

CIMAT A.C.

Selección <<automática>> del umbral

Cipriano C. Hdz.

cipriano.callejas@cimat.mx



CIMAT

July 25, 2022

MOTIVACIÓN

Eventos Climáticos Extremos (ECE)

¿Qué son?

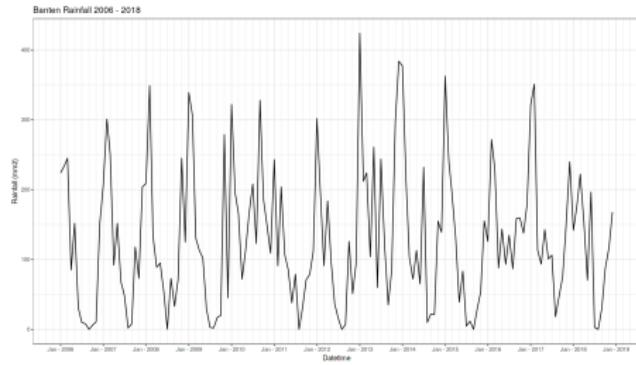
Estos son grandes desviaciones de condiciones o comportamientos promedio (Rypkema and Tuljapurkar 2021)



Figure: Inundación de Hospital en México/ Ola de calor en Paris.

Inundaciones

Las inundaciones pueden ser causadas por eventos extremos en precipitaciones pluviales. **• tener fuertes lluvias por un tiempo prolongado.**



Nos gustaría poder estimar cuándo sucederá un periodo de fuertes lluvias.

Teoría de Extremos

Eventos Extremos

Los eventos extremos (EE) intuitivamente se pueden pensar como aquellos para los cuales cierta variable aleatoria (v.a.) X excede un umbral (**usualmente grande**) u .

Esto es, calcular:

$$\mathbb{P}(X > u + y | X > u) = \frac{\bar{F}(u + y)}{\bar{F}(u)} \text{ para } y > 0. \quad (1)$$

Ejemplo Si $X \sim Exp(1)$ entonces sus eventos extremos:

$$\mathbb{P}(X > u + y | X > u) = \frac{\exp(-u - y)}{\exp(-u)} = e^{-y} \text{ para } y > 0.$$

también se distribuyen de forma exponencial. (**Propiedad de pérdida de memoria**).

F es usualmente desconocida.

Tenemos la siguiente equivalencia asintótica probada por Pickands-Balkema-de Haan.

$$F \in D(H_{\xi,a,b}) \Leftrightarrow \left[\frac{\bar{F}(u+y)}{\bar{F}(u)} \stackrel{u}{\approx} \bar{P}_{\xi,a(u)}. \right] \quad (2)$$

Donde $H_{\xi,a,b}$ denota la función de distribución generalizada de extremos (GEV) y $\bar{P}_{\xi,a(u)}$ es la cola de la distribución de Pareto Generalizada (GPD):

$$\bar{P}_{\xi,\sigma_u}(y) = \left[1 + \frac{\xi(y - u)}{\sigma_u} \right]^{-1/\xi} \quad (3)$$

Método de excesos sobre un umbral

A esta forma de estudiar eventos extremos se le conoce como **método de excesos sobre un umbral (POT)**, cuyo algoritmo es el siguiente:

- ① Probar que $F \in D(H_{\xi,a,b})$
- ② Ajustar la GPD (3) considerando cierto umbral u
- ③ Validar el ajuste (Gráficas de diagnóstico).
- ④ Estimar $\bar{F}(u + y)$ mediante la aproximación (2).

SELECCIÓN <<AUTOMÁTICA>> DEL UMBRAL

Selección del umbral

La elección del umbral u es un problema que (usualmente) se resuelve de dos formas:

- ☞ Una es mediante el conocimiento a priori del fenómeno (**experiencia**) u
- ☞ observando la estabilidad de la estimación de los parámetros $(\hat{\xi}, \hat{\sigma})$ respecto al umbral u , (**Gráficos de residuos promedios**).

Hay mucha incertidumbre en la elección del umbral.

