

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE MATEMÁTICA

Proyecto MAT-468

– Simulación Estocástica –

Estudiante:
Clemente Ferrer

Ayudante:
Gabriel Riffo

Profesor:
Francisco Cuevas

Noviembre 2023

TEORÍA DE VALORES EXTREMOS

CLEMENTE FERRER

ABSTRACT

En este trabajo se propone una exploración introductoria de la teoría clásica de valores extremos, la distribución generalizada de valores extremos (GEV) y las metodologías de inferencia asociadas. Se aborda tanto la inferencia clásica mediante máxima verosimilitud, como la inferencia bayesiana mediante MCMC para la distribución GEV. Junto a ello, se revisan herramientas de diagnóstico del modelo, proporcionando un marco sólido para el análisis de extremos climáticos. La aplicabilidad y robustez de los métodos propuestos se demuestran a través de tres conjuntos de datos.

1. INTRODUCCIÓN

La investigación sobre valores extremos, un área incipiente en estadística, se ha consolidado a lo largo de los años como un campo de estudio esencial para comprender y modelar eventos excepcionales. La teoría clásica de valores extremos, iniciada por Frechet, Fisher y Tippet, ha sentado las bases para explorar las propiedades de la distribución del máximo de variables aleatorias independientes e idénticamente distribuidas cuando el tamaño de la muestra tiende a infinito.

Las dos primeras secciones presentan las contribuciones de [Fisher y Tippet \(1928\)](#) sobre las distribuciones límite o asintóticas de máximos parciales, conocidas como Gumbel, Frechet y Weibull. Estas distribuciones, identificadas como Tipo I, Tipo II y Tipo III se engloban como casos particulares de la distribución generalizada de valores extremos, GEV por sus siglas en inglés.

Enseguida, basándonos en el trabajo de [Coles \(2001\)](#), se exploran las técnicas de inferencia para la distribución generalizada de valores extremos, junto con el diagnóstico de los modelos a presentar.

Lo anterior se pondrá en práctica a través de tres aplicaciones específicas: La temperatura invernal en Port Davis, la temperatura diaria anual en Nueva York y la temperatura anual en Santiago, Chile. Con las cuales, este trabajo busca evidenciar la versatilidad y aplicabilidad de las metodologías propuestas, contextualizando la investigación en el marco del cambio climático y su impacto potencial en distintas regiones geográficas.

Finalmente, se discutirá la estimación bayesiana de los parámetros mediante Metropolis-Hastings con caminata aleatoria.

2. TEORÍA CLÁSICA DE VALORES EXTREMOS

En este capítulo se desarrollará el modelo que representa la piedra angular de la teoría de valores extremos. Para ello, definimos los *máximos parciales* como

$$M_n = \max \{X_1, \dots, X_n\},$$

donde X_1, \dots, X_n es una sucesión de variables aleatorias independiente con función de distribución común F .

De ahora en adelante, nos centraremos solo en el análisis de los máximos parciales, pues los resultados pueden extenderse naturalmente al caso de *mínimos parciales* mediante la transformación

$$Y_i = -X_i, \quad i = 1, \dots, n.$$

Luego, es claro que

$$\min\{X_1, \dots, X_n\} = -\max\{Y_1, \dots, Y_n\}.$$

Observación 1. En aplicaciones, las variables aleatorias suelen representar valores de un proceso medidos en una escala de tiempo regular. Por ende, M_n representa el máximo de los eventos a lo largo de n unidades de tiempo observadas.

Note que la distribución de M_n puede ser obtenida de forma explícita

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

sin embargo, esto no es inmediatamente útil en la práctica, ya que la función de distribución F es desconocida. Una posibilidad es utilizar técnicas estándar para estimar F a partir de datos observados, \hat{F} , y luego sustituir esta estimación \hat{F}^n . Desafortunadamente, pequeñas discrepancias en la estimación pueden dar lugar a discrepancias sustanciales para \hat{F}^n .

Un enfoque alternativo es aceptar que F es desconocida y buscar familias aproximadas de modelos para F^n , que pueden estimarse únicamente en función de los datos extremos. Esto es similar a la práctica habitual de aproximar la distribución de los promedios parciales mediante la distribución normal, tal como lo justifica el teorema del límite central. En lo que sigue buscaremos un análogo del teorema anterior para valores extremos.

Ahora bien, estudiando el comportamiento asintótico de $F^n(z)$, un primer problema surge, ya que para cualquier $z < z_+$, donde $z_+ = \sup\{z : F(z) < 1\}$, $F^n(z) \rightarrow 0$ a medida que $n \rightarrow \infty$, por lo que la distribución de M_n se concentra en un punto, z_+ .

Ejemplo 1. Considere $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Exp}(\lambda)$, como $f_X(x) = \lambda e^{-\lambda x}$ se tiene que

$$F_{M_n}(z) = (1 - e^{-\lambda z})^n$$

y por ende $f_{M_n}(z) = \frac{d}{dz} F_{M_n}(z) = n(1 - e^{-\lambda z})^{n-1} \lambda e^{-\lambda z}$. A continuación generamos las curvas de f_{M_n} y F_{M_n} para diversos valores de n con $\lambda = 1$.

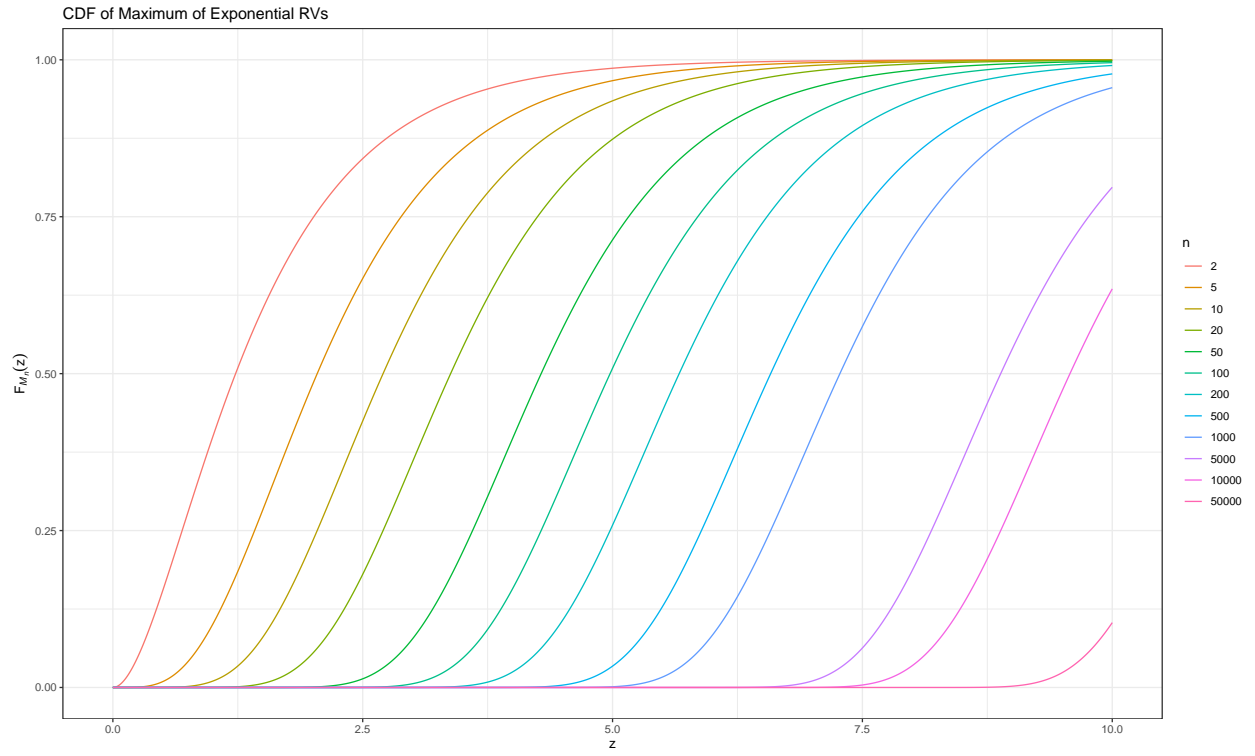


Figura 1. CDF del máximo de variables aleatorias exponenciales para múltiples valores de n .

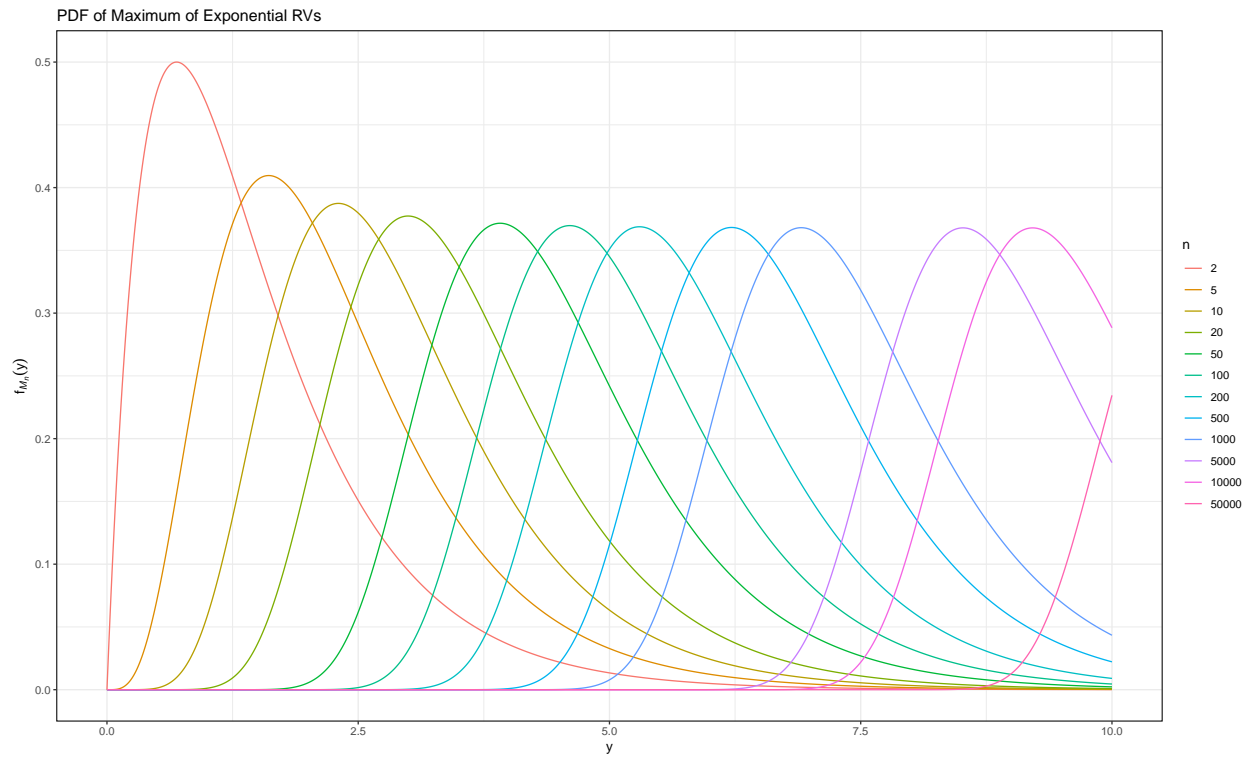


Figura 2. PDF del máximo de variables aleatorias exponenciales para múltiples valores de n .

