{i=1}^m \ln g_{\xi,\mu,\sigma}(M_{ni}) \\ &= -m \ln \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^m \ln \left(1 + \xi \frac{M_{ni} - \mu}{\sigma}\right) - \sum_{i=1}^m \left(1 + \xi \frac{M_{ni} - \mu}{\sigma}\right)^{-1/\xi} \end{aligned} \quad (3.1)$$

con $1 + \xi(M_{ni} - \mu)/\sigma > 0$ para $i = 1, \dots, m$. Cuando $\xi = 0$ la función de log-verosimilitud queda como

$$l(\mu, \sigma) = -m \ln \sigma - \sum_{i=1}^m \left(\frac{M_{ni} - \mu}{\sigma}\right) - \sum_{i=1}^m \exp \left\{ - \left(\frac{M_{ni} - \mu}{\sigma}\right) \right\}$$

Para la inferencia de los niveles de retorno se procede substituyendo las estimaciones obtenidas por (3.1) en la ecuación de los niveles de retorno (1.10), de tal forma que la estimación del nivel de retorno x_p asociado al periodo de retorno $1/p$ con $0 < p < 1$ es obtenido usando máxima verosimilitud para $\hat{\xi} \neq 0$ mediante

$$\hat{x}_p = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}} \left(1 - y_p^{-\hat{\xi}}\right),$$

y para $\hat{\xi} = 0$

$$\hat{x}_p = \hat{\mu} - \hat{\sigma} \ln y_p,$$

con $y_p = -\ln(1 - p)$. La interpretación de x_p es el nivel que se espera sea excedido en promedio una vez cada $1/p$ años.

Estimación de los parámetros de la DGP

De igual forma a la DGVE, se utiliza el método de máxima verosimilitud para estimar los parámetros de la DGP. Una vez elegido un umbral apropiado u , para las n_u observaciones x_j que exceden el umbral u se obtiene el valor de los excesos de dichas excedencias obteniendo $y_j = x_j - u$, posteriormente se estiman los parámetros de la de la función de distribución que modela a los excesos y_j . Por el Teorema de Pickands -Balkema-de Haan sabemos que la distribución de los excesos y_1, \dots, y_{n_u} puede aproximarse mediante la DGP.

Escribiendo $h_{\xi, \beta}$ para la función de densidad de la DGP, se tiene para $\xi \neq 0$ la siguiente función de log-verosimilitud:

$$\begin{aligned} l(\xi, \beta) &= \sum_{j=1}^{n_u} \ln h_{\xi, \beta}(y_j) \\ &= -n_u \ln \beta - \left(1 + \frac{1}{\xi}\right) \sum_{j=1}^{n_u} \ln \left(1 + \xi \frac{y_j}{\beta}\right) \end{aligned} \quad (3.2)$$

con $(1 + \xi y_j / \beta) > 0$. Para $\xi = 0$ la función (3.2) queda como:

$$l(\beta) = -\beta \ln \beta - \frac{1}{\beta} \sum_{j=1}^{n_u} y_j.$$

3.1.2. Diagnóstico y evaluación del modelo

Las técnicas de bondad de ajuste tratan de medir la conformidad o discrepancia de los datos muestrales con una distribución hipotética, es decir si la muestra fue extraída de una población que sigue esta distribución planteada. Las pruebas de bondad de ajuste utilizadas para evaluar los modelos univariados de valores extremos analizados en esta tesis fueron la prueba de Kolmogorov-Smirnov (KS) y la prueba de Anderson Darling. Además se emplearon herramientas gráficas como P-P plots y Q-Q plots.

Prueba de Kolmogorov-Smirnov (KS)

Esta prueba está basada en las diferencias entre la función de distribución empírica y la función de distribución especificada. Una característica atractiva de esta prueba es que no

depende de la función de distribución que se está probando. Sin embargo, presenta algunas desventajas importantes pues solo se aplica a distribuciones continuas, suele ser más sensible cerca del centro de la distribución que en las colas y la distribución debe de ser completamente especificada con sus parámetros correspondientes.

El estadístico de prueba Kolmogorov-Smirnov se define como

$$D_n = \sup_x |\tilde{F}(u_{(j)}) - F_0(u_{(j)})|,$$

donde $u_{(j)} = x_{(j)}$ y $x_{(1)} < x_{(2)} < \dots < x_{(n)}$ es la muestra ordenada, \tilde{F} es la función de distribución empírica o muestral y F_0 es la función de distribución puesta a prueba, que para este trabajo son la DGVE, $G_{\hat{\xi}, \hat{\mu}, \hat{\sigma}}$ y la DGP, $H_{\hat{\xi}, \hat{\beta}}$.

Se desea probar:

H_0 : Los datos representan observaciones provenientes de la distribución especificada F_0 .

H_a : Los datos no provienen de la distribución especificada F_0 .

La regla de decisión para aceptar o rechazar la prueba se lleva a cabo mediante el valor del estadístico de prueba y el nivel de significancia α o también mediante el *p-value*. Entonces la hipótesis nula H_0 es rechazada al nivel de significancia α si el valor del estadístico D_n excede el valor del cuantil $1 - \alpha$ de la distribución de D_n o valor crítico al nivel α . El *p-value* es el más pequeño nivel de significancia para el cual la hipótesis nula sería rechazada, si $p\text{-value} \leq \alpha$, se rechaza H_0 . Además también brinda una idea de qué tan fuerte es la evidencia en contra de la hipótesis nula.

Anderson-Darling

Esta prueba es similar a la prueba de Kolmogorov-Smirnov (KS), pero presenta una mejora para las colas de la distribución de F. Se considera más poderosa que la prueba de KS aunque hace uso de la distribución hipotética a probar en el cálculo de los valores críticos del estadístico de prueba, esto tiene la ventaja de permitir una prueba más sensible y la desventaja de que los valores críticos deben calcularse para cada distribución.

El estadístico de prueba Anderson-Darling se define como

$$A_n^2 = -n - \frac{1}{n} \sum_{j=1}^n (2j-1) [\ln u_{(j)} + \ln (1 - u_{(n-j+1)})]$$

donde $u_{(j)} = F_0(x_{(j)})$ y $x_{(1)} < x_{(2)} < \dots < x_{(n)}$ es la muestra ordenada.

Se desea probar:

H_0 : Los datos representan observaciones provenientes de la distribución especificada F_0 .

H_a : Los datos no provienen de la distribución especificada F_0 .

La regla de decisión para aceptar o rechazar la prueba se lleva a cabo de manera análoga a la prueba de KS.

Evaluación gráfica

Al ordenar las observaciones máximas de m bloques $x_{(1)} \leq \dots \leq x_{(m)}$ la función de distribución empírica evaluada en $x_{(i)}$ está dada por

$$\tilde{G}(x_{(i)}) = i/(m+1).$$

Sustituyendo las estimaciones obtenidas mediante (3.1) en la ecuación de la DGVE (1.9) se tiene el correspondiente modelo de máximos por bloque estimado

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

Si el modelo de la DGVE es adecuado, $\hat{G}(x_{(i)}) \approx \tilde{G}(x_{(i)})$ para cada i , entonces la gráfica de probabilidades consiste en los puntos de la forma

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

que deben alinearse con una recta diagonal. De igual forma se tienen los puntos de la gráfica de cuantiles

$$\left\{ \left(\hat{G}^{\leftarrow}(i/(m+1)), x_{(i)} \right), i = 1, \dots, m \right\}$$

donde

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

Para el caso de del modelo de excedencias, suponiendo o tomando un umbral lo suficientemente alto u , los excesos $y_{(1)} \leq \dots \leq y_{(n_u)}$ y el modelo estimado mediante la DGP \hat{H} , la gráfica de probabilidades consiste en los puntos:

$$\left\{ \left(j/(n_u + 1), \hat{H}(y_{(j)}) \right), j = 1, \dots, n_u \right\}$$

donde

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

para $\widehat{\xi} \neq 0$. La gráfica de cuantiles consiste en los puntos

$$\left\{ \left(\widehat{H}^{\leftarrow}(j/(n_u + 1)), y_{(j)} \right), j = 1, \dots, n_u \right\}$$

donde

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

Para más detalle de los modelos de diagnóstico y ajuste de DGVE se puede consultar [10].

3.2. Ajuste bivariado de valores extremos

En el caso de dos variables, los métodos gráficos pueden ser herramientas de exploración útiles para inferir en las características de los datos.

3.2.1. Estimación de parámetros

Existen varios enfoques para la estimación de parámetros de una función cópula, ejemplos pueden consultarse en [15]. El uso de máxima verosimilitud es el método de estimación más directo y utilizado donde se estiman simultáneamente todos los parámetros.

El método de inferencia para marginales (*Inference for margins IFM*) es un enfoque que se lleva acabo en dos etapas. Los marginales se estiman en la primera etapa y en la segunda etapa la distribución marginal estimada se evalúa en la función cópula para luego estimar el parámetro de dependencia. Una característica atractiva del método IFM es que uno puede obtener cópulas para las cuales la estructura de dependencia es independiente de las distribuciones marginales y es computacionalmente más atractivo. A continuación se describe brevemente este método.

Supongamos que una familia paramétrica de cópulas (C_θ) es considerada como un modelo para la dependencia entre dos variables aleatorias X y Y . Dada una muestra aleatoria $(X_1, Y_1), \dots, (X_n, Y_n)$ de (X, Y) . La estimación de θ se obtiene mediante máxima verosimilitud de la función de la forma

$$l(\theta) = \sum_{i=1}^n \log \left[C_\theta \{ \widehat{F}(X_i), \widehat{G}(Y_i) \} \right]$$

substituyendo $\widehat{F} = F_{\delta_n}$ y $\widehat{G} = G_{\eta_n}$ donde (\widehat{F}_{δ_n}) y (\widehat{G}_{η_n}) son familias paramétricas adecuadas para las marginales y δ_n, η_n estimaciones por máxima verosimilitud estándar de sus parámetros derivados por los valores observados de X y Y respectivamente y se utilizan métodos numéricos para obtener las estimaciones de θ . Mayor detalle del método se puede encontrar en [26] y [27].

3.2.2. Diagnóstico y evaluación del modelo

Pruebas de bondad de ajuste y medidas estadísticas

Bajo la suposición de que la estructura de dependencia de una distribución continua bivariada pertenece a la clase de distribuciones de valores extremos, sabemos que éstas se caracterizan por su función de dependencia de Pickands (2.20).

Una prueba de bondad de ajuste desarrollada para cópulas de valores extremos se basa en el estadístico de Cramér-von Mises midiendo la distancia entre la distribución empírica comparada con una estimación paramétrica de la función de dependencia de Pickands y dos estimadores no paramétricos de la misma estudiados y desarrollados en [16].

Sea $(X_1, Y_1), \dots, (X_n, Y_n)$ una muestra aleatoria de función de distribución conjunta desconocida H cuya cópula subyacente es de la forma (2.20) con función de dependencia de Pickands A . Para probar la hipótesis de que A pertenece a una clase paramétrica específica \mathcal{A} , es decir probar la hipótesis nula

$$H_0 : A \in \mathcal{A} = \{A_\theta : \theta \in \mathcal{O}\}$$

donde \mathcal{O} es un subconjunto abierto de \mathbb{R}^d para algún entero d se puede proceder comparando un estimador no paramétrico de A_n de A con un estimador paramétrico A_{θ_n} . Para este propósito se pueden utilizar varias medidas de distancia pero se considera que el estadístico de Cramér-von Mises definido como

$$S_n = \int_0^1 n |A_n(t) - A_{\theta_n}(t)|^2 dt \quad (3.3)$$

generalmente conduce a pruebas más poderosas que, por ejemplo, el estadístico de Kolmogorov-Smirnov [18].

Para la estimación paramétrica de A_{θ_n} (la función de Pickands que es utilizada en este trabajo) se emplea un desarrollo vía máxima pseudo verosimilitud considerada en ([17]). Existen otros métodos como los de inversión de las medidas de dependencia Tau de Kendall y rho de Spearman como se muestra en ([19]). Detalles de los estadísticos no paramétricos son desarrollados por Pickands ([32]).

Otra herramienta empleada para elegir un modelo paramétrico de cópula dentro de una serie de modelos candidatos es el criterio de información de Akaike (*Akaike information criterion*, AIC), desarrollado por Hirotugu Akaike (1973) definido como:

$$AIC = -2l(\hat{\theta}) + 2k$$

donde $l(\hat{\theta})$ es el valor maximizado de la función de máxima verosimilitud y k es el número de parámetros.