Selección <<automática>>

Es de interés estudiar formas en las que dicha elección tenga un procedimiento menos subjetivo, esto es, algún método <<automático>> de seleccionar el umbral .

Dos puntos importantes a considerar:

- ☞ Un umbral **demasiado pequeño** es probable que se violen las hipótesis teóricas de la aproximación asintótica de Pikands, Bakena y de Haan.
- ☞ Un umbral **demasiado grande** generará pocos excesos (o eventos extremos) y esto conlleva a tener datos insuficientes.

Descripción del algoritmo

- ① Seleccionar un conjunto de umbrales candidatos (equidistantes y diferentes) $\{u_1, u_2, \dots, u_n\}$.
- ② Tomamos $u = u_1$, y aplica la prueba de normalidad **chí cuadrada de Pearson** sobre las diferencias:

$$\tau_{u_j} - \tau_{u_{j-1}} \text{ para } j = 2, \dots, n.$$

donde $\tau_{u_j} = \hat{\sigma}_{u_j} - \hat{\xi}_{u_j} u_j$.

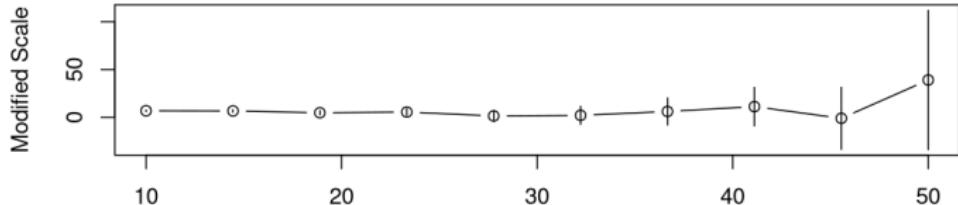
- ☞ Si la hipótesis nula **no es rechazada**, entonces u es considerado un buen candidato a umbral.
- Si la hipótesis nula **es rechazada**, entonces consideramos $u = u_2$, removemos la diferencia $\tau_{u_2} - \tau_{u_1}$, y de nuevo aplicamos la prueba de normalidad a las diferencias restantes.

- ③ El paso 2 es repetido hasta que las diferencias (restantes) efectivamente tengan una distribución normal con media cero.

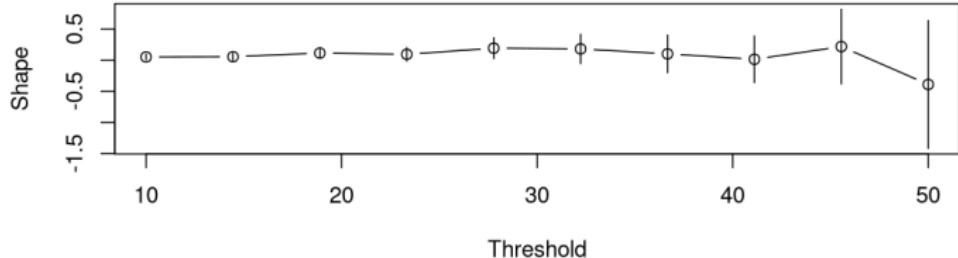
EJEMPLO DE CONTROL

Para ejemplificar este algoritmo, consideremos el siguiente conjunto de datos de Coles et al. 2001, el cual corresponde a la acumulación pluvial reportada diariamente de 1914 a 1962 en Inglaterra.