La Figura 1 muestra como para $z < z_+$ fijo, se tiene que $F^n(z) \rightarrow 0$ a medida que $n \rightarrow \infty$. Este fenómeno se justifica pues la densidad de M_n se va trasladando a la derecha como

se aprecia en la Figura 2. No obstante, para superar esta dificultad consideraremos una normalización lineal de M_n como sigue

$$M_n^* = \frac{M_n - b_n}{a_n},$$

donde $a_n > 0$ y $b_n \in \mathbb{R}$ son sucesiones de constantes elegidas para estabilizar la localización y escala de M_n . Para poner en práctica esta redefinición, prosigamos con el ejemplo exponencial previo.

Ejemplo 1 (Continuación). Para el máximo de variables aleatorias exponenciales independientes e idénticamente distribuidas $P(M_n \leq z) = (1 - e^{-z})^n$, ahora si consideramos los máximos parciales reescalados

$$F_{M_n^*}(z) = P(M_n^* \leq z) = P\left(\frac{M_n - b_n}{a_n} < z\right) = P(M_n \leq a_n z + b_n) = F_{M_n}(a_n z + b_n).$$

Es decir, $F_{M_n^*}(z) = (1 - e^{-a_n z - b_n})^n$ y tomando $a_n = 1$, $b_n = \log n$ se deduce que

$$F_{M_n^*}(z) = (1 - e^{-z - \log(n)})^n = \left(1 - \frac{e^{-z}}{n}\right)^n \rightarrow \exp(-e^{-z}),$$

teniendo una convergencia explícita y distinta de cero para $z < z_+$. La Figura 3 presenta la distribución acumulada para M_n^* bajo las suposiciones anteriores.

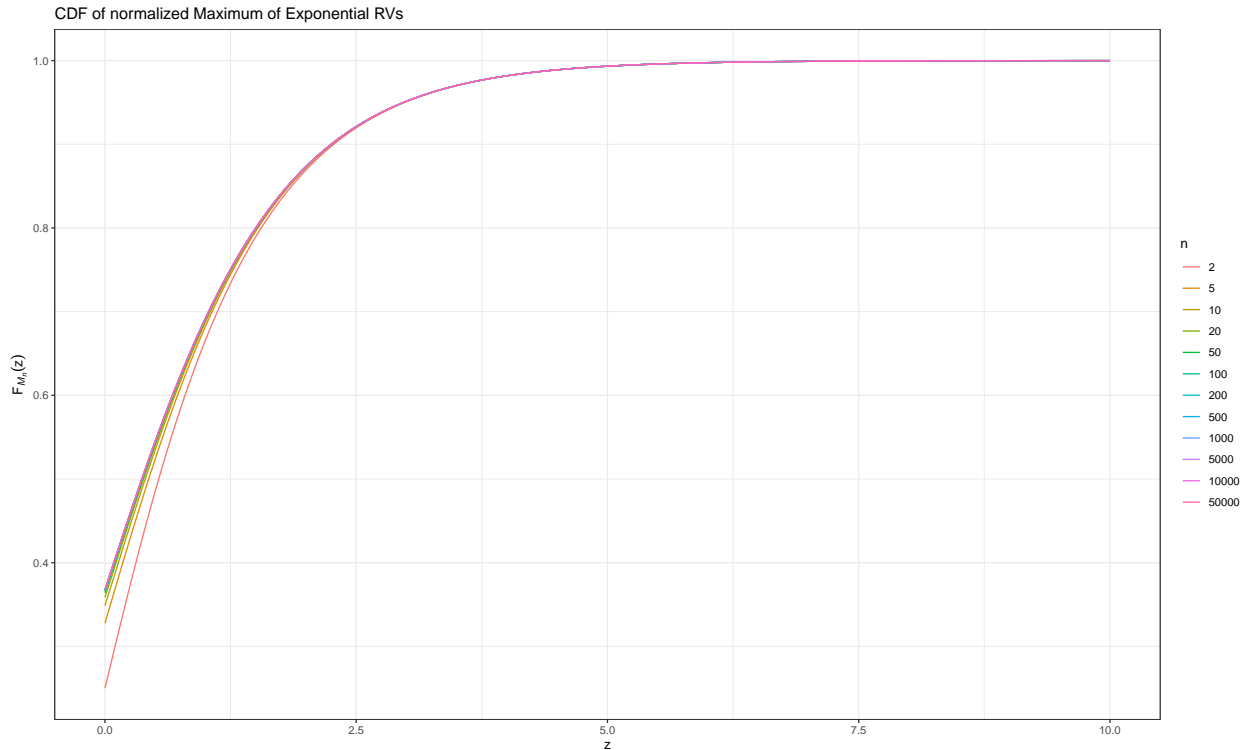


Figura 3. CDF del máximo normalizado de variables aleatorias exponenciales para múltiples valores de n .

Ahora, la pregunta natural es: ¿Cuál es el posible rango de distribuciones a los cuales puede converger M_n^* para a_n y b_n adecuados? El siguiente resultado establecido por [Fisher y Tippet \(1928\)](#) permitió responder esta inquietud.

Teorema 1 (*Ley asintótica del máximo*). Sean X_1, \dots, X_n variables aleatorias independientes e idénticamente distribuidas y defina $M_n := \max\{X_1, \dots, X_n\}$. Si existen sucesiones $\{a_n\}_{n=1}^{\infty}$ y $\{b_n\}_{n=1}^{\infty}$ con $b_n > 0$, para todo $n \in \mathbb{N}$ tales que

$$\lim_{n \rightarrow \infty} P\left(\frac{M_n - b_n}{a_n} \leq z\right) = G(z), \quad z \in \mathbb{R},$$

donde G es una distribución no degenerada, entonces G pertenece a uno de los tres tipos siguientes de distribuciones límite (o asintóticas) de máximos:

I) (Gumbel)

$$G(z) = \exp\left\{-\exp\left(-\frac{z-a}{b}\right)\right\}, \quad -\infty < z < \infty.$$

II) (Fréchet)

$$G(z) = \begin{cases} 0, & z \leq a, \\ \exp\left\{-\left(\frac{z-a}{b}\right)^{-\gamma}\right\}, & z > a. \end{cases}$$

III) (Weibull)

$$G(z) = \begin{cases} \exp\left\{-\left[\left(-\frac{z-a}{b}\right)^\gamma\right]\right\}, & z < a, \\ 1, & z \geq a. \end{cases}$$

Donde, $a \in \mathbb{R}$ es un parámetro de localización, $b > 0$ un parámetro de escala, y $\gamma > 0$ un parámetro de forma.

La característica notable de este resultado es que los tres tipos de distribuciones son los únicos límites posibles para las distribuciones de M_n^* , independientemente de la distribución F de la población. Es en este sentido que el teorema proporciona un análogo del teorema del límite central para valores extremos.

Puede observar que en el Ejemplo 1, la distribución límite de M_n^* corresponde a una Gumbel con $a = 0$ y $b = 1$.

Observación 3. El teorema anterior no implica que cualquier distribución F pertenece a algunos de los tres tipos anteriores, pues deben existir las sucesiones (a_n) y (b_n) para garantizarlo. Un contraejemplo clásico se obtiene al considerar una variable aleatoria X con CDF

$$F(x) = 1 - \frac{1}{\ln x}, \quad x > e.$$

De hecho, F tiene una cola más pesada que las distribuciones tipo Pareto y no existen sucesiones normalizantes. Esta familia de distribuciones es llamada *super-heavy tails*.

La Figura 4 compara las colas derechas para la densidad anterior frente a una normal estándar y una Pareto de parámetros $x_m = e$ y $\alpha = 2$.

Por otro lado, las Figuras 6, 7 y 8 presentan simulaciones que permiten visualizar las diferentes distribuciones Tipo I, II y III para múltiples combinaciones de parámetros.

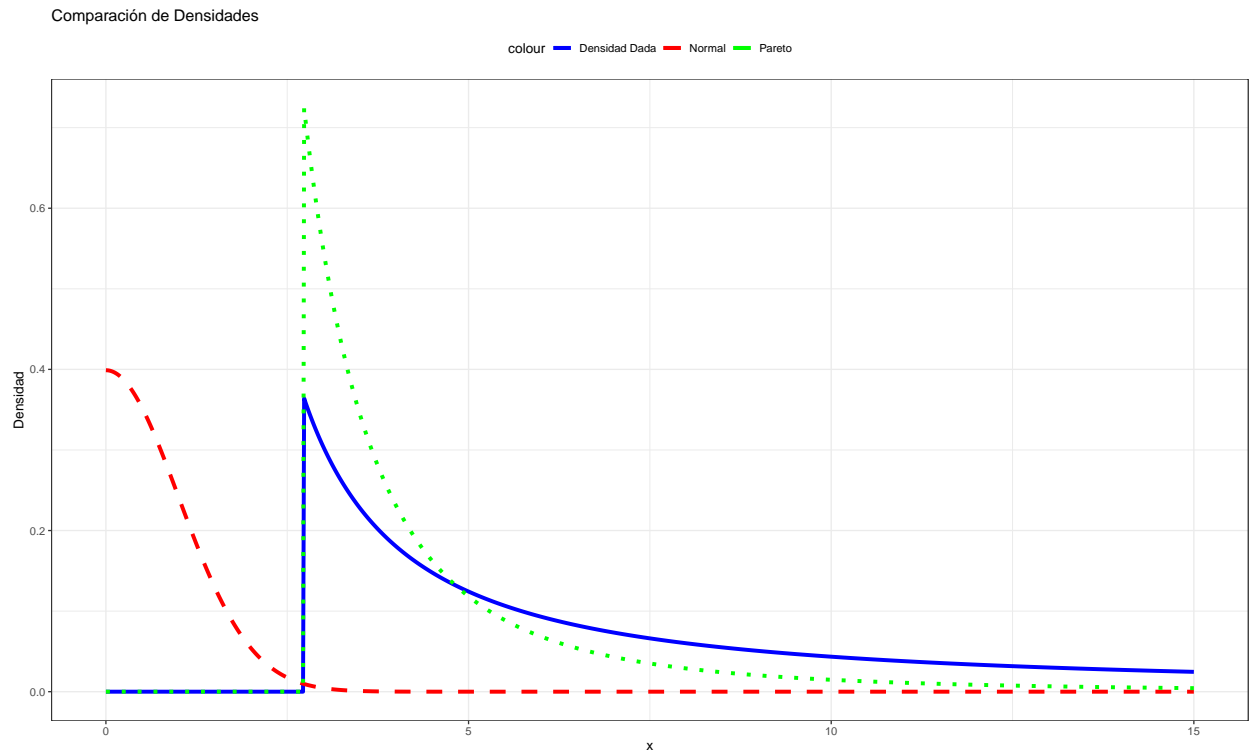


Figura 4. Cola derecha de densidad *super heavy tail*, frente a la densidad normal estándar y Pareto de parámetros $x_m = e$, $\alpha = 2$.