Uno de los beneficios del AIC es que es computacionalmente más eficiente que otros métodos de selección de cópula ([44]) y entonces cuando hay varios modelos de cópulas y queremos saber cuál ajusta mejor a los datos, se puede calcular el AIC para cada modelo y aquél con menor valor se considerará como el que brinda mejor ajuste.

Herramientas gráficas

Una manera útil de comparar el gráfico de una cópula, es a través de su diagrama de contorno que es la representación gráfica de sus curvas de nivel. Éstas últimas son conjuntos en $[0, 1]^n$ dados por $C(u, v) = k$ con k constante, es decir para cualquier cópula C y un $k \in [0, 1]$ dado, el gráfico de la curva de nivel es el conjunto $\{(u, v) \in [0, 1]^2 : C(u, v) = k\}$.

Para evaluar el ajuste de las cópulas estimadas, se puede comparar gráficamente la distancia entre las funciones de dependencia de Pickands teóricas de las cópulas de valores extremos estimadas y los estimadores no paramétricos de las funciones de dependencia de Pickands.

Capítulo 4

Análisis de valores extremos del Oleaje mediante altura de Ola y velocidad de viento

En este capítulo se aplicará la teoría expuesta en los capítulos anteriores para analizar los eventos extremos de un conjunto de datos bivariados de una estación localizada en el Golfo de México. Se utilizará un modelo de cópulas de valores extremos junto con distribuciones generalizadas de valores extremos para el ajuste de las distribuciones marginales.

Para contar con mediciones directas provenientes de instrumentos oceanográficos que midan dos o más datos meteorológicos distintos de manera independiente se consideró la información del banco histórico de la *National Oceanic and Atmospheric Administration* (NOAA) y la *National Data Buoy Center* (NDBC) que forman parte de la *National Weather Service*, en un punto geográfico localizado en el Océano Atlántico dentro del Golfo de México y reactivamente cercano a la costa de Tamaulipas y Texas.

La NDBC-NOAA diseña, desarrolla, opera y mantiene una red mundial de colección de datos de boyas y estaciones costeras. Todas las estaciones miden velocidad de viento, dirección y ráfaga, presión barométrica y la temperatura del aire, así como la temperatura superficial del mar, altura y periodo de olas.

Se utilizaron las bases de datos históricas de la estación 42002 WEST GULF 207 NM East of Brownsville, TX ubicada en las coordenadas $26^{\circ}5'$ Latitud Norte, $93^{\circ}45'$ Longitud Oeste que presenta registros cada hora durante los años de 1980 a 2014. Las variables ocupadas de las bases de datos para este trabajo son altura de ola significativa (metros) y velocidad de viento (metros sobre segundo). Las bases de datos son públicas y están disponibles en el sitio de internet de la NDBC [3].



Figura 4.1: Localización geográfica de la estación 42002 de la NDBC.

La altura un tercio o significativa, está definida como la media aritmética del 33% de las alturas de ola más altas, particularmente en la NDBC, durante un periodo de muestreo de 20 minutos.

Las variables que se ocuparán en el análisis se eligieron con base en los argumentos y propósitos planteados en la introducción de esta tesis, se buscaban componentes importantes para la caracterización del oleaje y la existencia de una relación mutua entre los mismos. Se eligió la altura de ola pues, de acuerdo con [37], es uno de los parámetros de mayor importancia en ingeniería de costas así como en oceanografía, y la velocidad del viento pues

de manera natural se espera estén muy correlacionados, esto de considerar que el oleaje se propaga principalmente en dirección del viento y que cuanto mayor es la altura de las olas, mayor es la cantidad de energía que pueden extraer del viento. Además la altura de las olas depende principalmente de tres parámetros del viento, que son su velocidad, su persistencia en el tiempo y su dirección. Ésto puede consultarse a mayor detalle en [1].

4.1. Análisis exploratorio de los datos

Como acción inicial al estudio de los datos ocuparemos gráficos y medidas estadísticas para tener un análisis descriptivo de la información.

Se comenzará analizando las variables de manera individual obteniendo un resumen estadístico de la información mediante el cálculo de la media, desviación estándar, sesgo y curtosis, así como los gráficos de histogramas, Box Plot y QQPlot para visualizar el comportamiento de los datos.

	Media	Sd	Curtosis	Sesgo
H	1.236	0.7316	7.0139	1.4989
V	6.421	3.1098	3.0733	0.3675

Altura de oleaje (H) y Velocidad de viento (V)

Tabla 4.1: Resumen estadístico de los datos.

Podemos destacar que la forma de la distribución de ambas variables son curvas con sesgo positivo lo que nos indica que la distribución se despliega hacia la derecha pues el mayor grado de dispersión de los datos está hacia el extremo derecho de ésta. Además, la curtosis de ambas es mayor a 3 haciendo referencia a distribuciones leptocúrticas que son aquellas cuya forma es más apuntada y con colas más anchas en comparación con una distribución normal.

En los histogramas y Box Plot de la figura 4.2 de las variables se puede visualizar lo explicado en el párrafo anterior, apreciándose el considerable grado de dispersión de las observaciones hacia la cola derecha de la distribución.

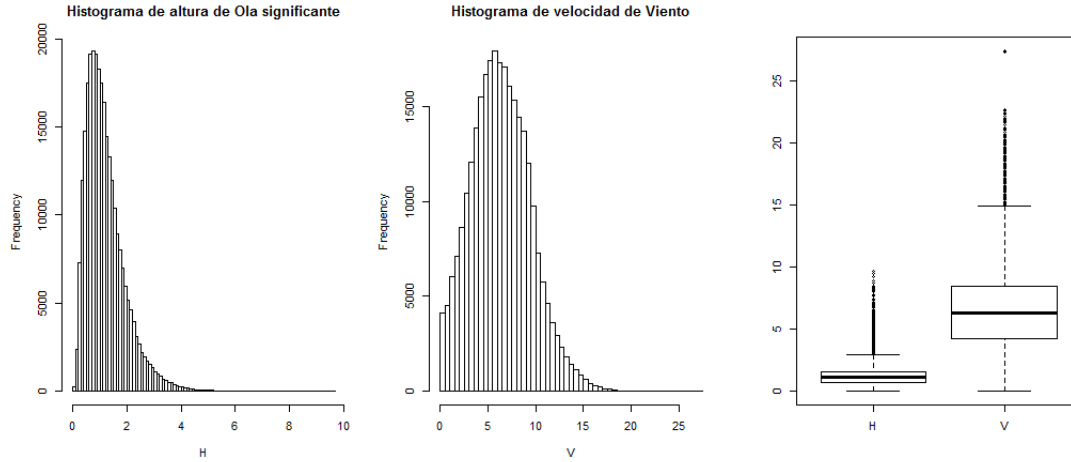


Figura 4.2: Histogramas de altura de Ola significativa y velocidad de Viento y Box Plot.

Sin perder de vista que buscamos analizar el comportamiento conjunto de la altura de ola significativa y la velocidad de viento en el fenómeno del oleaje, necesitamos explorar la relación entre ambas.

Del gráfico de dispersión 4.3 se puede suponer cierto grado de correlación positiva entre las variables cuando estas presentan niveles más altos, hecho que es de esperarse puesto que entre los factores que determinan la altura del oleaje están la fuerza del viento [37].

En primera instancia podemos afirmar que tanto la variable altura de ola significativa como la velocidad de viento presentan cola derecha pesada, esto puede vincularse teniendo presente que el oleaje es uno de los procesos derivados del sistema climático y, como tal, lleva asociado una ocurrencia de eventos extremos que procederemos a analizar basados en la Teoría de Valores Extremos.

4.2. Análisis de eventos extremos en el oleaje.

A continuación analizaremos los eventos extremos del oleaje con nuestra base de datos mediante la aplicación de la TVEB. Comenzaremos trabajando el enfoque de máximos por bloque el cual funje como el método eje para el desarrollo de la TVE. Primero se trabajará la parte univariada mediante el ajuste de las marginales con distribuciones generalizadas

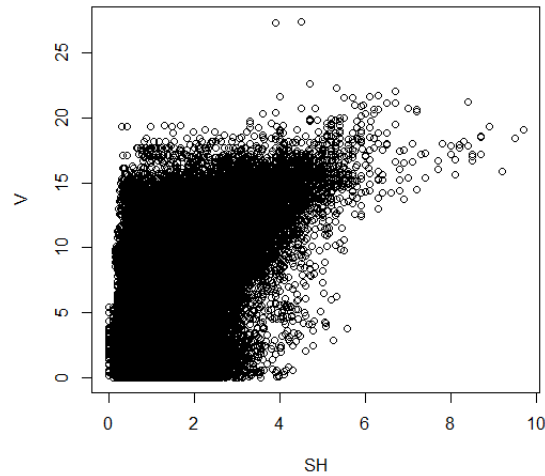


Figura 4.3: Diagrama de dispersión de Altura de ola y Velocidad de viento

de valores extremos y posteriormente se abordará el caso bivariado mediante un modelo de cópulas de valores extremos. Para finalizar se trabajará el enfoque de excedencias sobre un umbral siguiendo el mismo esquema, ajustando distribuciones generalizadas de Pareto a las marginales y después la cópula de valores extremos para la distribución conjunta.

4.2.1. Máximos por bloque

Para el desarrollo de este enfoque necesitamos elegir el tamaño de los bloques al que se tomarán las observaciones máximas. Usualmente se ocupan bloques con observaciones anuales ver [10] y [30], pero en nuestro caso debido a que contamos con un histórico de datos relativamente chico y con el propósito de tener con una submuestra de observaciones más robusta, se eligieron bloques semestrales. De esta forma, desde el año de 1980 hasta el 2014 se agruparon las observaciones en el primer y segundo semestre de cada año para contar con una submuestra con 70 observaciones máximas.

Es importante comprobar que contamos con una muestra indicada y congruente con los supuestos de la TVE. Efectivamente, la submuestra de observaciones máximas semestrales de ambas variables se comportan de manera independiente y estacionaria como se puede ver en los gráficos de autocorrelación ACF y PACF de la figura 4.4. Esto fue afortunado pues,

al elegir máximos de bloques con tamaño temporal menores a un año, se ven afectados por las condiciones ambientales de las estaciones a lo largo del año.

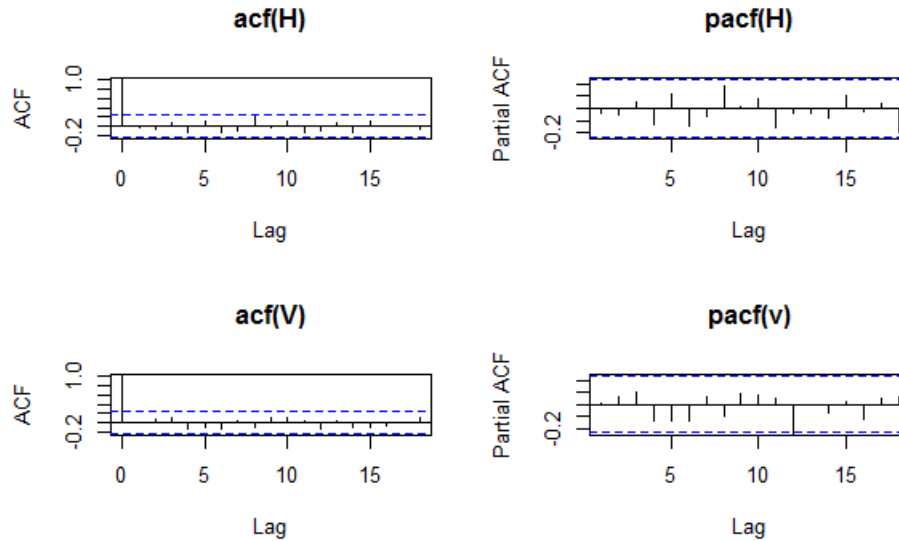


Figura 4.4: ACF y PACF de la submuestra de máximos semestrales de altura de Ola significativa y velocidad de Viento.

Ajuste marginal

El paso inicial para nuestra modelación es determinar las distribuciones paramétricas apropiadas para nuestros datos univariados, o mejor dicho, encontrar las distribuciones marginales que mejor los describen. Una vez contempladas las hipótesis anteriores y de acuerdo a las premisas de la TVE desarrolladas en el Capítulo 1, nuestras observaciones de máximos semestrales se comportan como observaciones independientes de una distribución generalizada de valores extremos.