Gráfica de Residuos



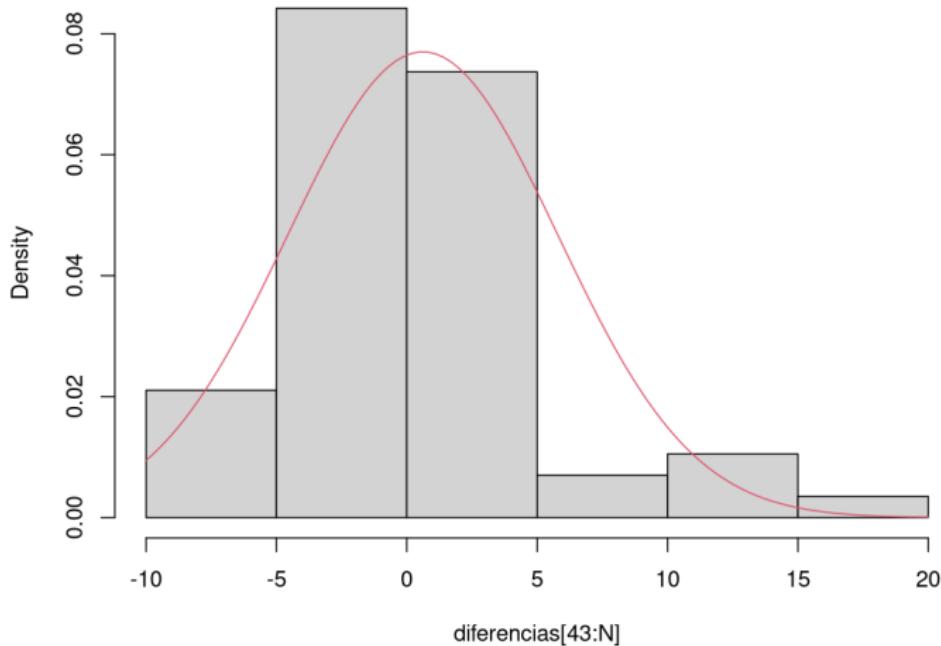
Modified Scale



Shape

Diferencias

Histogram of diferencias[43:N]



Detalles técnicos.

Como umbrales candidatos consideramos una partición equidistante del intervalo $[10, 50]$ de tamaño 100.

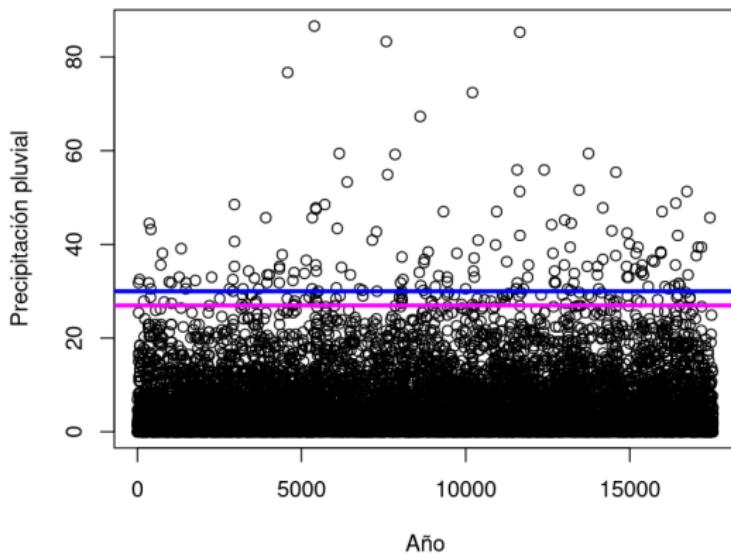
Comparamos los resultados:

	Umbral	ξ
Coles et al. 2001	30	0.1843027
AST	26.9697	0.1302254

Los detalles técnicos del algoritmo pueden ser consultados en Thompson et al. 2009.

Resultados

La línea azul corresponde al umbral sugerido por Coles et al. 2001, mientras que la color magenta usando el algoritmo. Lo cual concuerda con lo obtenido Thompson et al. 2009.