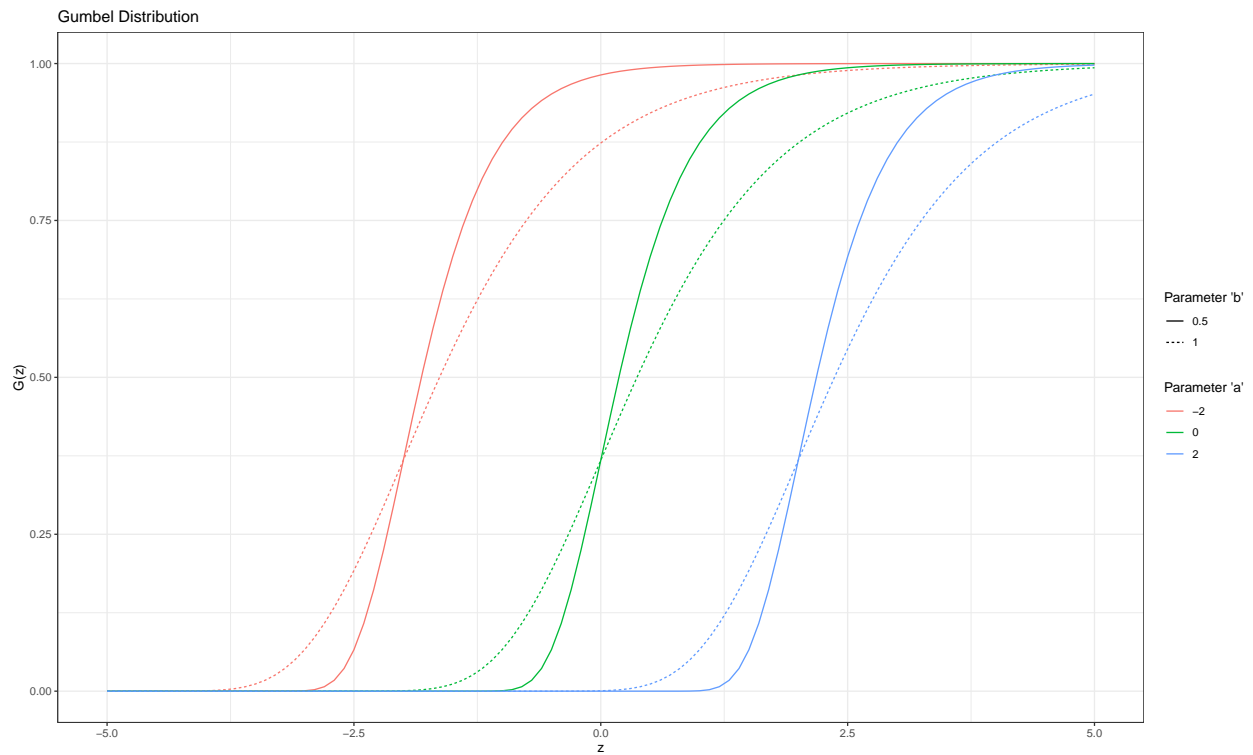


Figura 5. Distribución Gumbel para múltiples parámetros

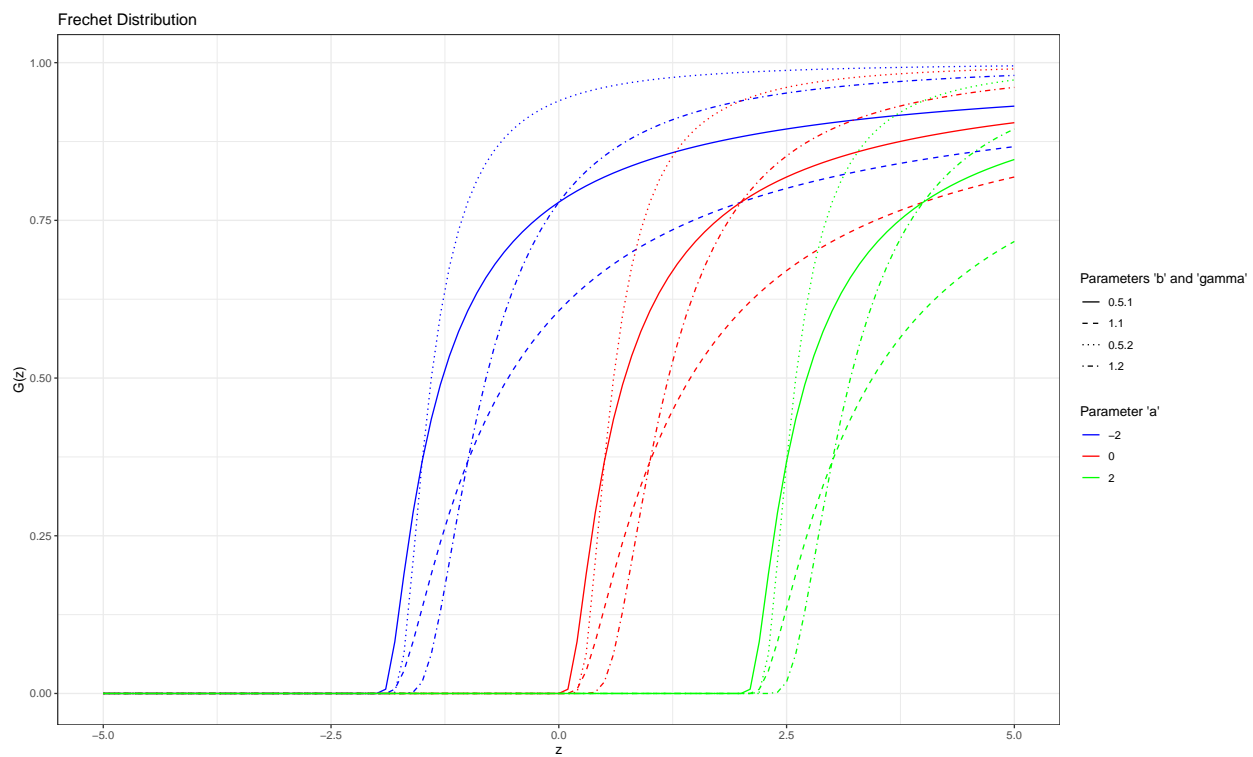


Figura 6. Distribución Fréchet para múltiples parámetros.

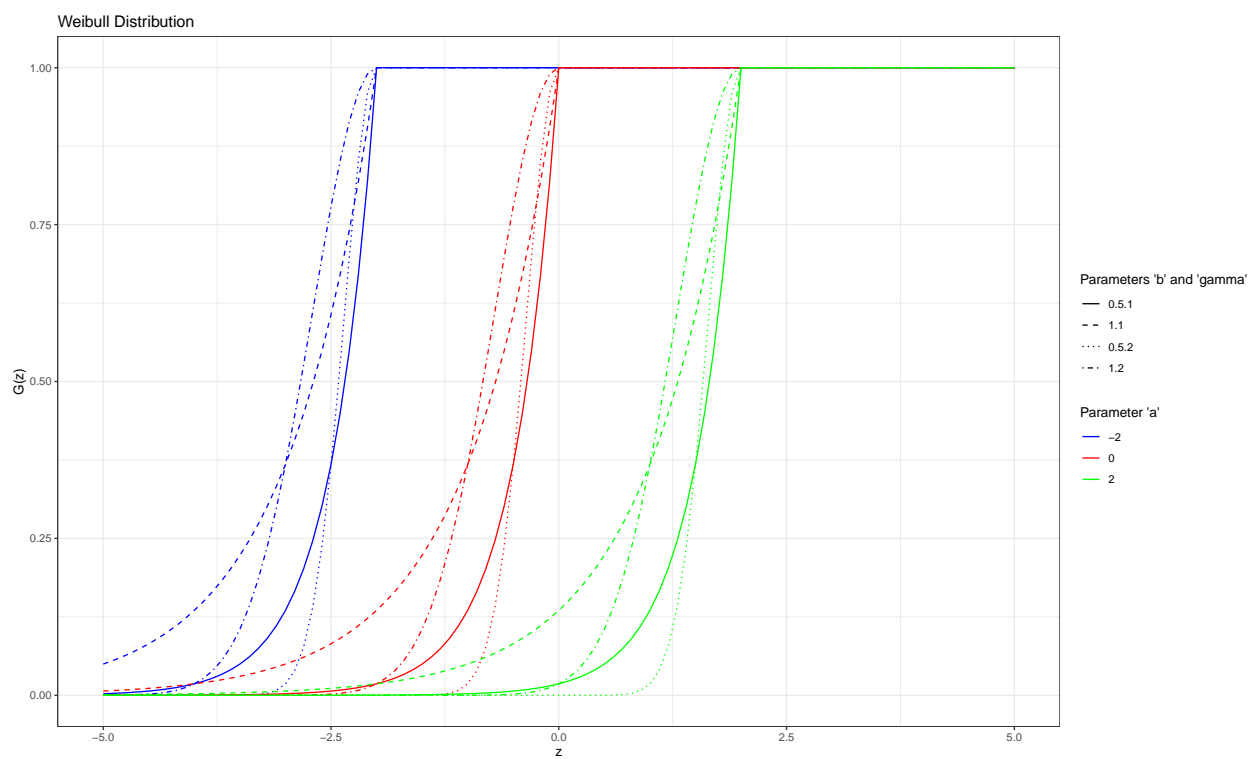


Figura 7. Distribución Weibull para múltiples parámetros.

3. DISTRIBUCIÓN GENERALIZADA DE VALORES EXTREMOS

En las primeras aplicaciones de la teoría de valores extremos, era común adoptar una de las tres familias y luego estimar los parámetros relevantes de esa distribución. Sin embargo, hay dos debilidades: en primer lugar, se necesita una técnica para elegir cuál de las tres familias es más apropiada para los datos en cuestión; en segundo lugar, una vez tomada esa decisión, las inferencias subsiguientes presuponen que esta elección es correcta y no permiten considerar la incertidumbre que dicha selección implica, a pesar de que esta incertidumbre puede ser sustancial.

No es difícil comprobar que las familias Gumbel, Fréchet y Weibull pueden combinarse en una única familia de modelos.

Definición. La distribución generalizada de valores extremos, GEV por sus siglas en inglés, es

$$G(z) = \exp \left\{ - \left[1 + \xi \left(\frac{z - \mu}{\sigma} \right) \right]^{\frac{-1}{\xi}} \right\},$$

definida en el conjunto $\{z : 1 + \xi(z - \mu)/\sigma > 0\}$, donde $\mu, \xi \in \mathbb{R}$ y $\sigma > 0$.

El modelo posee tres parámetros μ , σ y ξ siendo localización, escala y forma respectivamente. Las clases Tipo II y III corresponden a los casos $\xi > 0$ y $\xi < 0$ en esta parametrización. Además, $\xi = 0$ debe interpretarse como $\xi \rightarrow 0$ para $G(z)$, convergiendo al caso Gumbel. La Figura 8 presenta simulaciones que ejemplifican la discusión anterior.

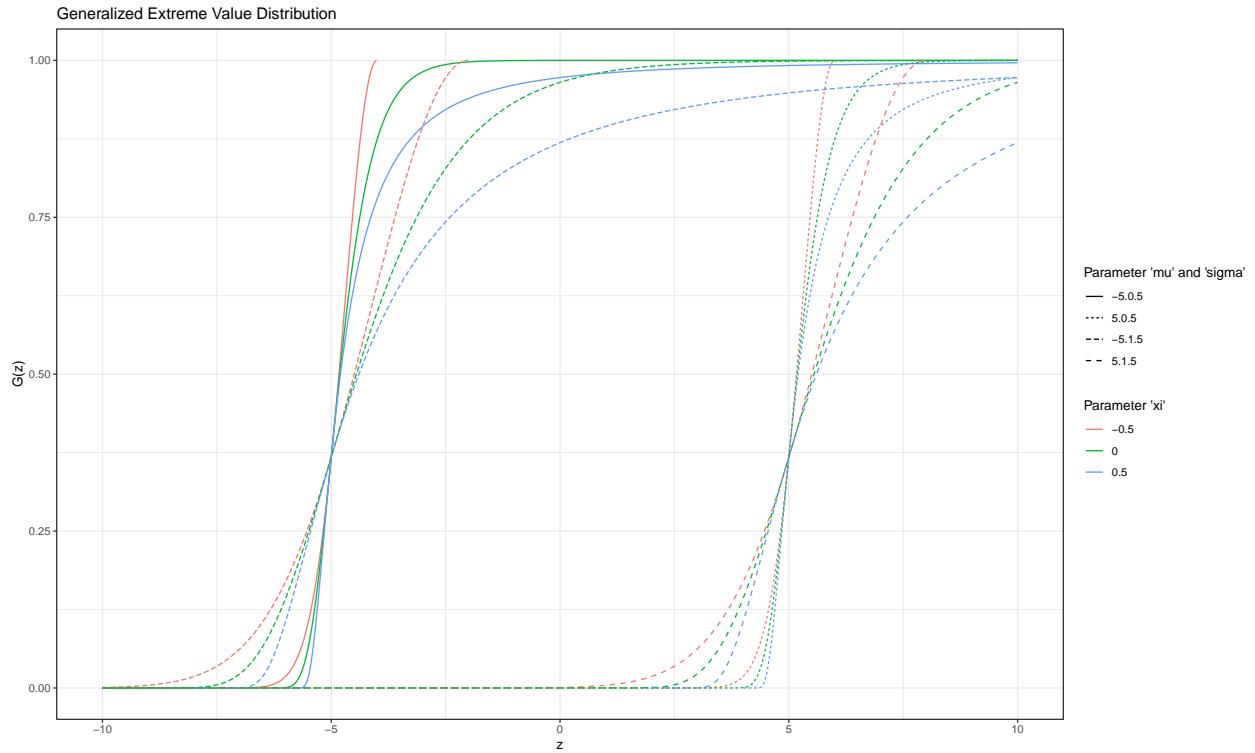


Figura 8. Distribución generalizada de valores extremos para múltiples parámetros.