La estimación de los parámetros de las DGVE para la altura de ola significativa y velocidad de viento mediante máxima verosimilitud y los intervalos de confianza para cada parámetro se presenta en la tabla 4.2. Se puede consultar [40], [23], [8] para mayor detalle sobre el paquete estadístico.

De acuerdo a las estimaciones para las marginales altura de ola significativa y velocidad de

		Estimadores	Intervalos
Altura	$\hat{\xi}$	-0.00084	(-0.12973 , 0.12805)
	$\hat{\mu}$	4.60387	(4.38619 , 4.82155)
	$\hat{\sigma}$	0.84983	(0.69723 , 1.00242)
Velocidad	$\hat{\xi}$	-0.2105	(-0.29699 , -0.12392)
	$\hat{\mu}$	16.4550	(15.66378 , 17.24621)
	$\hat{\sigma}$	3.1449	(2.64117 , 3.64863)

Tabla 4.2: Estimadores e Intervalos de Confianza al 95 % de los parámetros de la DGVE.

viento tenemos que el parámetro de forma ξ es negativo y correspondería a una distribución con cola acotada, tal es el caso de la distribución Weibull, sin embargo, los intervalos de confianza del parámetro de forma de la altura de ola contienen al cero lo que nos da una evidencia de un posible ajuste con la distribución Gumbel.

Para evaluar que el ajuste hecho explica adecuadamente nuestros datos se realizaron las pruebas de bondad de ajuste de Kolmogorov-Smirnov y Anderson-Darling a un nivel de significancia $\alpha = 0.05$, o nivel de confianza de 95 % y para ambas pruebas, por los resultados de la tabla 4.2 no se rechaza el modelo estimado, adicionalmente se presentan los gráficos de ajuste mediante las gráficas de cuantiles y probabilidad y se incluyen los gráficos de los niveles de retorno como se muestra en las figuras 4.5 y 4.6. Finalmente tenemos las funciones de distribución marginales de los máximos semestrales de la altura de ola G_{hs} y velocidad de viento G_v en las ecuaciones 4.1 y 4.2.

Prueba	Altura de ola (hs)	Velocidad de viento (v)
Kolmogorov-Smirnov	D = 0.0797 p-value = 0.7661	D = 0.1488 p-value=0.09001
Anderson-Darling	A = 0.4634 p-value = 0.7838	A = 2.2858 p-value = 0.06449

Tabla 4.3: Resultados del ajuste.

$$G_{hs}(x) = \exp \left\{ - \left(1 + -0.00084 \left(\frac{x - 4.60387}{0.84983} \right) \right)^{-1/-0.00084} \right\} \quad (4.1)$$

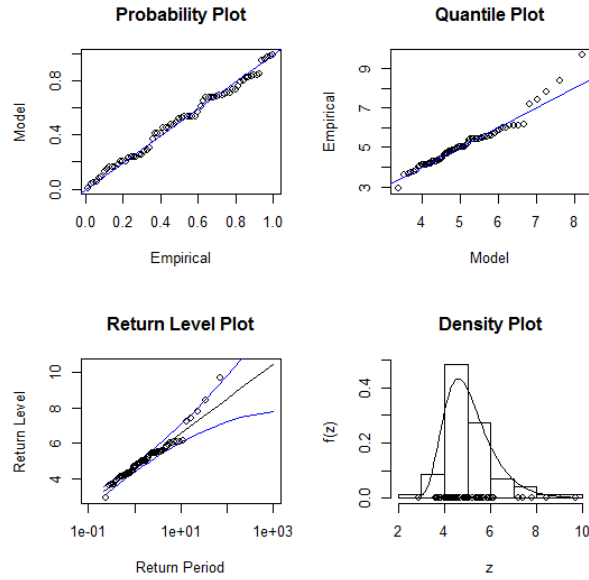


Figura 4.5: Gráficas de diagnóstico del ajuste de Máximas alturas de ola significativa.

$$G_v(x) = \exp \left\{ - \left(1 + -0.2105 \left(\frac{x - 16.455}{3.1449} \right) \right)^{-1/-0.2105} \right\} \quad (4.2)$$

4.2.2. Modelación bivariada de máximos por bloque

Procederemos a desarrollar un modelo bivariado para la altura de ola significativa y la velocidad de viento mediante la aplicación de la TVEB ajustando una cópula de valores extremos a nuestros datos. El paquete estadístico usado para el ajuste de las cópulas [24], trabaja con los modelos de un sólo parámetro, pues estima sólo el principal que captura la información de dependencia.

Antes de ajustar nuestro modelo bivariado exploraremos la estructura de dependencia existente entre los componentes de nuestra submuestra de máximos. Para comprobar la hipótesis nula de independencia de la altura de Ola significativa y la velocidad de viento, se realiza una prueba de independencia;

Global Cramer-von Mises statistic: 0.2133229 with p-value 0.0004995005

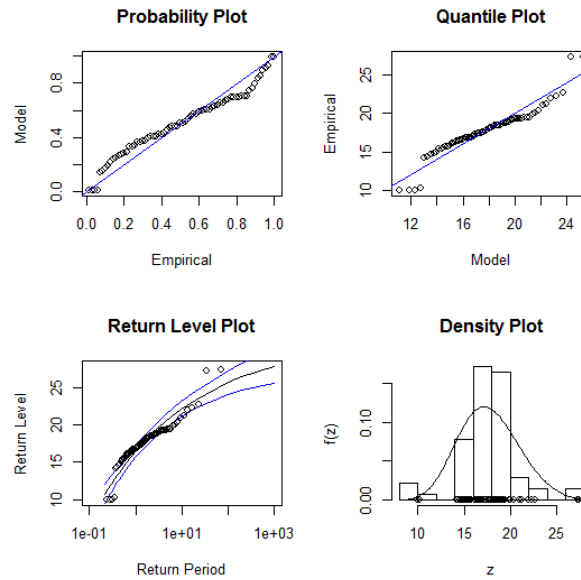


Figura 4.6: Gráficas de diagnóstico del ajuste de Máximos de velocidad de viento.

Combined p-values from the Mobius decomposition:
 0.0004995005 from Fisher's rule,
 0.0004995005 from Tippett's rule

Estos valores (p values) dan una fuerte evidencia en contra de la hipótesis nula de independencia. Por lo tanto, tiene sentido considerar modelar la dependencia entre la altura de Ola significativa y la velocidad de viento.

Del análisis exploratorio esperamos que exista una asociación positiva entre las variables de estudio hecho que se confirma en el gráfico Chi-plot de la figura 4.7. Además como la correlación lineal no es una medida de dependencia indicada en este contexto nos concentramos en la medida tau de Kendall y resulta ser significativamente mayor que cero.

Consideraremos la clase de Cópulas de valores extremos por ser modelos adecuados para la estructura de dependencia entre eventos extremos. Seleccionaremos entonces el modelo que mejor ajuste de entre los siguientes cuatro: cópulas Gumbel o modelo logístico, cópula Galambos o modelo negativo logístico, cópula Tawn o modelo asimétrico logístico y el modelo Husler-Reiss.

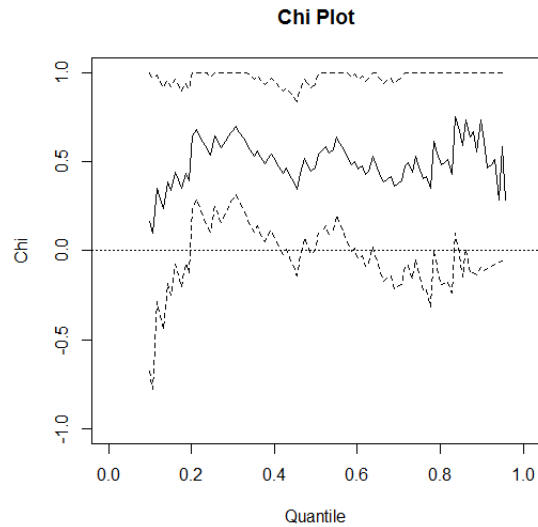


Figura 4.7: Gráfico Chi-plot de la altura de Ola significativa y Velocidad de viento.

A continuación se procede a estimar los parámetros de éstas. Como ya se estimaron los parámetros de las distribuciones marginales, la segunda etapa en la estimación corresponde a estimar el parámetro de dependencia de la cópula. Al unir la estimación univariada con la estimación del parámetro de dependencia se está utilizando el método IFM descrito en el capítulo 3 en la sección de estimación de parámetros. Una vez que el parámetro de dependencia de las cópulas ha sido estimado, queda por evaluar cuál de ellas proporciona el mejor ajuste a nuestros datos. Los métodos utilizados para este fin son pruebas de hipótesis de bondad de ajuste, criterios de información estadística y métodos gráficos.

Cópula	Parámetro	P-Value
Gumbel	1.4085	0.2972
Galambos	0.6308	0.1763
Tawn	0.7521	0.493
Husler-Reiss	0.9289	0.07942

Tabla 4.4: Resultados del ajuste.

Naturalmente, por ser Cópulas de valores extremos y por lo explicado anteriormente, las pruebas de hipótesis hechas indican que todos los modelos serían aceptables para el ajuste, pero para discernir entre áquel que mejor modele los datos consideramos el que haya tenido el menor valor para el AIC. En la tabla 4.8 se muestran los resultados del ajuste, siendo la cópula Gumbel la que muestra el menor valor de AIC así como el valor más grande en la estimación de máxima verosimilitud y después le sigue la cópula Tawn.

Cópula	Log-likelihood	AIC
Gumbel	12.24	-22.48
Galambos	11.73	-21.46
Tawn	11.88	-21.77
Husler-Reiss	11.32	-20.64

Tabla 4.5: Resultados del ajuste.

Del gráfico comparativo con la cópula empírica observamos (de manera intuitiva) que ambas tienen un ajuste muy similar (ver la figura 4.8).

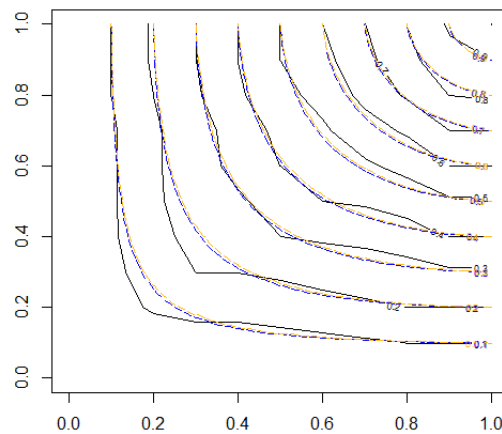


Figura 4.8: Gráfico de líneas de contorno: líneas punteadas azules para el ajuste de la Cópula Gumbel y líneas punteadas amarillas para el ajuste de la Cópula Tawn.

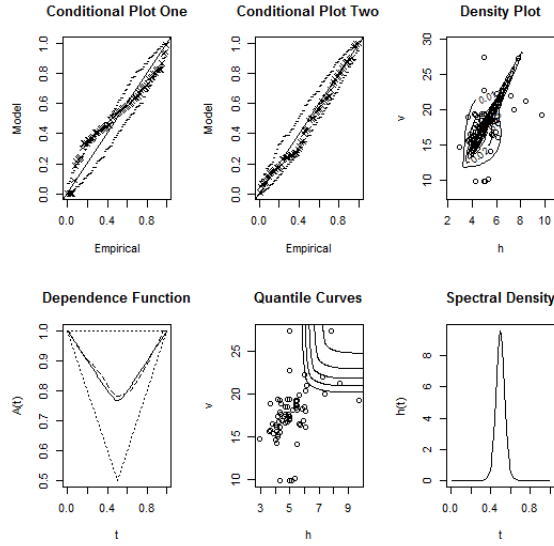


Figura 4.9: Ajuste Bivariado de Máximos por Bloque.

Del ajuste realizado tenemos la expresión de la Cópula Gumbel obtenida y su función de dependencia de Pickands de la siguiente forma

$$C(u_1, u_2) = \exp\{-[(-\log(u_1))^{1.4085} + (-\log(u_2))^{1.4085}]^{1/1.4085}\}, \quad (4.3)$$

$(u_1, u_2)^\top \in [0, 1]^2$. Por el capítulo 2 y (4.4),

$$A(t) = [t^{1.4085} + (1-t)^{1.4085}]^{1/1.4085}. \quad (4.4)$$

4.2.3. Excedencias sobre un umbral

Para el análisis de los datos mediante el método de picos o excedencias sobre un umbral se requiere primero determinar el tamaño adecuado del umbral u a partir del cual se tomarán todas aquéllas observaciones que lo excedan para ajustar, con base en los resultados expuestos en los capítulos anteriores, una distribución generalizada de Pareto.