Gráficas de Diagnóstico

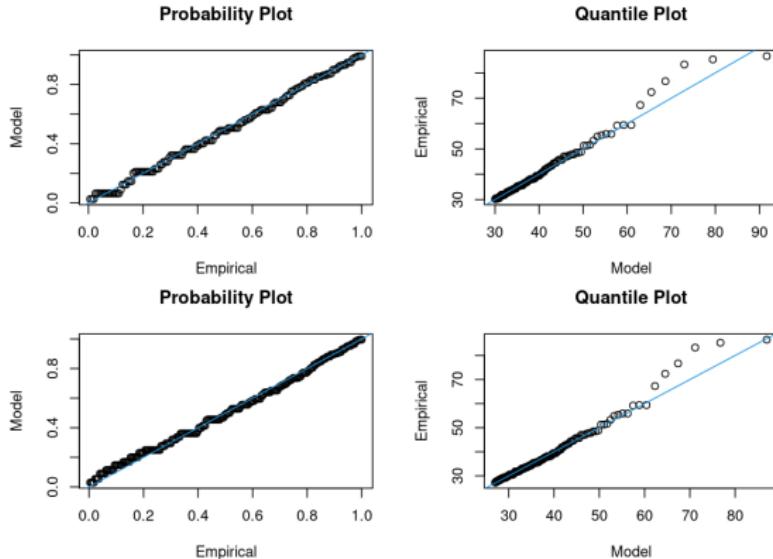


Figure: (Top) Diagnóstico con el umbral elegido **manualmente**.
(Bottom) Gráficas de diagnóstico con el umbral **obtenido automáticamente**.

APLICACIÓN

Motivación

La elección de estos datos es basado en el trabajo de Guermah and Rassoul 2020, para la región de **Khemis-Miliana en Algeria**, puesto que ha padecido cambios climáticos críticos en las últimas décadas, entre ellos inundaciones.



Figure: Bab El Oued (2001) / Ghardia (2008)

Nuestros datos

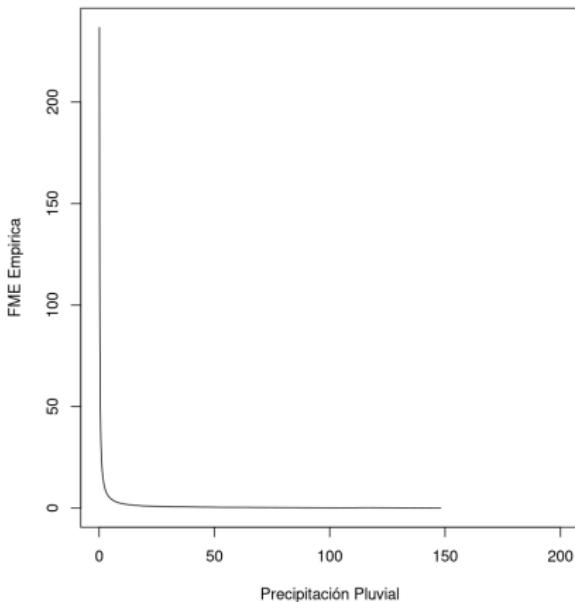
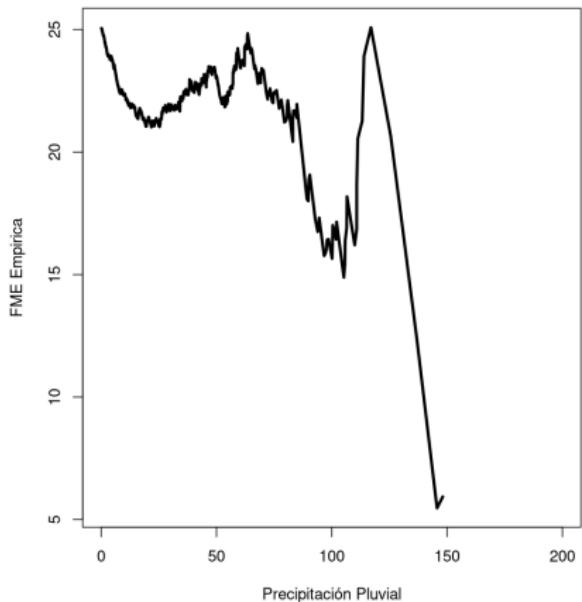
116 años de precipitaciones pluviales en Pakistán (Kaggle).