La unificación de las tres familias originales de distribución de valores extremos en una sola familia simplifica enormemente la implementación estadística. A través de la inferencia,

los propios datos determinan el tipo más apropiado de comportamiento de la cola, y no es necesario realizar juicios subjetivos a priori sobre cuál familia individual de valores extremos adoptar. Además, la incertidumbre en el valor inferido de ξ mide la falta de certeza sobre cuál de los tres tipos originales es más apropiado para un conjunto de datos dado.

Por conveniencia, volveremos a enunciar el Teorema 1 considerando la definición anterior.

Teorema 2. Si existen sucesiones $\{a_n\}_{n=1}^{\infty}$ y $\{b_n\}_{n=1}^{\infty}$ con $b_n > 0$, para todo $n \in \mathbb{N}$ tales que

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\frac{M_n - b_n}{a_n} \leq z \right) = G(z), \quad z \in \mathbb{R},$$

donde G es una distribución no degenerada, entonces

$$G(z) = \exp \left\{ - \left[1 + \xi \left(\frac{z - \mu}{\sigma} \right) \right]^{\frac{-1}{\xi}} \right\}, \quad (1)$$

definida en el conjunto $\{z : 1 + \xi(z - \mu)/\sigma > 0\}$, donde $\mu, \xi \in \mathbb{R}$ y $\sigma > 0$.

A continuación se presentan ejemplos de uso del teorema anterior.

Ejemplo 2. Considere $X_1, \dots, X_n \stackrel{iid}{\sim} U(0, 1)$. Note que $\mathbb{P}(M_n \leq z) = z^n$, cuyo límite es 0 si $z \neq 1$. Luego, reescalando

$$F_{M_n^*}(z) = F_{M_n}(a_n z + b_n) = (a_n z + b_n)^n.$$

Escogiendo $a_n = 1/n$ y $b_n = 1$ tenemos que

$$F_{M_n^*}(z) = \left(1 + \frac{z}{n} \right)^n \rightarrow e^z.$$

Comparando con la distribución GEV, notamos que $(\mu, \sigma, \xi) = (-1, 1, -1)$.

Ejemplo 3. Considere $X_1, \dots, X_n \stackrel{iid}{\sim} f(z)$, donde $f(z) = \frac{dF(z)}{dz}$ y $F(z) = e^{-1/z}$ es la distribución Fréchet estándar, $z > 0$. Note que $F^n(z) = e^{-n/z} \rightarrow 0$ para $n \rightarrow \infty$, se sigue reescalando que

$$F_{M_n^*}(z) = F_{M_n}(a_n z + b_n) = e^{-n/(a_n z + b_n)}.$$

Escogiendo $a_n = n$ y $b_n = 0$ tenemos que

$$F_{M_n^*}(z) = e^{-1/z} \rightarrow e^{-1/z}.$$

Comparando con la distribución GEV, notamos que $(\mu, \sigma, \xi) = (1, 0, 1)$.

Observación 4. La aparente dificultad de que las constantes de normalización sean desconocidas en la práctica se resuelve fácilmente asumiendo

$$\mathbb{P} \left(\frac{M_n - b_n}{a_n} \leq z \right) \approx G(z)$$

para n suficientemente grande. Equivalentemente

$$\mathbb{P}(M_n \leq z) \approx G \left(\frac{M_n - b_n}{a_n} \leq z \right) = G^*(z),$$

donde G^* es otra distribución perteneciente a la familia GEV. En otras palabras, si el Teorema 2 permite aproximar M_n^* por un miembro de la familia GEV para n grande, entonces M_n puede ser aproximado por un miembro diferente de la familia GEV. Además, dado que los parámetros de la distribución deben estimarse de todos modos, en la práctica es irrelevante que los parámetros de la distribución G sean diferentes de los de G^* .

Este argumento da el puntapié inicial para modelar los extremos de una serie de observaciones independientes. Los datos se agrupan en secuencias de observaciones de longitud n , para algún valor grande de n , generando una serie de máximos de bloques, $M_{n,1}, \dots, M_{n,m}$, a los cuales se puede ajustar la distribución GEV. A menudo, los bloques se eligen para corresponder a un período de tiempo de un año, en cuyo caso n es el número de observaciones en un año y los bloques son máximos anuales. Las estimaciones de los cuantiles extremos de la distribución de máximos anuales se obtienen invirtiendo la Ecuación (1):

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi} [1 - \{-\log(1-p)\}^{-\xi}], & \text{si } \xi \neq 0, \\ \mu - \sigma \log\{-\log(1-p)\}, & \text{si } \xi = 0, \end{cases}$$

donde $G(z_p) = 1-p$. En términos comunes, z_p es el nivel de retorno asociado con el periodo de retorno $T = 1/p$, ya que con un grado razonable de precisión, se espera que el nivel z_p sea superado en promedio una vez cada $1/p$ años. Más precisamente, z_p es superado por el máximo anual en cualquier año específico con probabilidad p .

Si se representa z_p en función de $\log(T) = \log(1/p) = -\log(p)$ se obtiene un gráfico de niveles de retorno como el que se muestra en la Figura 8.

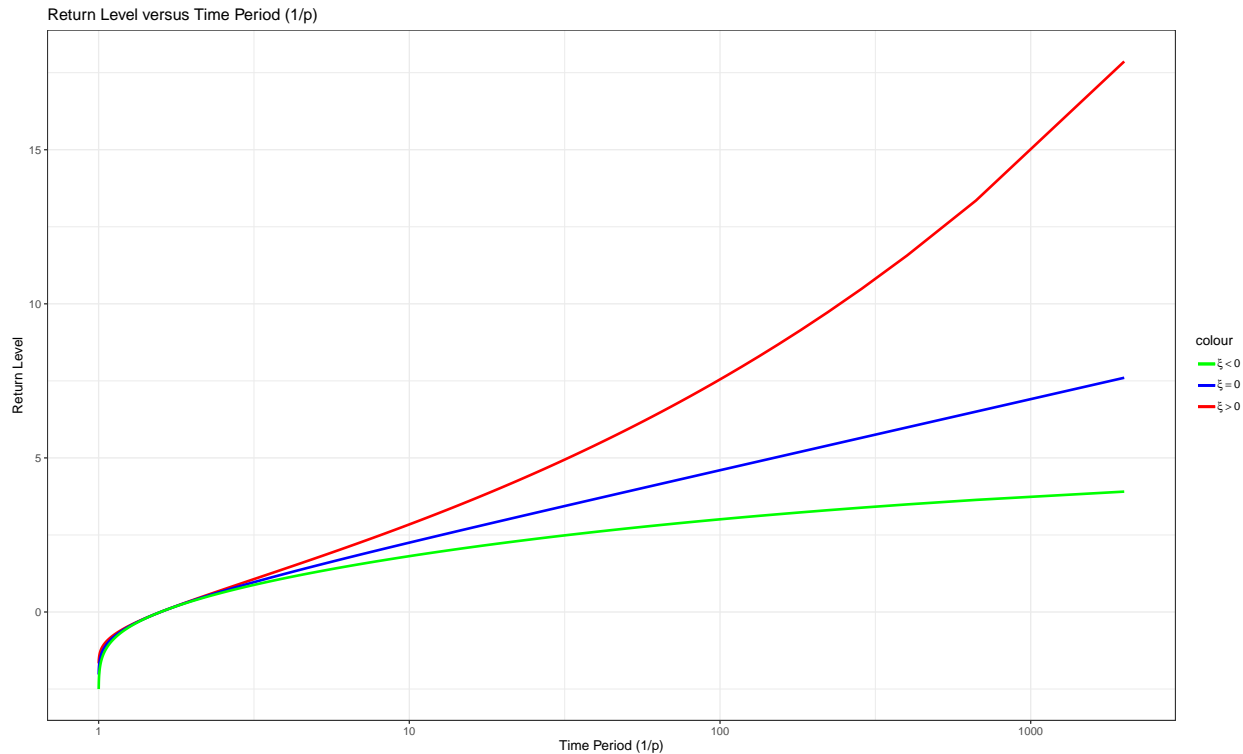


Figura 9. Gráfico de niveles de retorno para $\xi < 0$, $\xi = 0$ y $\xi > 1$ considerando $\mu = 0$ y $\sigma = 1$.

4. INFERENCIA PARA LA DISTRIBUCIÓN GEV

Se han propuesto muchas técnicas para la estimación de parámetros en modelos de valores extremos. Estas incluyen gráficos de probabilidad; técnicas basadas en momentos en las cuales las funciones de los momentos del modelo se igualan con sus equivalentes empíricos; procedimientos en los cuales los parámetros se estiman como funciones especificadas de estadísticas de orden; y métodos basados en verosimilitud. Cada técnica tiene sus pros y contras, pero la utilidad general y la adaptabilidad a la construcción de modelos complejos de las técnicas basadas en verosimilitud hacen que este enfoque sea particularmente atractivo.

Ahora bien, una dificultad potencial con el uso de métodos de verosimilitud para el modelo de valores extremos generalizados (GEV) se refiere a las condiciones de regularidad que son necesarias para que las propiedades asintóticas habituales asociadas con el estimador de máxima verosimilitud sean válidas. Como menciona [Smith \(1985\)](#) existen tres casos:

- 1) Cuando $\xi > -0,5$, los estimadores de máxima verosimilitud son regulares, en el sentido de tener las propiedades asintóticas usuales.
- 2) Cuando $-1 < \xi < -0,5$, los estimadores de máxima verosimilitud son generalmente obtenibles, pero no tienen las propiedades asintóticas estándar.
- 3) Cuando $\xi < -1$, es poco probable que se obtengan estimadores de máxima verosimilitud.

El caso $\xi \leq -0,5$ corresponde a distribuciones con una cola superior muy corta y acotada. Esta situación rara vez se encuentra en aplicaciones de modelado de valores extremos, por lo que las limitaciones teóricas del enfoque de máxima verosimilitud generalmente no son un obstáculo en la práctica.