Se elegirán entonces los umbrales adecuados para las variables, altura de ola y velocidad de viento y se desarrollará para cada una el modelo de picos sobre un umbral, posteriormente se ajustarán cópulas de valores extremos para construir el modelo bivariado.

Ajuste marginal

Para determinar el umbral adecuado de nuestras variables en estudio se ocuparon herramientas gráficas como la gráfica de media de excesos descrita en el capítulo 3, que muestra un umbral adecuado en los valores de u en los que presenta un comportamiento lineal y gráficas de las estimaciones de los parámetros de escala modificada y forma a distintos valores de umbrales en donde el umbral adecuado se indica en los valores para los que las gráficas de las estimaciones presentan un comportamiento constante.

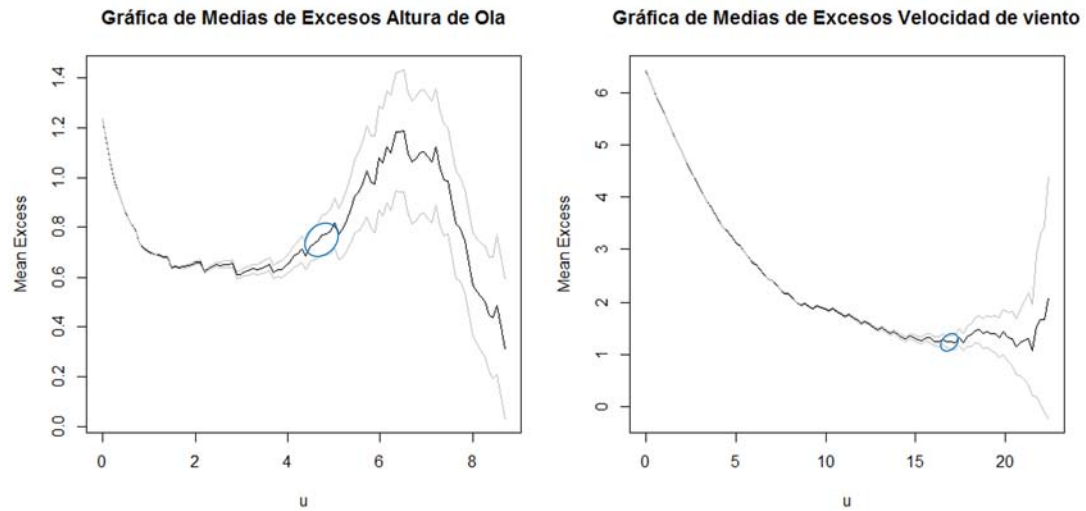


Figura 4.10: Gráfico de Función de medias de excesos de altura de ola significativa y velocidad de viento.

Para el caso de la altura de ola significativa, se puede ver en la gráfica de media de excesos 4.10 que comienza a mostrar un comportamiento más lineal para valores de u mayores de 4 y alrededor de 5, mientras que para la velocidad de viento la gráfica de medias de excesos comienza a mostrar un comportamiento lineal para valores de u alrededor de 16 y 17.

El siguiente paso consiste en analizar las estimaciones de los parámetros de forma ξ y escala modificada β a distintos valores de umbrales, considerándose como umbrales adecuados aquellos valores para los cuales las gráficas de estimación de los parámetros de forma y escala modificada se comporten de forma aproximadamente constante, y si hay estabilidad

en uno o más rangos de valores de umbrales se elige por lo general el menor de ellos o el que presente el mejor ajuste al modelo.

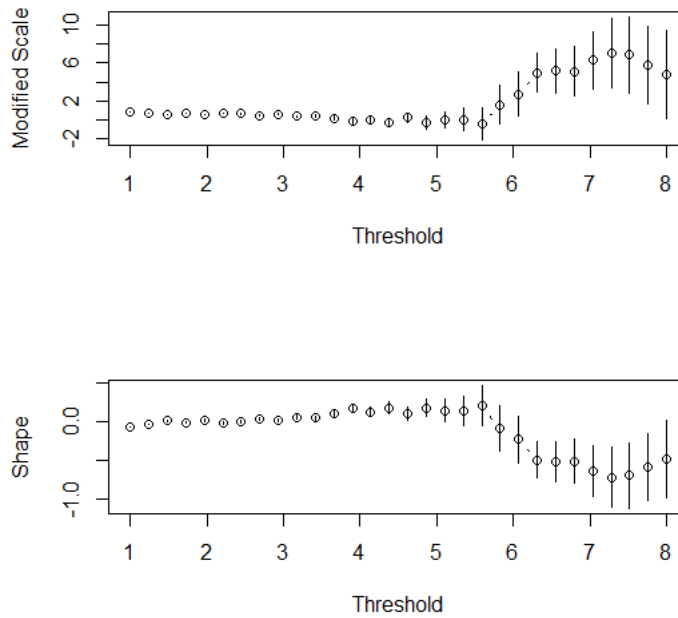


Figura 4.11: Gráfico de estimaciones de los parámetros de la DGP de escala modificada y forma para distintos umbrales de altura de ola.

En la figura (4.11) se muestra la gráfica de los estimadores de ξ y β para la altura de ola y se puede ver que esta presenta un comportamiento aproximadamente constante en valores de u entre 4 y 5. Por su parte, en la figura (4.12) se tiene la gráfica de los estimadores de ξ y β para la velocidad de viento que presenta un comportamiento aproximadamente constante en valores de u entre 16 y 17.

Teniendo los valores de umbrales sugeridos mediante las herramientas anteriores, se realizó el ajuste de los modelos para varios valores dentro de los intervalos considerados y se evaluó cada uno mediante pruebas de bondad de ajuste de Kolmogorov-Smirnov y Anderson Darling para seleccionar finalmente aquél que proporcionará el mejor ajuste de ellos. Con base en lo anterior se tomó el valor de $u = 4.3$ para la altura de ola y de $u = 16.7$ para la

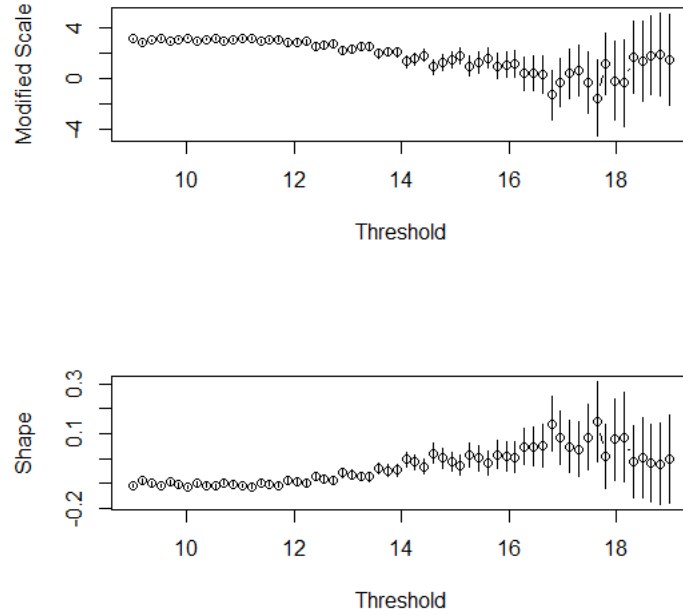


Figura 4.12: Gráfico de estimaciones de los parámetros de la DGP de escala modificada y forma para distintos umbrales de velocidad de viento.

velocidad de viento.

Una vez que se tienen considerados los valores aproximados de los umbrales adecuados debemos cuidar que las observaciones excedentes a dichos umbrales sean lo menos correlacionadas posibles.

En las aplicaciones y como se puede ver en [10], se suele asumir alguna condición que limite el grado de dependencia entre las excedencias eligiendo, por ejemplo, un tiempo específico de distancia de tal modo que los eventos $X_i > u$ y $X_j > u$ sean aproximadamente independientes. Para este trabajo se eligió una condición temporal de 7 días.

Una vez que se procuró tener una muestra de excedencias sobre los umbrales elegidos aproximadamente independientes, se procede a estimar los parámetros de la DGP para la altura de ola y la velocidad de viento para valores de umbrales en los rangos sugeridos por las

		N	$\hat{\theta}$	Estimadores	Intervalos
Altura	$u = 4.3$	111	$\hat{\beta}$	0.73462	(0.5391816 , 0.9300624)
			$\hat{\xi}$	0.07079	(-0.1199077 , 0.2614786)
			$\hat{\xi}$	0.2547	(0.01716, 0.59393)
Velocidad	$u = 16.7$	127	$\hat{\beta}$	1.1722	(0.8793654 , 1.4650991)
			$\hat{\xi}$	0.1409	(-0.04008965 , 0.32196652)

Tabla 4.6: Estimadores e Intervalos de Confianza al 95 % de los parámetros de la DGP.

	Umbral	Estadístico	p-value
Altura	$u = 4.3$	D = 0.0716,	0.6206
		A = 1.2015,	0.2669
Velocidad	$u = 16.7$	D = 0.0813	0.3707
		A = 0.8761	0.4291

Tabla 4.7: 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.

herramientas gráficas usando el método de máxima verosimilitud y eligiendo el menor valor de umbral que ajusta bien al modelo. En la tabla (4.6) se presentan el número de excedencias N , las estimaciones obtenidas y sus correspondientes intervalos de confianza.

Se presentan en la tabla (4.7) las pruebas de bondad de ajuste de Kolmogorov-Smirnov y Anderson-Darling que verifican que se tiene un ajuste adecuado de los excesos sobre el umbral establecido pues ninguna de las pruebas es rechazada. Además se puede apreciar en las gráficas de diagnóstico de ambos ajustes que efectivamente los modelos describen correctamente a los datos.

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 (4.5) y (4.6) respectivamente, y en las ecuaciones (4.7) y (4.8) la expresión de sus respectivos cuantiles en años. Como la base comprende de 1980 a 2014 se consideraron 34 años para la estimación de $n_y = N/34$, donde N es el número de excedencias de cada variable.

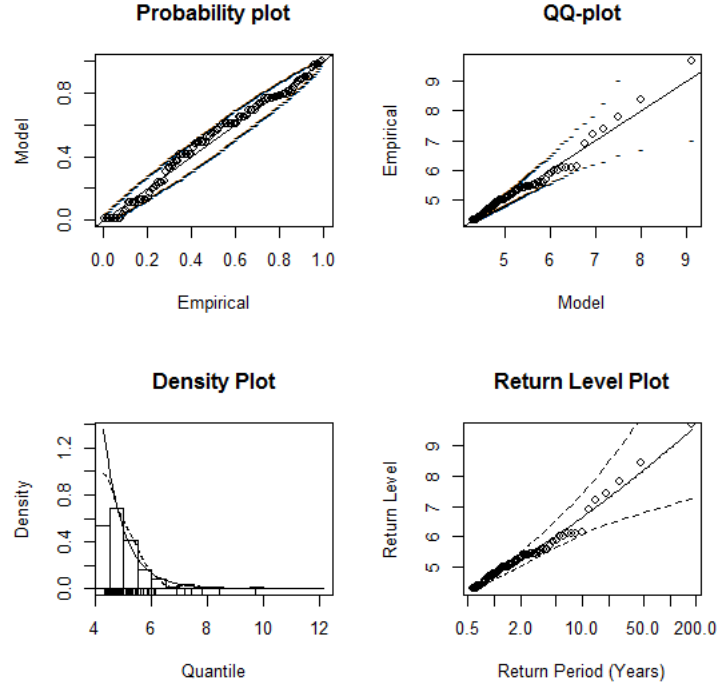


Figura 4.13: Gráficas de diagnóstico del ajuste de alturas de ola por encima de un umbral: $u = 4.3$.

$$H_a(x) = 1 - \left(1 + 0.07079 \left(\frac{x - 4.3}{0.73462} \right) \right)^{-1/0.07079}, \quad x > 4.3 \quad (4.5)$$

$$H_v(x) = 1 - \left(1 + 0.1409 \left(\frac{x - 16.7}{1.1722} \right) \right)^{-1/0.1409}, \quad x > 16.7 \quad (4.6)$$

$$x_a(T) = 4.3 + \frac{0.73462}{0.07079} \left(\left(\frac{111}{34} T \right)^{0.07079} - 1 \right) \quad (4.7)$$

$$x_v(T) = 16.7 + \frac{1.1722}{0.1409} \left(\left(\frac{127}{34} T \right)^{0.1409} - 1 \right) \quad (4.8)$$