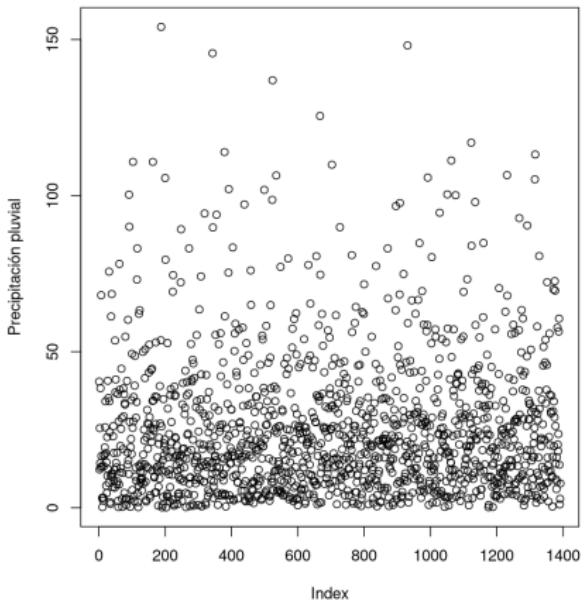
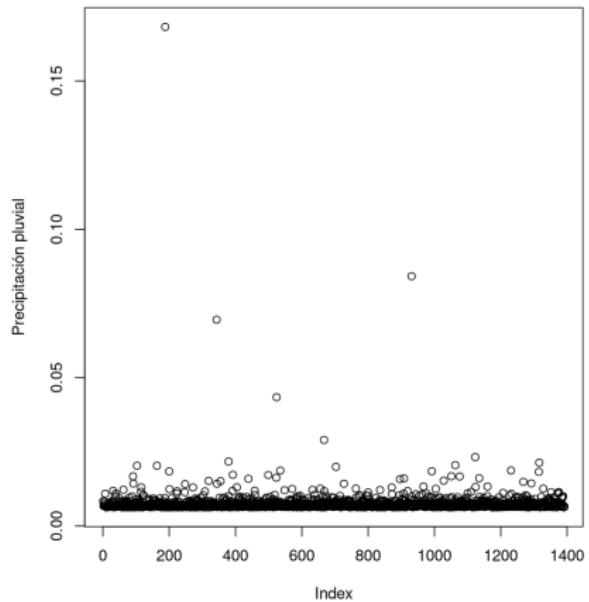
Figure: Inundación en el 2010 Pakistán por fuertes lluvias.

Identificación del dominio de atracción

☞ Evidencia de que la distribución vive en el dominio Gumbel.

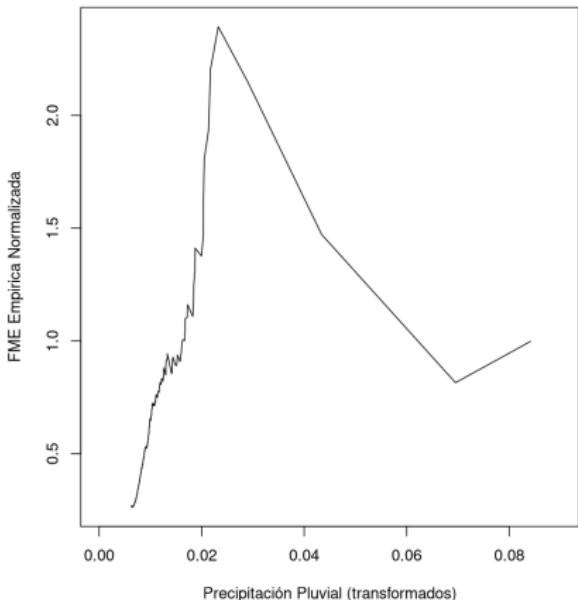
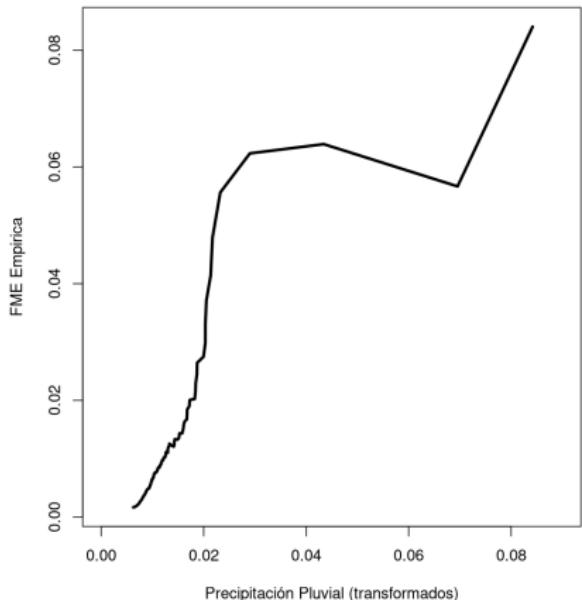