En lo que sigue, usaremos la notación $Z_i := M_{n,i}$, de modo que $Z_i \stackrel{iid}{\sim} \text{GEV}(\mu, \sigma, \xi)$, para $i = 1, \dots, m$. La independencia viene del hecho de asumir que la muestra original X_i lo es. Luego la log-verosimilitud de la muestra viene dada por

$$\ell(\mu, \sigma, \xi; \mathbf{z}) = -m \log \sigma - (1 + 1/\xi) \sum_{i=1}^m \log \left[1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) \right] - \sum_{i=1}^m \left[1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) \right]^{-1/\xi}$$

siempre que

$$1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) > 0, \quad \text{para } i = 1, \dots, m$$

Si esta condición no se cumple, la verosimilitud es cero y por ende $\ell(\mu, \sigma, \xi; \mathbf{z}) = -\infty$. Si $\xi = 0$, la log-verosimilitud debe ser estudiada considerando el modelo Gumbel. En este caso

$$\ell(\mu, \sigma; \mathbf{z}) = -m \log \sigma - \sum_{i=1}^m \left(\frac{z_i - \mu}{\sigma} \right) - \sum_{i=1}^m \exp \left\{ - \left(\frac{z_i - \mu}{\sigma} \right) \right\}.$$

Mediante métodos numéricos, es posible obtener $\hat{\boldsymbol{\theta}}_{\text{ML}} = (\hat{\mu}_{\text{ML}}, \hat{\sigma}_{\text{ML}}, \hat{\xi}_{\text{ML}})^\top$. Más aún, sujeto a las limitaciones de ξ discutidas previamente, la distribución asintótica de $\hat{\boldsymbol{\theta}}_{\text{ML}}$ es normal multivariada con vector de medias $\boldsymbol{\mu} = \boldsymbol{\theta}$ y matriz de covarianza $\boldsymbol{\Sigma} = \mathcal{I}(\hat{\boldsymbol{\theta}}_{\text{ML}})^{-1}$, los cálculos explícitos pueden ser hallados en [Coles \(2001\)](#).

Por otro lado, dada la invarianza de los estimadores máximo verosímiles es posible obtener $\hat{z}_{p\text{ML}}$ como sigue

$$\hat{z}_{p\text{ML}} = \begin{cases} \hat{\mu}_{\text{ML}} - \frac{\hat{\sigma}_{\text{ML}}}{\hat{\xi}_{\text{ML}}} \left[1 - y_p^{-\hat{\xi}_{\text{ML}}} \right], & \text{si } \hat{\xi}_{\text{ML}} \neq 0, \\ \hat{\mu}_{\text{ML}} - \hat{\sigma}_{\text{ML}} \log y_p, & \text{si } \hat{\xi}_{\text{ML}} = 0, \end{cases}$$

donde $y_p = -\log(1 - p)$. Además, mediante el método delta,

$$\text{Var}(\hat{z}_p) \approx \nabla z_p^\top \Sigma \nabla z_p$$

donde Σ es la matriz de covarianza de $\hat{\theta}_{\text{ML}}$ y

$$\nabla z_p^\top = \left(\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \xi} \right) = (1, -\xi^{-1} (1 - y_p^{-\xi}), \sigma \xi^{-2} (1 - y_p^{-\xi}) - \sigma \xi^{-1} y_p^{-\xi} \log y_p)$$

evaluada en $\hat{\theta}_{\text{ML}}$.

5. DIAGNÓSTICO DEL MODELO

Como es bien sabido en cursos básicos de estadística, el uso de gráficos de probabilidad (o PP) y gráficos de cuantiles (o QQ) puede revelar información muy útil sobre el grado de concordancia entre una distribución de probabilidad asumida o estimada y la distribución empírica de los datos. Estas son también herramientas altamente útiles para realizar una verificación visual de modelos GEV ajustados en casos particulares. Para una discusión exhaustiva sobre el uso de estos gráficos, ver [Beirlant et al. \(2004\)](#).

Un gráfico de probabilidad o PP es una comparación directa del modelo de distribución ajustado con la distribución empírica. Supongamos que la muestra de máximos por bloque ha sido ordenada de forma creciente $z_{(1)} \leq z_{(2)} \leq \dots \leq z_{(m)}$. La función de distribución empírica, \tilde{G} , evaluada en $z_{(i)}$ es dada por

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

Mediante sustitución de los estimadores máximo verosímiles en G y evaluada en $z_{(i)}$, se obtiene $\hat{G}_{\text{ML}}(z_{(i)})$. Ahora, si el modelo GEV es una buena aproximación para los datos, $\hat{G}_{\text{ML}}(z_{(i)}) \approx \tilde{G}(z_{(i)})$ para $i = 1, \dots, m$. Por lo tanto, el gráfico de probabilidad consiste en

$$\{(\hat{G}_{\text{ML}}(z_{(i)}), \tilde{G}(z_{(i)})), i = 1, \dots, m\}$$

y debe estar cercano a la diagonal.

Sin embargo, no todo es perfecto, una debilidad del gráfico de probabilidad para modelos de valores extremos es que tanto $\hat{G}_{\text{ML}}(z_{(i)})$ como $\tilde{G}(z_{(i)})$ tienden a aproximarse a 1 a medida que $z_{(i)}$ aumenta, lo que dificulta visualizar la concordancia para valores altos en los datos, lo cual es primordial para el estudio de valores extremos. Es decir, el gráfico de probabilidad proporciona menos información en la región de mayor interés. Esta deficiencia se evita mediante el gráfico de cuantiles, que consiste en los puntos:

$$\left\{ \left(\hat{G}_{\text{ML}}^{-1}(i/(m + 1)), z_{(i)} \right), i = 1, \dots, m \right\},$$

donde

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

Las desviaciones de la diagonal en el gráfico de cuantiles también indican un fallo en el modelo.

6. APLICACIÓN I: TEMPERATURA MÁXIMA INVERNAL EN PORT DAVIS

Para demostrar el ajuste de la función de distribución de valores extremos generalizados (GEV) a datos de máximos por bloque, utilizaremos datos de temperatura máxima invernal desde 1927 hasta 1995 en Port Jervis, Nueva York ([Thompson y Wallace 1998](#); [Wettstein y Mearns 2002](#)). En esta y en las aplicaciones póstumas haremos uso de las librerías `extRemes` y `extRemes 2.0` desarrolladas por [Gilleland y Katz \(2016\)](#).

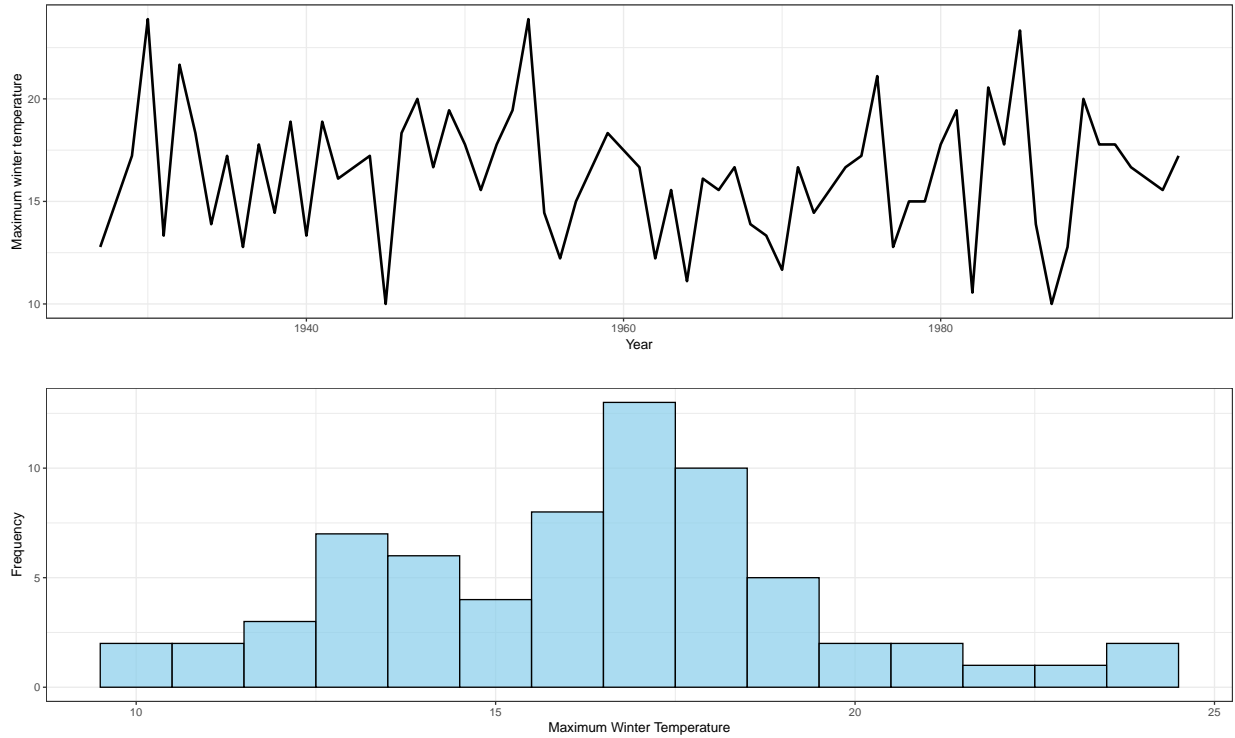


Figura 10. Análisis exploratorio de los datos extremos. Temperatura máxima invernal anual desde 1927 hasta 1995 en Port Jervis, Nueva York (arriba), junto con su distribución (abajo). El histograma ilustra la distribución de frecuencia de las temperaturas máximas invernales anuales a lo largo de los años, ofreciendo perspectivas sobre la variabilidad térmica dentro del período estudiado.

Del gráfico notamos una lateralización de la observaciones, presentando una zona de soporte en las temperaturas máximas entre los 10°C y 24°C aproximadamente, lo cual se condice al observar el histograma que posee mayor concentración de extremos alrededor de las temperaturas de 15°C y 20°C.

	μ	σ	ξ
Estimaciones	15.14	2.97	-0.22
Error estándar	0.40	0.28	0.07
IC asintóticos al 95 %	(14,36, 15,92)	(2,43, 3,51)	(-0,36, -0,07)
Matriz de covarianza			
μ	0.16	0.01	-0.01
σ		0.08	0.01
ξ			0.01
Negative log-likelihood	172.74		
AIC	351.49		
BIC	358.14		
Niveles de retorno estimados (°C)		IC asintótico al 95 %	
2 años	16,19	(15,38, 16,99)	
20 años	21,65	(20,40, 22,89)	
100 años	23,79	(21,79, 25,78)	

Cuadro 1. Parámetros estimados de la función de distribución de valores extremos generalizados (GEV) y niveles de retorno ajustados a la temperatura máxima invernal anual (grados centígrados) en los datos de Port Jervis, Nueva York, que se muestran en la Figura 10.