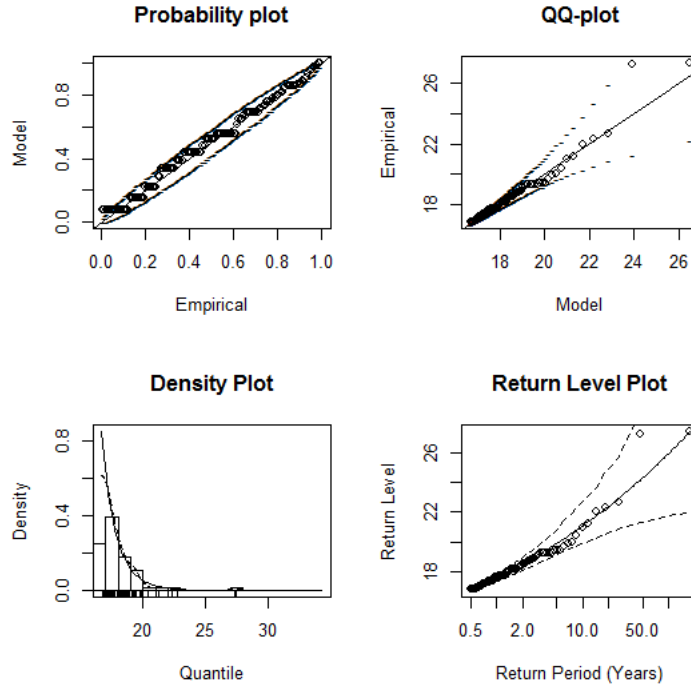


Figura 4.14: Gráficas de diagnóstico del ajuste de velocidades de viento por encima de un umbral: $u = 16.7$.

4.2.4. Modelación bivariada de excedencias sobre un umbral

Procedemos a desarrollar el modelo bivariado de excedencias sobre un umbral para la altura de ola significativa y la velocidad de viento mediante la aplicación de la TVEB y el ajuste de una cópula de valores extremos a los datos. De igual forma que en el método bivariado de máximos por bloques, se utilizó el paquete estadístico Copula [24], el cual trabaja con los modelos de un sólo parámetro.

Consideraremos la clase de Cópulas de valores extremos para seleccionar el modelo que mejor ajuste de entre los siguientes cuatro: cópulas Gumbel o modelo logístico, cópula Galambos o modelo negativo logístico, cópula Tawn o modelo asimétrico logístico y el modelo Husler-Reiss.

A continuación se procede a estimar los parámetros. En forma análoga al método de máximos por bloque, como ya se estimaron los parámetros de las distribuciones marginales,

Cópula	Parámetro	P-Value
Gumbel	1.17126	0.04745
Galambos	0.39099	0.03347
Tawn	0.40887	0.1024
Husler-Reiss	0.70141	0.02248

Tabla 4.8: Resultados del ajuste.

la segunda etapa en la estimación corresponde a estimar el parámetro de dependencia de la cópula. Al unir la estimación univariada con la estimación del parámetro de dependencia se está utilizando el método IFM. Una vez que el parámetro de dependencia de las cópulas ha sido estimado, queda por evaluar cuál de ellas proporciona el mejor ajuste a nuestros datos. Los métodos utilizados para este fin son pruebas de hipótesis de bondad de ajuste, criterios de información estadística y métodos gráficos.

Las pruebas de hipótesis hechas indican que el modelo aceptable para el ajuste de los datos es el de la Cópula Tawn pues los demás modelos fueron rechazados por las pruebas de bondad de ajuste. En la Tabla 4.8 se muestran los resultados del ajuste y por último, del gráfico comparativo con la cópula empírica observamos que el modelo presenta un ajuste adecuado (ver la Figura 4.15).

Del ajuste realizado, tenemos que la expresión de la Cópula Tawn obtenida queda de la siguiente forma

$$C(u_1, u_2) = u_1 u_2 \exp\{-[(-\log(u_1))^{0.40887} + (-\log(u_2))^{0.40887}]^{1/0.40887}\}. \quad (4.9)$$

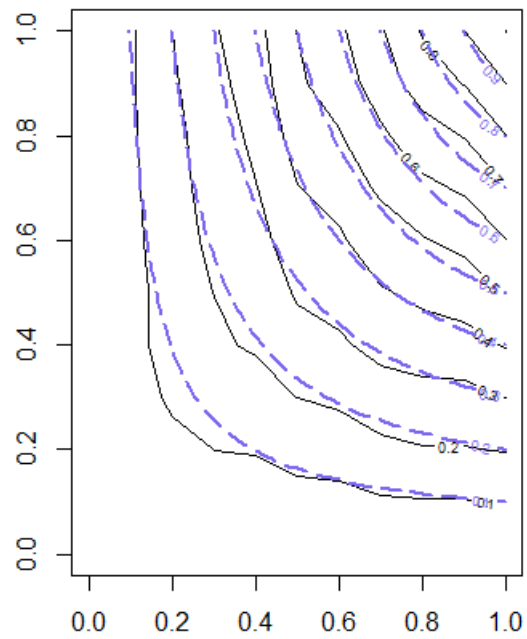


Figura 4.15: Gráfico de líneas de contorno: Ajuste de la Cópula Tawn estimada por el método de excedencias sobre un umbral.

Conclusiones

El objetivo de esta tesis fue estudiar eventos extremos bivariados del oleaje en una zona cercana a la costa de México, se observaron la altura de ola significativa y la velocidad del viento, que son dos de las variables principales en la caracterización de este fenómeno hidrometeorológico.

Se aplicaron los resultados de la teoría clásica de valores extremos para modelar univariadamente las distribuciones marginales de cada variable en estudio mediante 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 que nos dice que la Distribución Generalizada de Pareto surge como distribución límite para los excesos sobre un umbral. También se aplicó teoría de valores extremos multivariada junto con un enfoque de cópulas usando uno de los cuatro modelos de la familia de cópulas de valores extremos para capturar la dependencia entre las marginales.

Una importante motivación de este trabajo fue poder realizar comparaciones con ajustes realizados previamente en el Instituto de Ingeniería de la UNAM y poder calcular distribuciones de probabilidad para eventos extremos bivariados del oleaje y de los que no se cuentan con estudios previos de comparación. Cabe destacar que al realizar un modelo bivariado se busca tener modelos que estudien el efecto que pueden ocasionar los diversos factores y variables que determinan al oleaje y cuyos eventos extremos combinados podrían resultar importantes en la cuantificación riesgo hidrometeorológico y cuya manifestación causaría pérdidas humanas, económicas y daños materiales considerables.

El análisis univariado es suficientemente útil cuando solo los eventos extremos de una variable aleatoria son importantes en los criterios de diseño de una estructura o cuando la

correlación entre otras variables no es significativa, es decir que sean menos dependientes entre sí. Sin embargo, un análisis separado de variables aleatorias X y Y no puede revelar la relación de dependencia entre ellas cuando la correlación es una información importante en los criterios de diseño.

El periodo de retorno de eventos hidrometeorológicos extremos es utilizado como un criterio común empleado en el diseño de estructuras. Tradicionalmente son usadas distribuciones univariadas de valores extremos para describir los fenómenos hidrometeorológicos extremos, sin embargo, estos fenómenos son de cierta complejidad y a menudo están caracterizados por diversos aspectos y factores, por ejemplo, las inundaciones se describen como eventos multivariados caracterizadas por la duración, volumen y pico. Es por esto que el mejor enfoque para el análisis de este tipo de eventos complejos es a través de distribuciones conjuntas de dos o más variables aleatorias teniendo en cuenta la correlación entre las mismas.

Después de seleccionar una distribución bivariada para modelar un evento extremo hidrometeorológico podrá buscarse la forma de los períodos de retorno y las relaciones entre sus funciones de distribución univariada y bivariada aunque la forma bivariada resulta más compleja que la de los periodos de retorno univariados, pueden definirse los períodos de retorno conjuntos de X y Y o definirse los periodos de retorno condicionales para X dado $Y \leq y$ y viceversa.

Es importante resaltar que entre los primeros pasos realizados en nuestro análisis fue una prueba estadística que nos permitiera concluir si nuestras observaciones extremas conjuntas presentaban una relación de dependencia. Los resultados de la misma nos permitieron rechazar la hipótesis nula que asumía un comportamiento independiente y así dar sentido al estudio bivariado de valores extremos de este trabajo. En caso contrario, si el resultado de la prueba hubiese sido aceptar la independencia de nuestras parejas de datos, entonces el análisis univariado de valores extremos hubiese sido suficiente para estudiar el comportamiento conjunto de la altura de ola y la velocidad de viento en el fenómeno del oleaje.

Existe un amplio campo de estudio en la teoría de valores extremos, particularmente los extremos multivariados están son de reciente desarrollo y motivadores de diversas aplicaciones y trabajos de estudio. Otras de las aplicaciones, por ejemplo, han sido estudiar la

dependencia entre extremos de diferentes espacios y localizaciones geográficas. Por lo que el trabajo realizado en esta tesis podría utilizarse para modelar la dependencia espacial de alguna(s) variables importantes en la caracterización del oleaje u otro fenómeno hidrometeorológico.

Se espera también que este trabajo pueda ser un preámbulo para el aprovechamiento de información al cuantificar al Oleaje como fenómeno hidrometeorológico y su riesgo asociado, así como una posible alternativa para la inferencia de datos faltantes entre las variables correlacionadas con el objetivo de realizar predicciones numéricas deterministas.

Apéndice A

Proceso de Ajuste

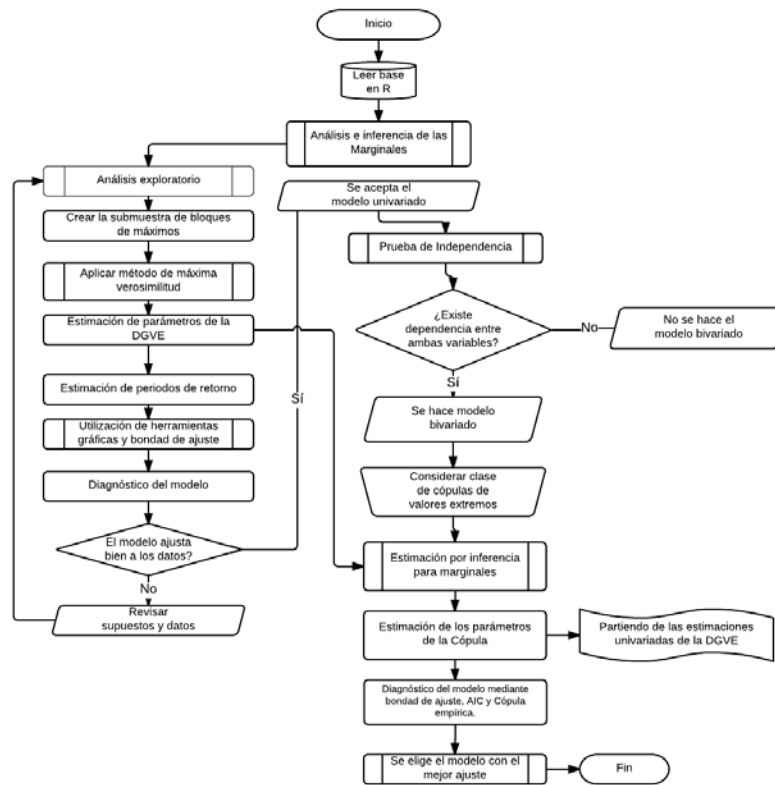


Figura A.1: Diagrama de flujo del procedimiento para estimar un modelo bivariado de máximos por bloque.