Datos transformados



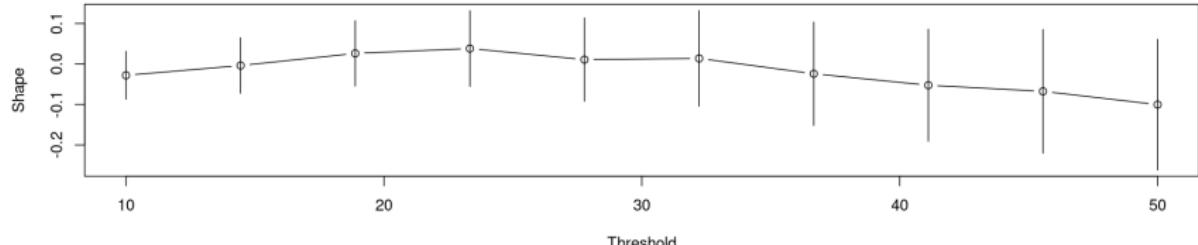
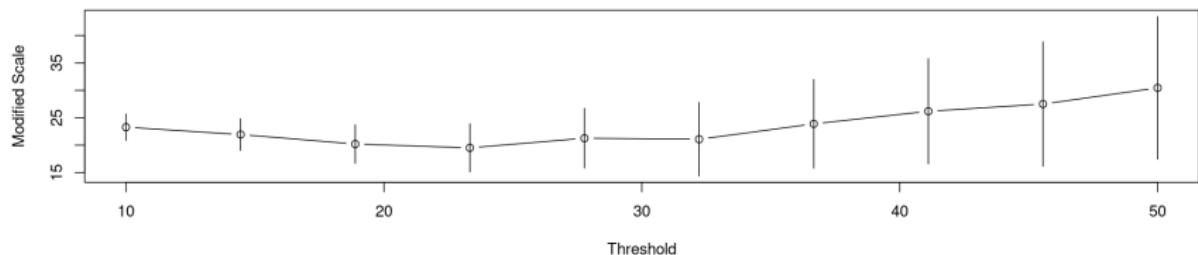
No evidencia en contra del dominio Weibull

☞ Evidencia de que los datos viven en el dominio Fréchet.



Diagra de Residuos

Como lo sugiere Coles et al. 2001, consideramos un rango grande.



Resultados

La línea azul corresponde al umbral obtenido manualmente, mientras que la color magenta fue seleccionado <<automáticamente>>.