Es decir, $\hat{\xi}_{ML} < 0$ y por ende los máximos se ajustan mejor a una densidad Weibull. Hagamos el chequeo del modelo mediante los gráficos discutidos previamente:

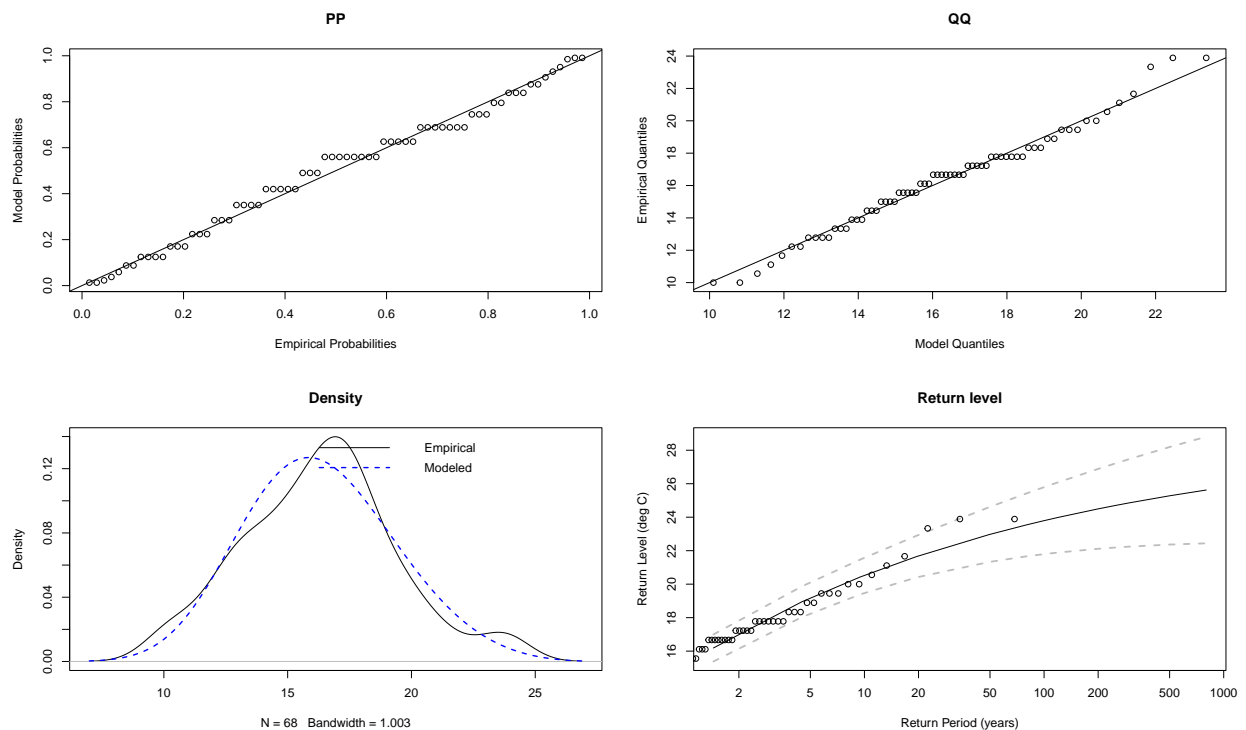


Figura 11. Diagnóstico de la función de distribución de valores extremos generalizados (GEV) ajustada a los datos de la temperatura máxima invernal en Port Jervis. Gráfico de probabilidad (arriba a la izquierda), gráfico de cuantiles (arriba a la derecha), gráficos de densidad de los datos empíricos y la función de distribución GEV ajustada (abajo a la izquierda), y gráfico de niveles de retorno con intervalos de confianza del 95 % mediante una aproximación asintóticamente normal (abajo a la derecha).

Es decir, el modelo Weibull se ajusta de buena manera a los datos extremos. Además, se espera que en promedio el máximo de la temperatura invernal sea de $16,19^{\circ}\text{C}$ una vez cada 2 años, $21,65^{\circ}\text{C}$ una vez cada 20 años y $23,79^{\circ}\text{C}$ una vez cada 100 años.

7. APLICACIÓN II: TEMPERATURA MÁXIMA ANUAL EN NUEVA YORK

Como segunda aplicación, emplearemos la base de datos de temperaturas diarias de Nueva York entre 1951 y 2017, proporcionada por la Administración Nacional Oceánica y Atmosférica de Estados Unidos (NOAA por sus siglas en inglés). Nos enfocaremos en calcular el máximo anual de la temperatura promedio diaria, con el objetivo de obtener una muestra de tamaño $m = 67$ de valores extremos. Posteriormente, procederemos a ajustar un modelo de valores extremos generalizado.

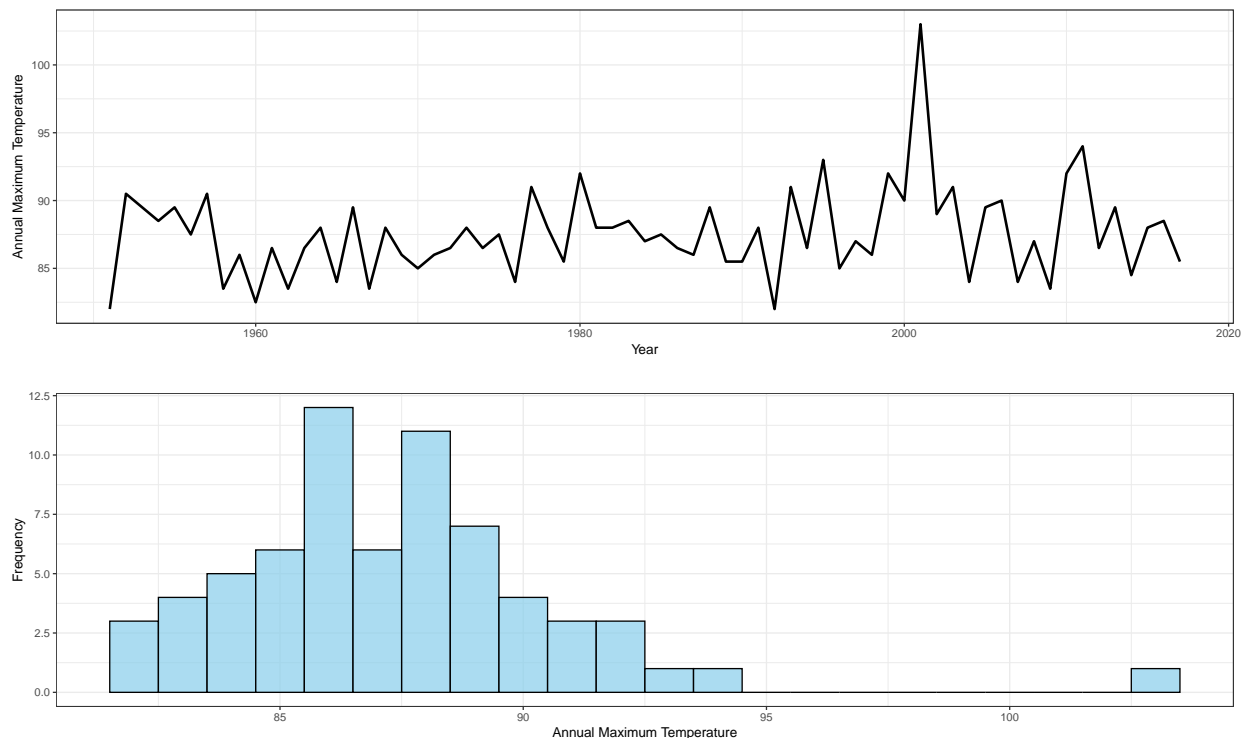


Figura 12. Análisis exploratorio de los datos extremos. Máximo anual de la temperatura promedio diaria en Nueva York entre 1951 y 2017 (arriba), junto con su distribución (abajo). El histograma ilustra la distribución de frecuencia de las temperaturas máximas anuales a lo largo de los años, ofreciendo perspectivas sobre la variabilidad térmica dentro del período estudiado.

Observando los gráficos generados, emerge un patrón distintivo: la presencia de un valor singular que se aparta notablemente del resto del conjunto de datos, destacándose como un valor atípico. Este fenómeno es característico de los valores extremos, sugiriendo la ocurrencia de un evento o condición excepcional que influye significativamente en la distribución general. Por ello, ajustaremos una distribución de valores extremos generalizada a los datos.

	μ	σ	ξ
Estimaciones	86.14	2.64	-0.02
Error estándar	0.35	0.25	0.06
IC asintóticos al 95 %	(85,44, 86,84)	(2,15, 3,13)	(-0,15, 0,10)
Matriz de covarianza			
μ	0.12	0.02	-0.01
σ		0.06	-0.01
ξ			0.01
Negative log-likelihood	169.16		
AIC	344.32		
BIC	350.93		
Niveles de retorno estimados (°F)		IC asintótico al 95 %	
2 años	87,11	(86,34, 87,87)	
20 años	93,70	(91,80, 95,60)	
100 años	97,60	(94,06, 101,1)	

Cuadro 2. Parámetros estimados de la función de distribución de valores extremos generalizados (GEV) y niveles de retorno ajustados a la temperatura máxima anual de la temperatura diaria (grados Fahrenheit) en Nueva York, que se muestran en la Figura 12.

Es decir, $\hat{\xi}_{MV} \approx 0$ y por ende los datos siguen una densidad Gumbel. Más aún, el diagnóstico presente en la Figura 13 denota un buen ajuste del modelo.

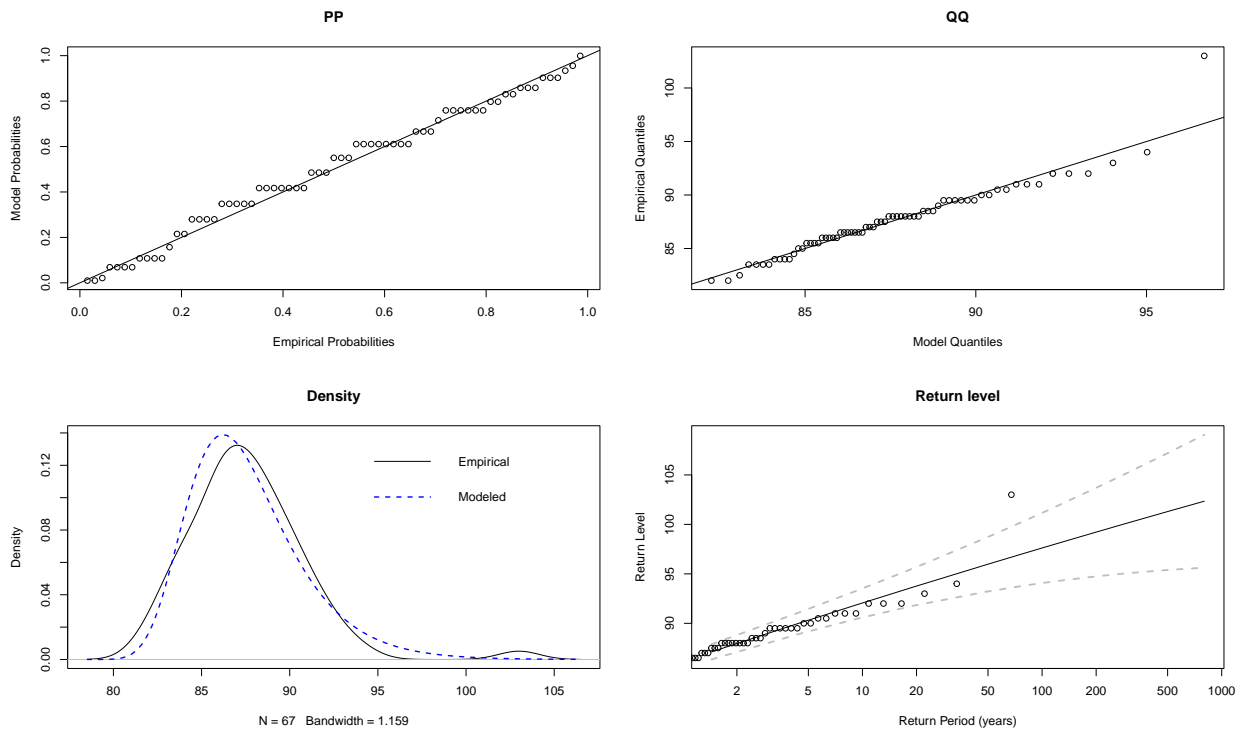


Figura 13. Diagnóstico de la función de distribución de valores extremos generalizados (GEV) ajustada a los datos de la temperatura máxima diaria en Nueva York. Gráfico de probabilidad (arriba a la izquierda), gráfico de cuantiles (arriba a la derecha), gráficos de densidad de los datos empíricos y la función de distribución GEV ajustada (abajo a la izquierda), y gráfico de niveles de retorno con intervalos de confianza del 95 % mediante una aproximación asintóticamente normal (abajo a la derecha).