Apéndice B

Proceso de Ajuste

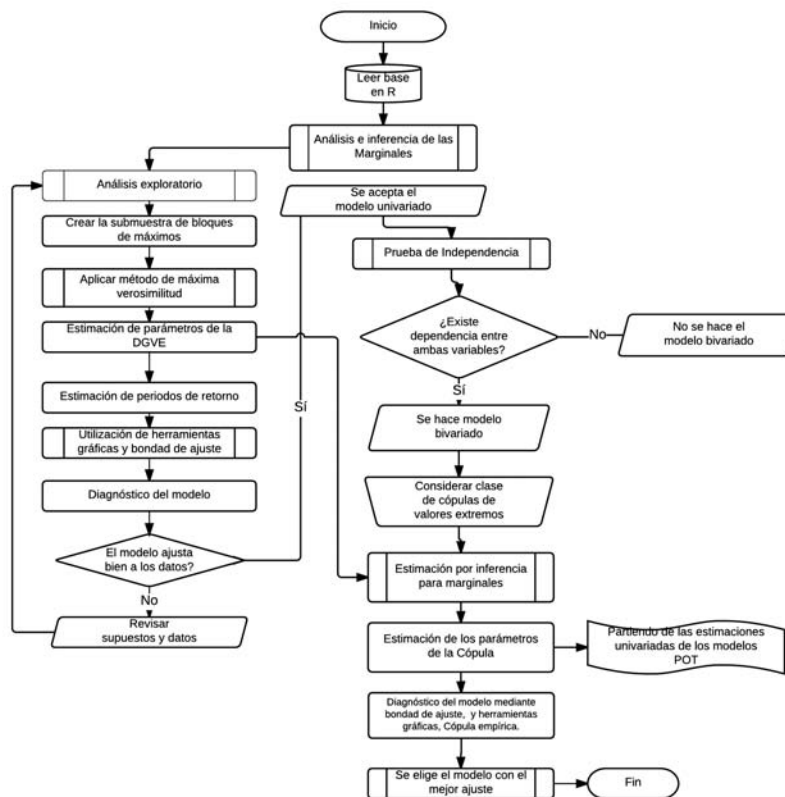


Figura B.1: Diagrama de flujo del procedimiento para estimar un modelo bivariado de excepciones sobre un umbral.

Apéndice C

Resultados varios de Probabilidad

Definición. Función Inversa Generalizada Sea H una función no decreciente en \mathbb{R} , se define la inversa generalizada de H como:

$$H^{\leftarrow}(y) = \inf\{s : H(s) \geq y\}$$

Propiedades:

1. $H^{\leftarrow}(y) = \inf\{s : H(s) \geq y\}$ es continua por la izquierda en \mathbb{R} supongamos que $x_n \uparrow x$ pero que $H^{\leftarrow}(x_n) \uparrow H^{\leftarrow}(x-) < H(x)$ entonces $\exists \delta > 0$, y tal que $\forall n$

$$H^{\leftarrow}(x_n) < y < H^{\leftarrow}(x) - \delta$$

La desigualdad de la izquierda y la definición de H^{\leftarrow} producen que $H(y) \geq x_n \forall n$, de aquí haciendo que $n \rightarrow \infty$ tenemos $H(y) \geq x$ de donde otra vez por definición de H^{\leftarrow} tenemos que $y \geq H^{\leftarrow}(x)$ que junto con $y < H^{\leftarrow}(x) - \delta$ nos lleva a la contradicción deseada.

En caso de que la función H sea continua por la derecha, se tiene

Proposición C.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) $F(F^{\leftarrow}(y)) \geq y$.

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

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

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

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 $\lim_{n \rightarrow \infty} F^{\leftarrow}(y_n) < F^{\leftarrow}(y)$. Entonces existe $\delta > 0$ y t tales que para toda n ,

$$F^{\leftarrow}(y_n) < t < F^{\leftarrow}(y) - \delta. \quad (\text{C.1})$$

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

c) Para demostrar que $A(y)$ es cerrado, tomamos una sucesión $s_n \in A(y)$ para toda n , tal que $s_n \rightarrow s_0$. 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

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

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

$$s_0 - \delta_0 < s_n < s_0 + \delta_0, \quad \forall n \geq n_0. \quad (\text{C.3})$$

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 (C.3), $s_n \in (s_0, s_0 + \delta_0)$, que por (C.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 $F^{\leftarrow}(y) = \inf A(y) \in A(y)$, 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

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

□

Proposición C.2. Convergencia de Funciones inversas. Si $\{F_n, n \geq 0\}$ es una sucesión de funciones no decrecientes y $F_n(x) \rightarrow F_0(x)$ para todo $x \in \mathcal{C}(F_0)$, entonces $F_n^{\leftarrow}(y) \rightarrow F_0^{\leftarrow}(y)$ 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 $F_n(x_1) \rightarrow F_0(x_1)$, 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,

$$F_0^{\leftarrow}(y) \leq \liminf_{n \rightarrow \infty} F_n^{\leftarrow}(y). \quad (\text{C.4})$$

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 $F_n(x_2) \rightarrow F_0(x_2)$, 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

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

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

$$\limsup_{n \rightarrow \infty} F_n^{\leftarrow}(y) \leq F_0^{\leftarrow}(y). \quad (\text{C.5})$$

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

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

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

Teorema C.1. Convergencia a tipos. Sean X_1, X_2, \dots variables aleatorias con función de distribución F_n para cada $n \geq 1$, y sean U, V variables aleatorias con función de distribución F_U y F_V respectivamente (no degeneradas) y $\mathcal{C}(F_U), \mathcal{C}(F_V)$ los conjuntos de puntos de continuidad de F_U y F_V respectivamente. Sean $a_n > 0, \alpha_n > 0, b_n \in \mathbb{R}, \beta_n \in \mathbb{R}$ constantes.

a) Si

$$F_n(a_n x + b_n) \rightarrow F_U(x) \quad y \quad F_n(\alpha_n x + \beta_n) \rightarrow F_V(x) \quad (\text{C.6})$$

para $x \in \mathcal{C}(F_U)$, $x \in \mathcal{C}(F_V)$ respectivamente. Equivalentemente

$$\frac{X_n - b_n}{a_n} \xrightarrow{d} U \quad y \quad \frac{X_n - \beta_n}{\alpha_n} \xrightarrow{d} V \quad (\text{C.7})$$

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

$$\lim_{n \rightarrow \infty} \frac{\alpha_n}{a_n} = A > 0 \quad y \quad \lim_{n \rightarrow \infty} \frac{\beta_n - b_n}{\alpha_n} = B \in \mathbb{R} \quad (\text{C.8})$$

y

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

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

Demostración. b) Suponemos (C.8) y que $G_n := F_n(a_n x + b_n) \rightarrow F_U(x)$ para todo $x \in \mathcal{C}(F_U)$.

Entonces

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

Sea $x \in \mathcal{C}(F_U(A \cdot + B))$, supongamos que $x > 0$ y sea $\epsilon > 0$, por (C.8) $\exists n_0$ tal que $\forall n \geq n_0$

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

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

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

Entonces

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

Por lo que, $\forall z \in \mathcal{C}(F_U)$ tal que $z > (A + \epsilon)x + (B + \epsilon)$ se tiene

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

por (C.6), entonces

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

como $\epsilon > 0$ es arbitrario y F_U es función de distribución y por tanto continua por la derecha tenemos

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

Por otro lado, tenemos que

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

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

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

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

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

Entonces tenemos que

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

Por lo tanto,

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

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

a) Supongamos que $F_n(a_n x + b_n) \rightarrow F_U(x)$ y $F_n(\alpha_n x + \beta_n) \rightarrow F_V(x)$. Entonces, por la Proposición C.2 tenemos que

$$\frac{F_n^{\leftarrow}(y) - b_n}{a_n} \rightarrow F_U^{\leftarrow}(y) \quad , \quad \forall y \in \mathcal{C}(F_U^{\leftarrow}) \tag{C.10}$$

$$\frac{F_n^{\leftarrow}(y) - \beta_n}{\alpha_n} \rightarrow F_V^{\leftarrow}(y) \quad , \quad \forall y \in \mathcal{C}(F_V^{\leftarrow})$$

Como F_U y F_V son no degeneradas, y monótonas, podemos tomar dos puntos y_1, y_2 tales que $y_1, y_2 \in \mathcal{C}(\mathcal{F}_U^{\leftarrow}) \cap \mathcal{C}(\mathcal{F}_V^{\leftarrow})$, $y_1 < y_2$

$$F_U^{\leftarrow}(y_1) < F_U^{\leftarrow}(y_2) \quad y \quad F_V^{\leftarrow}(y_1) < F_V^{\leftarrow}(y_2)$$

Por lo tanto, por (C.10) para $i = 1, 2$ tenemos que

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

restando las expresiones para $i = 1$ de las de $i = 2$ tenemos

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

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

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

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

Además, de (C.11) tenemos

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

Y restando estas dos últimas expresiones ,

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

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

Definición. Convergencia Débil o en Distribución. *Una sucesión de variables aleatorias X_1, X_2, \dots , con funciones de distribución F_1, F_2, \dots respectivamente, se dice que converge en distribución a la variable aleatoria X , con función de distribución F si*

$$F_n(x) \rightarrow F(x)$$

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

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

Definición. Convergencia en Probabilidad. *Se dice que una sucesión de variables aleatorias (X_n) converge en Probabilidad a una constante $c \in \mathbb{R}$ si*

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - c| < \epsilon) = 1 \quad \forall \epsilon > 0$$

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

Teorema C.2. Transformación integral de probabilidad *Sea X una variable aleatoria con función de distribución continua F . Entonces la variable aleatoria $U = F(X)$ se distribuye uniformemente en $(0, 1)$.*

Demostración. Como F es una función de distribución por lo tanto es no decreciente. Sea $F^{\leftarrow} : (0, 1) \rightarrow (-\infty, \infty)$, $F^{\leftarrow}(u) = \inf\{x : H(x) \geq u\}$ la generalizada inversa de F . Entonces

$$\mathbb{P}(U \leq u) = \mathbb{P}(F(X) \leq u) = \mathbb{P}(X \leq F^{\leftarrow}(u)) = F(F^{\leftarrow}(u)) = u \quad 0 \leq u \leq 1$$

que es la función de distribución de una variable aleatoria que se distribuye uniforme en $(0, 1)$, por tanto $U \sim U(0, 1)$. \square

Definición. Ecuación de Cauchy. Una función $f : \mathbb{R} \rightarrow \mathbb{R}$ satisface la ecuación de Cauchy (o es aditiva) si

$$f(x + y) = f(x) + f(y), \quad x, y \in \mathbb{R}. \quad (\text{C.14})$$

Una función $g : (0, \infty) \rightarrow (0, \infty)$ satisface la ecuación de Hamel (o es multiplicativa) si

$$g(xy) = g(x)g(y) \quad x, y > 0. \quad (\text{C.15})$$

Funciones de Variación Regular. La teoría de Funciones de Variación Regular fue desarrollada por J. Karamata en 1930 y tiene un papel importante en el desarrollo de la Teoría de Valores Extremos, como por ejemplo, en la caracterización de los dominios de atracción de las distribuciones límite para máximos.

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

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

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 C.3. (Karamata).

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

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

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

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

(b) Si F satisface

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

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

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

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

Corolario C.1. (Representación de Karamata). \mathcal{L} es de variación lenta si y sólo si puede representarse como

$$\mathcal{L}(x) = c(x) \exp \left\{ \int_1^x \frac{\epsilon(t)}{t} dt \right\} \quad (\text{C.16})$$

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

$$\lim_{x \rightarrow \infty} c(x) = c_0 \in (0, \infty) \quad (\text{C.17})$$

$$\lim_{x \rightarrow \infty} \epsilon(x) = 0. \quad (\text{C.18})$$

La demostración tanto del Teorema C.3 como la del corolario C.1 pueden encontrarse en [34].

Cóputas. La teoría de Funciones de Variación Regular fue desarrollada por J. Karamata en 1930 y tiene un papel

Apéndice D

Herramientas Estadísticas

Media muestral

La media aritmética o media muestral \bar{x} de una muestra (x_1, \dots, X_n) es normalmente considerada como un estimador de la media de la población de la que proviene la muestra. Entonces para x_1, \dots, x_n un conjunto de n observaciones de una variable aleatoria X , la media aritmética o muestral \bar{x} se define como:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Desviación estándar

La desviación estándar es una medida de dispersión de los datos. En la práctica, la desviación estándar σ de una población se estima por la desviación estándar S de una muestra de esta población.

Para una muestra de n observaciones x_1, \dots, x_n , la desviación estándar S se define como:

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

donde \bar{x} es la media muestral.

Coefficiente de asimetría y curtosis

Una manera de identificar la forma de la distribución que siguen los datos es mediante la ayuda de coeficientes apropiados que permiten investigar si la forma de la distribución se asemeja a una distribución normal o si difiere de esta. Estas medidas de forma son el

coeficiente de asimetría o también conocido como sesgo y la curtosis.