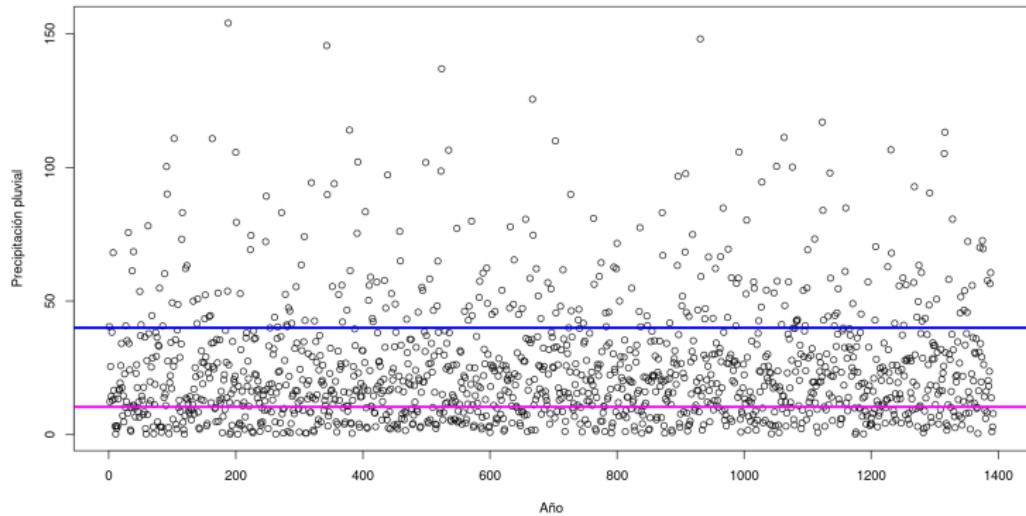


Figure: (man) $\hat{\xi} = -0.02817635$, (alg) $\hat{\xi} = -0.02598995$

Gráficas de Diagnóstico

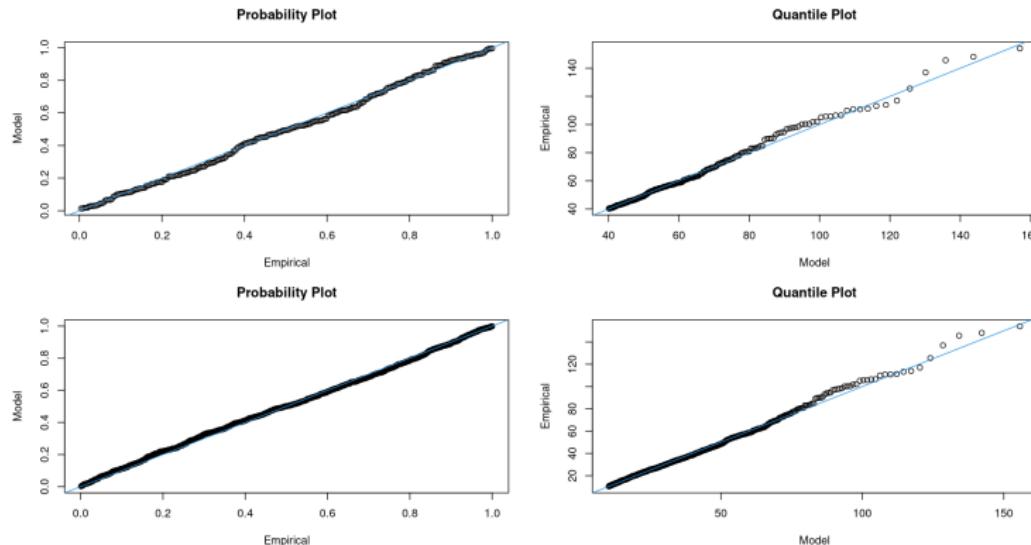


Figure: (Top) Diagnóstico con el umbral elegido manualmente.
(Bottom) Gráficas de diagnóstico con el umbral obtenido automáticamente.

CONCLUSIONES

- ☞ (Objetivo TE) Desarrollar procedimientos (con evidencia estadística) que ayuden a predecir el comportamiento extremo de fenómeno de este estilo, una primera extensión sería estudiar los periodos de retorno.
- ☞ El algoritmo ciertamente redujo la subjetividad de la elección (usual) del umbral. No obstante, aún hay gran dependencia en cuanto a la elección de los umbrales candidatos.
- ☞ En Orsini et al. 2020 estudiaron los accidentes automovilísticos mediante el método de POT usando el algoritmo descrito aquí.

Bibliografía

- Stuart Coles et al. (2001). *An introduction to statistical modeling of extreme values*. Vol. 208. Springer.
- Toufik Guermah and Abdelaziz Rassoul (2020). "Study of extreme rainfalls using extreme value theory (case study: Khemis-Miliana region, Algeria)". In: *Communications in Statistics: Case Studies, Data Analysis and Applications* 6.3, pp. 364–379.
- Federico Orsini et al. (2020). "Large-scale road safety evaluation using extreme value theory". In: *IET Intelligent Transport Systems* 14.9, pp. 1004–1012.
- Diana Rypkema and Shripad Tuljapurkar (2021). "Modeling extreme climatic events using the generalized extreme value (GEV) distribution". In:
- Paul Thompson et al. (2009). "Automated threshold selection methods for extreme wave analysis". In: *Coastal Engineering* 56.10, pp. 1013–1021.