8. APLICACIÓN III: TEMPERATURA MÁXIMA ANUAL EN SANTIAGO, CHILE

En la tercera y última aplicación clásica, se utiliza la base de datos de temperaturas diarias de Santiago, registradas en el Aeropuerto Arturo Merino Benítez desde 1973 hasta 2023. La obtención de los datos se llevó a cabo mediante una solicitud formal a la Administración Nacional Oceánica y Atmosférica de Estados Unidos. Siguiendo la metodología empleada en la sección anterior, se procederá al cálculo del máximo anual de la temperatura máxima diaria, generando así una muestra de $m = 49$ valores extremos, los cuales se utilizarán para la posterior adecuación de un modelo de distribución de valores extremos generalizada (GEV).

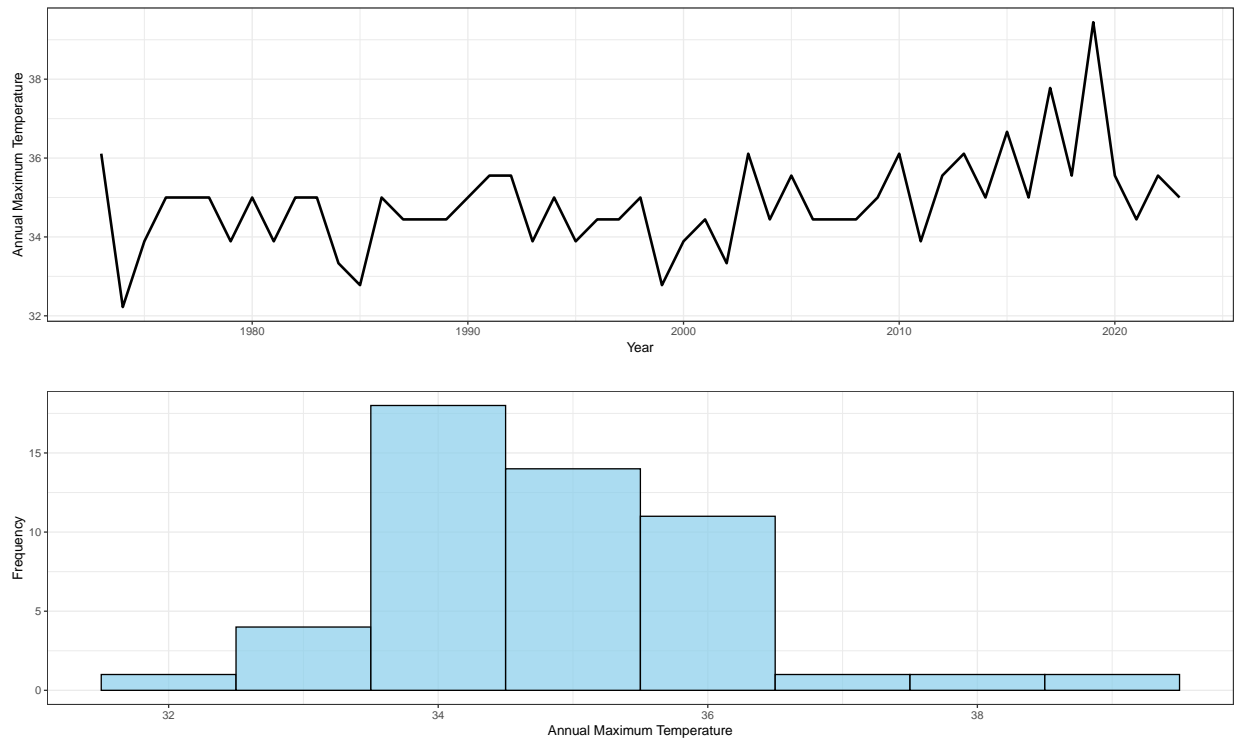


Figura 14. Análisis exploratorio de los datos extremos. Máximo anual de la temperatura promedio diaria en Santiago, Chile entre 1973 y 2023 (arriba), junto con su distribución (abajo). El histograma ilustra la distribución de frecuencia de las temperaturas máximas anuales a lo largo de los años, ofreciendo perspectivas sobre la variabilidad térmica dentro del período estudiado.

La representación gráfica de las temperaturas máximas anuales en Santiago, Chile, revela una tendencia notoria hacia un aumento constante. En particular, los últimos cuatro años exhiben temperaturas que rozan los 40 grados Celsius, marcando un fenómeno preocupante que parece ser exclusivo de este período reciente. Este incremento térmico pronunciado contrasta con la estabilidad observada en décadas previas, donde tales valores extremos eran prácticamente inexistentes. Este patrón ascendente sugiere un impacto significativo del cambio climático en la región, manifestándose en un aumento alarmante de las temperaturas anuales. Este fenómeno no solo destaca la vulnerabilidad climática de Chile, sino que también subraya la importancia de comprender y abordar los cambios ambientales que podrían tener consecuencias a largo plazo en la región.

Enseguida, un ajuste de valores extremos GEV nos provee el siguiente modelo:

	μ	σ	ξ
Estimaciones	34.36	1.06	-0.10
Error estándar	0.16	0.10	0.06
IC asintóticos al 95 %	(34,05, 34,68)	(0,85, 1,27)	(-0,22, 0,01)
Matriz de covarianza			
μ	0.02	0.01	-0.01
σ		0.01	-0.01
ξ			0.01
Negative log-likelihood	79.39		
AIC	164.78		
BIC	170.58		
Niveles de retorno estimados (°F)		IC asintótico al 95 %	
2 años	34,75	(34,41, 35,08)	
20 años	37,08	(36,41, 37,74)	
100 años	38,25	(37,18, 39,31)	

Cuadro 3. Parámetros estimados de la función de distribución de valores extremos generalizados (GEV) y niveles de retorno ajustados a la temperatura máxima anual de la temperatura diaria (grados celsius) en Santiago de Chile, que se muestran en la Figura 14.

Es decir, $\hat{\xi}_{ML} < 0$ y por ende los máximos se ajustan mejor a una densidad Weibull. No obstante, el diagnóstico del modelo muestra que no es adecuado para los datos.

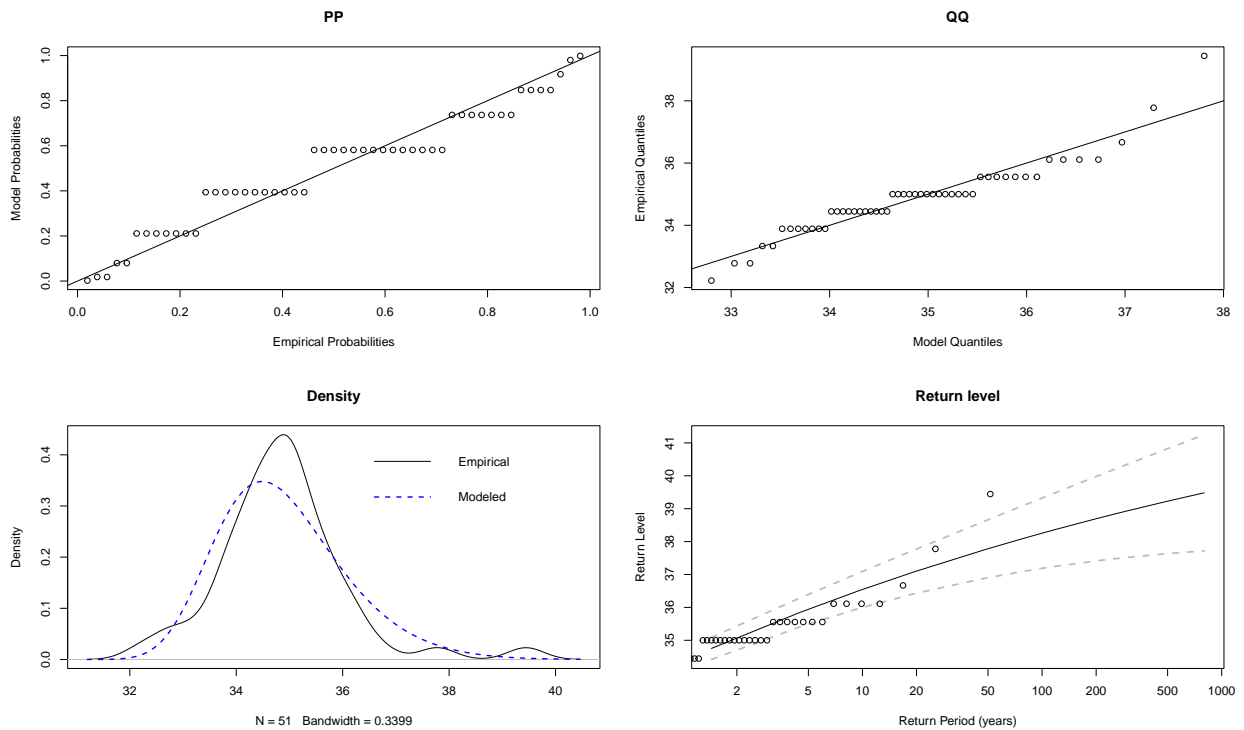


Figura 15. Diagnóstico de la función de distribución de valores extremos generalizados (GEV) ajustada a los datos de la temperatura máxima diaria en Santiago de Chile. Gráfico de probabilidad (arriba a la izquierda), gráfico de cuantiles (arriba a la derecha), gráficos de densidad de los datos empíricos y la función de distribución GEV ajustada (abajo a la izquierda), y gráfico de niveles de retorno con intervalos de confianza del 95 % mediante una aproximación asintóticamente normal (abajo a la derecha).

9. INFERENCIA BAYESIANA PARA LA DISTRIBUCIÓN GEV

Suponga que se tiene una muestra aleatoria $Z_1, \dots, Z_m \stackrel{iid}{\sim} \text{GEV}(\mu, \sigma, \xi)$. Con el objetivo de aplicar métodos MCMC, como menciona [Coles \(2001\)](#), es común reparametrizar el parámetro de escala para mantener su positividad $\eta = \log \sigma$. Además, en ausencia de información para cualquier priors de los parámetros, adoptamos un enfoque ingenuo y utilizamos principalmente priors no informativos e independientes para estos, es decir,

$$\pi(\boldsymbol{\theta}) = \pi(\mu, \eta, \xi) = \pi(\mu)\pi(\eta)\pi(\xi),$$

donde $\mu \sim N(0, \sigma_\mu^2)$, $\eta \sim N(0, \sigma_\eta^2)$ y $\xi \sim N(0, \sigma_\xi^2)$. Observe que los priors puede ser tan planos como se desee al ajustar los parámetros de varianza lo suficientemente grandes. Por lo demás, la elección de densidades normales es arbitraria.

Ahora, para realizar inferencia es necesario establecer algún algoritmo MCMC. En este caso usamos Metropolis-Hastings. Después de establecer valores iniciales para $\boldsymbol{\theta} = (\mu, \eta, \xi)$, utilizamos una regla de probabilidad arbitraria $p(\boldsymbol{\theta}_{i+1}, \boldsymbol{\theta})$ para la simulación iterativa de valores sucesivos en la cadena. Una vez que esta regla se ha utilizado para generar un valor candidato $\boldsymbol{\theta}'$ para $\boldsymbol{\theta}_{i+1}$, aceptamos esto con probabilidad

$$A(\boldsymbol{\theta}', \boldsymbol{\theta}_i) = \min \left\{ 1, \frac{\pi(\boldsymbol{\theta}')f(\boldsymbol{\theta}' | \mathbf{z})p(\boldsymbol{\theta}', \boldsymbol{\theta})}{\pi(\boldsymbol{\theta}_i)f(\boldsymbol{\theta}_i | \mathbf{z})p(\boldsymbol{\theta}_i, \boldsymbol{\theta}')} \right\},$$

de lo contrario $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}$. En particular, usaremos Metropolis-Hastings con caminatas aleatorias:

$$\mu' = \mu_i + \varepsilon_\mu,$$

$$\eta' = \eta_i + \varepsilon_\eta,$$

$$\xi' = \xi_i + \varepsilon_\xi,$$

donde $\varepsilon_\mu \sim N(0, \sigma_{\varepsilon_\mu}^2)$, $\varepsilon_\eta \sim N(0, \sigma_{\varepsilon_\eta}^2)$ y $\varepsilon_\xi \sim N(0, \sigma_{\varepsilon_\xi}^2)$.