El coeficiente de **asimetría** permite estudiar el grado de asimetría que presenta una distribución de probabilidad de una variable aleatoria. Como la distribución normal es simétrica, el coeficiente de asimetría es cero. Valores negativos del coeficiente de asimetría indican que los datos son valores sesgados a la izquierda y valores positivos indican que los datos son sesgados a la derecha. Por sesgada a la izquierda, nos referimos a que la cola izquierda es más larga en relación con la cola derecha. Del mismo modo, sesgada a la derecha significa que la cola derecha es más larga en relación con la cola izquierda.

El coeficiente de asimetría se define como:

$$\gamma = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{S} \right)^3$$

donde μ_3 es el tercer momento respecto a la media. Según [2], se hace el siguiente ajuste para tener un estimador insesgado del coeficiente de asimetría

$$\gamma = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{S} \right)^3$$

El coeficiente de **curtosis** es una medida que representa la forma de la cúspide de la curva de una función de distribución. Cuando una función de distribución tiene una forma más puntiaguda se le llama leptocúrtica y tienden a tener altos valores de curtosis, especialmente si las colas de la distribución son más grandes que las de la distribución normal, mientras que las distribuciones con forma más aplanada o más redondeada se le conoce como platicúrtica, estas tienen colas más delgadas y presentan valores de curtosis menores. La distribución normal presenta un valor de curtosis de 3.

El coeficiente de curtosis se define como:

$$\kappa = \frac{\hat{\mu}_4}{\hat{\sigma}^4} - 3 = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{S} \right)^4$$

donde μ_4 es el cuarto momento respecto a la media. Según [2], se hace el siguiente ajuste para tener un estimador insesgado del coeficiente de curtosis

$$\kappa = a \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{S} \right)^4 - b$$

donde

$$a = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \quad y \quad b = \frac{3(n-1)^2}{(n-2)(n-3)}$$

Histograma

El histograma es una representación gráfica de la distribución de frecuencias o función de densidad de los datos que han sido agrupados en clases y representada mediante barras. Revisar la forma de un histograma puede permitirnos detectar el modelo de probabilidad seguido por los datos.

Box Plot

El diagrama de caja es una forma de representar características de los datos mediante medidas estadísticas alienadas sobre una caja vertical como la mediana, el primer cuartil y el tercer cuartil, outliers o calores atípicos y los valores máximos y mínimos. A menudo se utiliza para comparar varios conjuntos de observaciones.

Además de las técnicas estadísticas ya mencionadas para conocer las características de los datos, se tiene que una forma de obtener una primera aproximación a la distribución real de una muestra de datos es mediante la función de distribución empírica. Una forma de definir a la función de distribución empírica es la siguiente: sean x_1, \dots, x_n realizaciones independientes provenientes de una población con función de distribución desconocida F y sea $x_{(1)}, \dots, x_{(n)}$ que denote a la muestra ordenada, $x_{(1)} \leq x_{(2)} \leq \dots x_{(n)}$. Para cada observación $x_{(i)}$, exactamente i de las n observaciones tienen un valor menor o igual a $x_{(i)}$, entonces un estimador empírico de la probabilidad de que una observación sea menor o igual a $x_{(i)}$ es $\tilde{F}(x_{(i)}) = i/n$. De acuerdo a [10], se suele hacer el ajuste $\tilde{F}(x_{(i)}) = i/(n+1)$, siguiendo esto se tiene la definición siguiente que se puede consultar en [10].

Función empírica

Dada una muestra de tamaño n de observaciones ordenadas $x_{(1)} \leq x_{(2)} \leq \dots x_{(n)}$ de una población con función de distribución F , la función de distribución empírica se define como:

$$\tilde{F} = \frac{i}{n+1} \tag{D.1}$$

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

Para juzgar la exactitud de un modelo de acuerdo a su ajuste con los datos en estudio existen algunas herramientas gráficas que permiten observar que tan adecuadamente describe el modelo estimado a la distribución real que siguen los datos. Varios procedimientos de bondad de ajuste se basan en comparaciones del modelo propuesto con respecto a la función empírica de la distribución F .

Gráfica de probabilidades, P-P plot

Dada una muestra ordenada de n observaciones independientes $x_{(1)} \leq x_{(2)} \leq \dots x_{(n)}$ de una población con función de distribución estimada \hat{F} , una gráfica de probabilidades o P-P plot (Probability Plot) consiste en los puntos:

$$\left\{ \left(\hat{F}(x_{(i)}), \frac{i}{n+1} \right) : i = 1, \dots, n \right\}$$

Si \tilde{F} es un modelo adecuado, los puntos graficados deben de encontrarse alineados con respecto a una recta diagonal, y si se alejan demasiado de ella nos indicaría que no es un modelo adecuado.

Gráfica de cuantiles, Q-Q plot

Dada una muestra ordenada de n observaciones independientes $x_{(1)} \leq x_{(2)} \leq \dots x_{(n)}$ de una población con función de distribución estimada \hat{F} , una gráfica de cuantiles o Q-Q plot (Quantile Plot) consiste en los puntos:

$$\left\{ \left(\hat{F}^{\leftarrow} \left(\frac{i}{n+1} \right), x_{(i)} \right) : i = 1, \dots, n \right\}$$

Aquí, $x_{(i)}$ provee un estimador del cuantil $i/(n+1)$ de la distribución F mediante \hat{F}^{\leftarrow} la función inversa generalizada de la función de distribución estimada. Si \hat{F} es una estimación razonable de F entonces los puntos de la gráfica de cuantiles también deben de alinearse a una línea recta diagonal.

Apéndice E

Código R

```
library(evd)
library(evir)
library(fExtremes)
library(ismev)
library(POT)
library(copula)
library(Kendall)
library("fCopulae")
library(stats)
library(ADGofTest)

#Leemos la base de datos y la guardamos una variable 'Consol' para manipularla en R
Consol<-read.table("C:/Users/Maye/Dropbox/Aplicacion/ConsolNA.txt", header = FALSE,
                  sep = "\t", dec = ".", col.names=c("llave","YEAR", "MONTH", "DAY", "HOUR",
                  "V", "SH", "AT", "DP", "AP"))

Data <- data.frame(Consol$SH, Consol$V)
#### Renombrando variables
names(Data)[c(1)] <- c("H")
names(Data)[c(2)] <- c("V")
```

```
##### Analisis exploratorio #####
#### Gráficas de las series
par(mfrow=c(2,1))
plot(Consol$V,type="l",main="Velocidad del viento en m/s",xlab="Año",ylab="Velocidad")
plot(Consol$SH,type="l",main="Altura de ola significativa en m", xlab="Año",ylab="Altura")

#### Histogramas
Utilizaremos la regla de Scott para determinar le tamaño de los intervalos

n <- length(Data$H)
p <- 1/3

par(mfrow=c(1,3))
hist(Data$H, (2*n)^p, main="Histograma de altura de Ola significativa", xlab="H" )
hist(Data$V, (2*n)^p, main="Histograma de velocidad de Viento", xlab="V" )
boxplot(Data)

#### Diagrama de cajas y brazos
boxplot(Data)

#### Summary
summary(Data)
var(Data$H, na.rm = TRUE)
var(Data$H,Data$Y, na.rm = TRUE, "pairwise.complete.obs")
var(Data$V,Data$Y, na.rm = TRUE, "pairwise.complete.obs")
sd(Data$H,na.rm = TRUE)
sd(Data$V,na.rm = TRUE)
kurtosis(Data$H, na.rm=TRUE, method="moment")
kurtosis(Data$V, na.rm=TRUE, method="moment")
skewness(Data$H, na.rm="TRUE", method="moment")
skewness(Data$V, na.rm="TRUE", method="moment")
qqnorm(Data$H)
```

```

qqline(Data$H, col = 2)
qqnorm(Data$V)
qqline(Data$V, col = 2)

#### Bloques ####
#Separación Semestral
datos_1<-subset(Consol,(MONTH==1) | MONTH==2 | MONTH==3 | (MONTH==4) | MONTH==5 | MONTH==6)
atos_2<-subset(Consol,(MONTH==12) | MONTH==7 | MONTH==8 | (MONTH==9) | MONTH==10 | MONTH==11)

#Máximos anuales por estaciones
h_1<-c()
v_1<-c()
p_1<-c()
h_2<-c()
v_2<-c()

for(i in 1980:2014)
{
h_1[i-1979]<-max(subset(datos_1,YEAR==i)$SH, na.rm=TRUE)
v_1[i-1979]<-max(subset(datos_1,YEAR==i)$V,na.rm=TRUE)
h_2[i-1979]<-max(subset(datos_2,YEAR==i)$SH,na.rm=TRUE)
v_2[i-1979]<-max(subset(datos_2,YEAR==i)$V,na.rm=TRUE)

}

h <- c(h_1,h_2)
v <- c(v_1,v_2)
Biv <- data.frame(h,v)
par(mfrow=c(2,2))
acf(h, main="acf(H)")
pacf(h, main="pacf(H)")
acf(v, main="acf(V)")
pacf(v, main="pacf(v)")
fgev(Biv$h)

```

```

confint(fgev(Biv$h))
plot(fgev(Biv$h))
fgev(Biv$v)
confint(fgev(Biv$v))
plot(fgev(Biv$v))
gev.diag(gev.fit(Biv$h))
gev.diag(gev.fit(Biv$v))
ad.test(Biv$h,pgev,xi=-0.0008747279,mu=4.6038639557,beta=0.8497702006)
ks.test(Biv$h,'pgev',xi=0,mu=4.6038639557,beta=0.8497702006)
ks.test(Biv$v,'pgev',xi=0,mu=16.4553733,beta=3.1445349)
ad.test(Biv$v,pgev,xi=-0.2104531,mu=16.4553733,beta=3.1445349)
gevh <- pgev(Biv$h,-0.0008393 ,4.6038715,0.8498257 )
gevv <- pgev(Biv$v,-0.2105 ,16.4550 , 3.1449)
gh <- gevh[1:length(Biv$h)]
gv <- gevv[1:length(Biv$v)]
newBiv <- data.frame(gh,gv)

#### Prueba de hipótesis de independencia
La prueba de independencia se da en dos pasos:

Paso 1: Simulaciones para el estadístico de prueba
paso1 <- indepTestSim(length(Biv$h), 2, N = 1000, verbose = TRUE, print.every = NULL)

n:tamaño de la muestra
p: dimensión del vector aleatorio
N: número de simulaciones del estadístico de prueba
print.every: va reportando el número de simulaciones (100 por default)

Paso 2:
Prueba de Hipótesis Ho:  $H(x,y) = F(x)G(y)$ 
paso2_1 <- indepTest(Biv, paso1 , alpha=0.05)
par(mfrow=c(2,1))
chiplot(Biv, which=1)

```

```

abline(h=0, lty=3)

x: matriz de datos
d: resultado del paso 1
alpha: nivel de la prueba (0.05) por default

### Conclusión: Se rechaza la hipotesis nula de independencia
Kendall(Biv$h,Biv$v)

#Método inference functions from margins

gofEVCopula(gumbelCopula(),newBiv)
gofEVCopula(galambosCopula(),newBiv)
gofEVCopula(tawnCopula(),newBiv)
gofEVCopula(huslerReissCopula(),newBiv)
fitCopula(gumbelCopula(),newBiv,method="ml")
fitCopula(galambosCopula(),newBiv,method="ml")
fitCopula(tawnCopula(),newBiv,method="ml")
fitCopula(huslerReissCopula(),newBiv,method="ml")
AIC(fitCopula(gumbelCopula(),pobs(Biv)))
AIC(fitCopula(galambosCopula(),pobs(Biv)))
AIC(fitCopula(huslerReissCopula(),pobs(Biv)))
AIC(fitCopula(tawnCopula(),pobs(Biv)))

##### Copula Empirica #####
n <- 70
d <- 2
family <- "Gumbel"
theta <- 1.615603
cop <- gumbelCopula(1.615603)
u <- matrix(runif(n*d), n, d)
U <- pobs(Biv)
ec <- C.n(u, U=U)

```

```

## compare with true distribution function
mean(abs(pCopula(u, copula=cop)-ec)) # increase n to decrease this error
## compare the empirical copula and the true copula
## on the diagonal of the unit square
Cn. <- function(x) C.n(do.call(cbind, rep(list(x), d)), U=U)
curve(Cn., 0, 1, main=paste("Diagonal of a", family, "copula"),
      xlab="u", ylab=expression(italic(C)[n](italic(u),...,italic(u))))
pC <- function(x) pCopula(do.call(cbind, rep(list(x), d)), cop)
curve(pC, lty=2, add=TRUE)
legend("topleft", lty=1:2, bty="n", inset=0.02,
      legend=c(expression(italic(C)[n]), expression(italic(C))))

### ANÁLISIS DE EXCEDENCIAS SOBRE UN UMBRAL###

#Gráficas de medias de excesos "Mean Residual Life Plot" para elección del umbral
mrlplot(Base$SH,main="Gráfica de Medias de Excesos Altura de Ola",xlab="u")
mrlplot(Base$V,main="Gráfica de Medias de Excesos Velocidad de Viento",xlab="u")

ola_ts<-ts(Base$SH,start=1980,frequency=(365*24))
viento_ts<-ts(Base$V,start=1980,frequency=(365*24))

#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=Base$SH,time=1:length(Base$SH))
viento<-data.frame(obs=Base$V,time=1:length(Base$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=7*24,clust.max=TRUE)
altura<-clust(ola,u=4.3,tim.cond=5*24,clust.max=TRUE)
velocidad<-clust(viento,u=16.7,tim.cond=5*24,clust.max=TRUE)

```



```
#Revisión de la elección del umbral por estabilidad de las estimaciones
#Estimaciones de los parámetros para altura de ola a diferentes umbrales
par(mfrow=c(2,1))
tcplot(Base$SH, u.range=c(1,8),cmax = TRUE, r = 5*24,ask=FALSE,nt=30)
#Estimaciones de los parámetros para velocidad de viento a diferentes umbrales
par(mfrow=c(2,1))
tcplot(Base$V, u.range=c(12,19),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_altura<-fitgpd(altura[,"obs"],4.3)
pot_velocidad<-fitgpd(velocidad[,"obs"],16.7)

#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, npx=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

exc<-altura[, "obs"]
exc_v<-velocidad[, "obs"]

PASA LA PRUEBA CON UMBRAL DE 4.5 en altura.
#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(exc, pot_altura)
ad_t(exc, pot_altura)
#Bondad de ajuste para u=9 de velocidad de viento
ks_t(exc_v, pot_velocidad)
ad_t(exc_v, pot_velocidad)

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])
}

#Obtenemos la nueva muestra con las excedencias sobre u
newbase <- subset(Base, Base$SH > 4.3 & Base$V > 16.7)

fitted.values(pot_altura)
fitted.values(pot_velocidad)

dgph <- pgpd(newbase$SH, pot_altura$threshold
,fitted.values(pot_altura)[1],
fitted.values(pot_altura)[2] )
dgpv <- pgpd(newbase$V,pot_velocidad$threshold
,fitted.values(pot_velocidad)[1] ,
fitted.values(pot_velocidad)[2])

gph <- dgph[1:length(dgph)]
gpv <- dgpv[1:length(dgpv)]

Base de excedencias
newBiv_exc <- data.frame(gph,gpv)

```

```

gofCopula(gumbelCopula(),newBiv_exc)
gofEVCopula(huslerReissCopula(),newBiv_exc)
gofCopula(gumbelCopula(),newBiv_exc)
gofEVCopula(tawnCopula(),newBiv_exc)

gumbel<- fitCopula(gumbelCopula(),newBiv_exc,method="ml")
galambos <- fitCopula(galambosCopula(),newBiv_exc, method="ml")
hr<- fitCopula(huslerReissCopula(),newBiv_exc,method="ml")
tawn <- fitCopula(tawnCopula(),newBiv_exc, method="ml")

n <- length(newbase$SH)
d <- 2
family <- "Gumbel"
theta <- 0.40887
cop <- Tawn(0.40887)
u <- matrix(runif(n*d), n, d)
U <- newBiv_exc
U <- pobs(Biv)
ec <- C.n(u, U=U)

## compare with true distribution function
mean(abs(pCopula(u, copula=cop)-ec)) # increase n to decrease this error
## compare the empirical copula and the true copula
## on the diagonal of the unit square
Cn. <- function(x) C.n(do.call(cbind, rep(list(x), d)), U=U)
curve(Cn., 0, 1, main=paste("Diagonal of a", family, "copula"),
xlab="u", ylab=expression(italic(C)[n](italic(u),...,italic(u))))
pC <- function(x) pCopula(do.call(cbind, rep(list(x), d)), cop)
curve(pC, lty=2, add=TRUE)
legend("topleft", lty=1:2, bty="n", inset=0.02,
legend=c(expression(italic(C)[n]), expression(italic(C))))
cCopula 27

```

```
r <- pempiricalCopula(newbase$SH,newbase$V)
empg <- pempiricalCopula(pobs(newbase))
contour(empg, pCopula)
e <- C.n(u,pobs(Biv))
ecs <- F.n(u,U)

persp (norm.cop, dCopula)
contour(norm.cop, pCopula)

gumbel.cop <- gumbelCopula(s$parameter)
tawn.cop <- tawnCopula(0.40887)

persp (tawn.cop, dCopula)
contour(tawn.cop, pCopula, lty=5, col="mediumslateblue",lwd= 2,add=TRUE)

#Obtenemos la nueva muestra con las excedencias sobre u
newbase <- subset(Base, Base$SH > 2.2 & Base$V > 10.4)

contour(gumbelp.cop, pCopula, lty=5, col="darkgoldenrod1", add=TRUE)
```


Bibliografía

- [1] Manual de diseño de obras civiles CFE. 2010, ch. Aspectos generales de meteorología y descripción de las oscilaciones del mar.
- [2] *Statistical Analysis Handbook - (c) 2015 Dr M J de Smith*, [Fecha de consulta: 22 Agosto 2015]. <http://www.statsref.com/HTML/index.html?anderson-darling.html>.
- [3] *NOAA-NDBC Web Site*, [Fecha de consulta: 22 Febrero 2015]. <http://www.ndbc.noaa.gov>.
- [4] ANNE DUTFOY, S. P., AND ROCHE, N. *Multivariate Extreme Value Theory - A Tutorial with Applications to Hydrology and Meteorology*. Dependence modeling review article. De Gruyter Open, 2014.
- [5] BALAKRISHNA, N., AND LAI, C. D. *Continuous Bivariate Distributions*. Springer New York, 2009.
- [6] BEIRLANT, J., G. Y. T. J., AND SEGERS, J. *Statistics of Extremes: Theory and Applications*. John Wiley and Sons. 2004.
- [7] BERMAN, S. M. *Convergence to bivariate limiting extreme value distributions*. Annals of the Institute of Statistical Mathematics. 1961.
- [8] BERNHARD PFAFF, A. M., AND STEPHENSON, A. G. *Package evir , Extreme Values in R*, 2015. <https://cran.r-project.org/web/packages/evir/evir.pdf>.
- [9] CAMPA ROJAS, M. A. *Teoría de Valores Extremos con Aplicaciones a Medidas de Riesgo*. Tesis de Licenciatura, Universidad Nacional Autónoma de México, Facultad de Ciencias. 244 páginas, 2001.

- [10] COLES, S. *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics, University of Bristol UK, 2001.
- [11] E. KOLE, K. K., AND VERBEEK, M. *Selecting copulas for risk management*, vol. 31. J. Bank. Financ., 2007.
- [12] EMBRECHTS, P., L. F., AND MCNEIL, A. *Modelling dependence with copulas and applications to risk management*. Handbook of heavy tailed distributions in finance. 2003.
- [13] EMBRECHTS, P., KLÜPPELBERG, C., AND MIKOSCH, T. *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, 1997.
- [14] ESCHENBURG, P. Properties of extreme value copulas, 2013.
- [15] GENEST, C., AND CATHERINE FAVRE, A. *Everything you always wanted to know about copula modeling but were afraid to ask*. Journal of Hydrologic Engineering, 2007.
- [16] GENEST, C., AND SEGERS, J. *Rank-based inference for bivariate extreme-value copulas*, vol. 37. The Annals of Statistics, 2009.
- [17] GENEST, C., G. K., AND RIVEST, L.-P. *A semiparametric estimation procedure of dependence parameters in multivariate families of distributions*, vol. 82. Biometrika, 1995.
- [18] GENEST, C. KOJADINOVIC, I. G. N. J., AND YAN, J. *A goodness-of-fit test for bivariate extreme-value copulas*, vol. 17. Bernoulli, 2011.
- [19] GHOUDI, K., K. A., AND RIVEST, L.-P. *Multivariate extreme value distributions Propriétés statistiques des copules de valeurs extrêmes bidimensionnelles*, vol. 26. Canad. J. Statist., 1998.
- [20] GIL BELLOSTA, C. *Package ADGofTest , Anderson-Darling GoF test*, 2011. <http://cran.r-project.org/web/packages/ADGofTest/ADGofTest.pdf>.
- [21] GÜLPINAR, N., AND KATATA, K. *Modelling oil and gas supply disruption risks using extreme-value theory and copula.*, vol. 41. Journal of Applied Statistics, 2014.
- [22] GUDENDORF, G., AND SEGERS, J. Extreme-value copulas. In *Copula Theory and Its Applications*, P. Jaworski, F. Durante, W. K. Härdle, and T. Rychlik, Eds., vol. 198 of *Lecture Notes in Statistics*. Springer Berlin Heidelberg, 2010, pp. 127–145.

- [23] 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>.
- [24] HOFERT, M., KOJADINOVIC, I., MAECHLER, M., AND YAN, J. *Package Copula, Multivariate Dependence with Copulas*, 2015. <http://cran.r-project.org/web/packages/copula/copula.pdf>.
- [25] HÜSLER, J., AND REISS, R. *Maxima of normal random vectors: Between independence and complete dependence*. *Statistics and Probability Letters*, 1989.
- [26] JOE, H. *Asymptotic efficiency of the two-stage estimation method for copula-based models*. *Journal of Multivariate Analysis*, 2005.
- [27] JOE, H., AND XU, J. J. *The Estimation Method of Inference Functions for Margins for Multivariate Models*. University of British Columbia, 2005.
- [28] JONATHAN, P., AND EWANS, K. *Statistical modelling of extreme ocean environments for marine design: A review*, vol. 62 of *Ocean Engineering. An International Journal of Research and Development*. Elsevier.
- [29] KAMPHUIS, J. *Introduction to Coastal Engineering and Management*, vol. 30 of *Advanced Series on Ocean Engineering*. World Scientific Publishing, 2010.
- [30] MCNEIL, A. J., FREY, R., AND EMBRECHTS, P. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, Princeton, 2005.
- [31] NELSEN, R. *An Introduction to Copulas*. Springer., 2006.
- [32] PICKANDS III, J. *Multivariate extreme value distributions*, vol. 2. International Statistical Institute, 1981.
- [33] REISS, R., AND THOMAS, M. *Statistical Analysis of Extreme Values: with Applications to Insurance, Finance Hydrology and Other Fields*, 2 ed. Birkhäuser Verlag, 2001.
- [34] RESNICK, S. I. *Extreme Values, Regular Variation and Point Processes*. Springer-Verlag Series in Operations, Research and Financial Engineering, Cornell University, 1987.
- [35] RIBATET, M. *The POT Package*, 2007. <http://benz.nchu.edu.tw/finmyc/POT.pdf>.

- [36] SIBUYA, M. *Bivariate extreme statistics, I*. Annals of the Institute of Statistical Mathematics, 1959.
- [37] SILVA CASARÍN, R. *Análisis y descripción estadística del Oleaje*. Instituto de Ingeniería, Universidad Nacional Autónoma de México., México, 2005.
- [38] 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.
- [39] SKAARUP ANDERSEN, H., AND SLOTH PEDERSEN, D. *Extreme value theory with applications in quantitative risk management*, 2010.
- [40] STEPHENSON, A., AND FERRO, C. *Package evd (Version 2.3-0)*, 2012. <http://cran.r-project.org/web/packages/evd/evd.pdf>.
- [41] TAWN, J. *Bivariate extreme value theory: models and estimation*. Biometrika, 1988.
- [42] THE R GRAPHICS PACKAGE. *Documentation for package ‘graphics’ (Version 2.15.3)*. <http://stat.ethz.ch/R-manual/R-patched/library/graphics/>.
- [43] 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>.
- [44] YAN FANG, L. M., AND LIU, L. *Comparison of Two Methods to Check Copula Fitting*. IAENG International Journal of Applied Mathematics, 2014.