10. APLICACIÓN I BAJO UN CONTEXTO BAYESIANO

Para ejemplificar lo anterior, se realizará nuevamente la primera aplicación de datos de temperatura máxima invernal desde 1927 hasta 1995 en Port Jervis, Nueva York, pero desde una mirada bayesiana.

Nuevamente el paquete `extRemes` provee una implmentación del método anterior, dónde los parámetros de los priors y de las caminatas aleatorias son fijados vía optimización.

En la Figura 16 se aprecian las series temporales y la densidad a posteriori de los parámetros del modelo. Se consideró una cadena de largo 10^5 con burn-in de 500, $\sigma_\mu^2 = \sigma_\eta^2 = \sigma_\xi^2 = 10^2$ y $\sigma_{\varepsilon_\mu}^2 = \sigma_{\varepsilon_\eta}^2 = \sigma_{\varepsilon_\xi}^2 = 0,01$.

Más aún, la moda de la distribuciones a posteriori es $(\mu_{\text{MO}}, \sigma_{\text{MO}}, \xi_{\text{MO}}) = (15, 11, 2, 97, -0,21)$ lo que concuerda con los valores obtenidos en el Cuadro 1 para los estimadores máximo verosímiles.

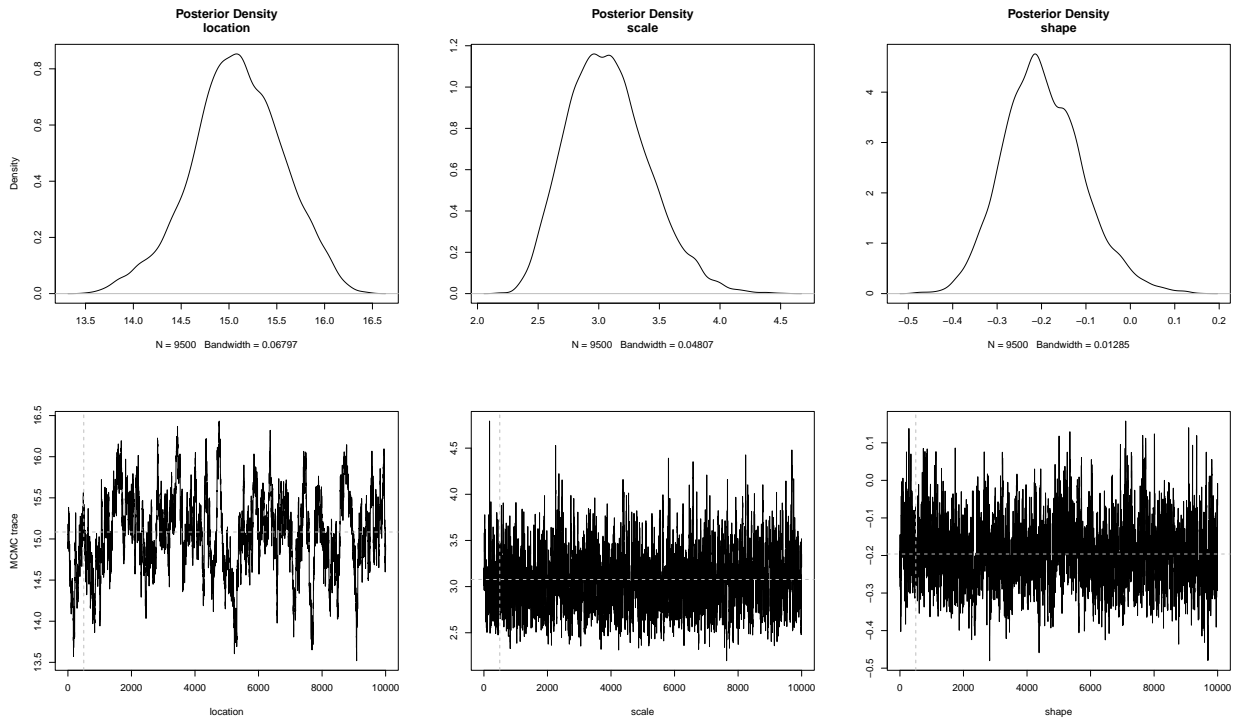


Figura 16. Densidad a posteriori de los parámetros del modelo GEV, junto a las series temporales propuestas por la caminata aleatoria como probabilidad de transición para Metropolis-Hastings.

11. CONCLUSIONES

La investigación realizada permitió abordar la teoría de valores extremos desde diversas aristas, comenzando desde la teoría clásica hasta la aplicación de algoritmos MCMC para inferencia bayesiana. La exploración de la distribución generalizada de valores extremos (GEV) y las técnicas de inferencia tradicionales han sentado las bases para un análisis más profundo.

Junto a ello, la aplicación de estos métodos a conjuntos de datos reales, como la temperatura máxima invernal en Port Davis, la temperatura máxima diaria anual en Nueva York y la temperatura máxima anual en Santiago, Chile, han proporcionado contextualización valiosa de la potencia que este enfoque de modelamiento provee en contextos climáticos particulares. Los diagnósticos han permitido evaluar la idoneidad de la distribución GEV, destacando la importancia de comprender y abordar las limitaciones del modelo.

Finalmente, la incorporación de la inferencia bayesiana mediante el algoritmo de Metropolis-Hastings con caminatas aleatorias, ha enriquecido el análisis al proporcionar una perspectiva diferente a la estadística clásica.

12. MATERIAL SUPLEMENTARIO

Los códigos utilizados a lo largo de este proyecto pueden hallarse libremente en el repositorio público [Extreme Value Theory](#).

Índice de figuras

1.	CDF del máximo de variables aleatorias exponenciales para múltiples valores de n	4
2.	PDF del máximo de variables aleatorias exponenciales para múltiples valores de n	4
3.	CDF del máximo normalizado de variables aleatorias exponenciales para múltiples valores de n	5
4.	Cola derecha de densidad <i>super heavy tail</i> , frente a la densidad normal estándar y Pareto de parámetros $x_m = e, \alpha = 2$	7
5.	Distribución Gumbel para múltiples parámetros	7
6.	Distribución Fréchet para múltiples parámetros.	8
7.	Distribución Weibull para múltiples parámetros.	8
8.	Distribución generalizada de valores extremos para múltiples parámetros. .	9
9.	Gráfico de niveles de retorno para $\xi < 0, \xi = 0$ y $\xi > 1$ considerando $\mu = 0$ y $\sigma = 1$	11
10.	Análisis exploratorio de los datos extremos. Temperatura máxima invernal anual desde 1927 hasta 1995 en Port Jervis, Nueva York (arriba), junto con su distribución (abajo). El histograma ilustra la distribución de frecuencia de las temperaturas máximas invernales anuales a lo largo de los años, ofreciendo perspectivas sobre la variabilidad térmica dentro del período estudiado.	14
11.	Diagnóstico de la función de distribución de valores extremos generalizados (GEV) ajustada a los datos de la temperatura máxima invernal en Port Jervis. Gráfico de probabilidad (arriba a la izquierda), gráfico de cuantiles (arriba a la derecha), gráficos de densidad de los datos empíricos y la función de distribución GEV ajustada (abajo a la izquierda), y gráfico de niveles de retorno con intervalos de confianza del 95 % mediante una aproximación asintóticamente normal (abajo a la derecha).	15
12.	Análisis exploratorio de los datos extremos. Máximo anual de la temperatura promedio diaria en Nueva York entre 1951 y 2017 (arriba), junto con su distribución (abajo). El histograma ilustra la distribución de frecuencia de las temperaturas máximas anuales a lo largo de los años, ofreciendo perspectivas sobre la variabilidad térmica dentro del período estudiado. . .	16
13.	Diagnóstico de la función de distribución de valores extremos generalizados (GEV) ajustada a los datos de la temperatura máxima diaria en Nueva York. Gráfico de probabilidad (arriba a la izquierda), gráfico de cuantiles (arriba a la derecha), gráficos de densidad de los datos empíricos y la función de distribución GEV ajustada (abajo a la izquierda), y gráfico de niveles de retorno con intervalos de confianza del 95 % mediante una aproximación asintóticamente normal (abajo a la derecha).	17
14.	Análisis exploratorio de los datos extremos. Máximo anual de la temperatura promedio diaria en Santiago, Chile entre 1973 y 2023 (arriba), junto con su distribución (abajo). El histograma ilustra la distribución de frecuencia de las temperaturas máximas anuales a lo largo de los años, ofreciendo perspectivas sobre la variabilidad térmica dentro del período estudiado. . .	18

15.	Diagnóstico de la función de distribución de valores extremos generalizados (GEV) ajustada a los datos de la temperatura máxima diaria en Santiago de Chile. Gráfico de probabilidad (arriba a la izquierda), gráfico de cuantiles (arriba a la derecha), gráficos de densidad de los datos empíricos y la función de distribución GEV ajustada (abajo a la izquierda), y gráfico de niveles de retorno con intervalos de confianza del 95 % mediante una aproximación asintóticamente normal (abajo a la derecha).	19
16.	Densidad a posteriori de los parámetros del modelo GEV, junto a las series temporales propuestas por la caminata aleatoria como probabilidad de transición para Metropolis-Hastings.	21

Índice de cuadros

1.	Parámetros estimados de la función de distribución de valores extremos generalizados (GEV) y niveles de retorno ajustados a la temperatura máxima invernal anual (grados centígrados) en los datos de Port Jervis, Nueva York, que se muestran en la Figura 10.	15
2.	Parámetros estimados de la función de distribución de valores extremos generalizados (GEV) y niveles de retorno ajustados a la temperatura máxima anual de la temperatura diaria (grados Fahrenheit) en Nueva York, que se muestran en la Figura 12.	17
3.	Parámetros estimados de la función de distribución de valores extremos generalizados (GEV) y niveles de retorno ajustados a la temperatura máxima anual de la temperatura diaria (grados celsius) en Santiago de Chile, que se muestran en la Figura 14.	19

Bibliografía

1. Fisher, R.A. and Tippett, L.H.C. (1928) *Limiting Forms of the Frequency Distribution of the Largest or Smallest Members of a Sample*. Proceedings of the Cambridge Philosophical Society, 24, 180-190.
2. Coles, S. (2001). *An introduction to statistical modeling of extreme values*. London: Springer-Verlag. ISBN: 1-85233-459-2.
3. Smith, R. (1985) *Statistics of Extreme Values*. International Statistical Institute, Amsterdam.
4. Beirlant, J., Goegebeur, Y., Segers, J. J. J., and Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley.
5. Thompson, D.W.J., Wallace, J.M. 1998. *The Arctic Oscillation signature in the wintertime geopotential height and temperature fields*. Geophysical Research Letter 25:1, 297-1,300.
6. Wettstein, J. J., and Mearns, L. O. (2002). *The influence of the North Atlantic-Arctic oscillation on mean, variance, and extremes of temperature in the northeastern United States and Canada*. Journal Of Climate, 15, 3586-3600.
7. Gilleland, E., and Katz, R. W. (2016). *extRemes 2.0: An Extreme Value Analysis Package in R*. Journal of Statistical Software, 72(8), 1–39.