Ingeniería Hidráulica en México. Vol. XII, Núm. 2, págs. 57-62, mayo-agosto de 1997

Nota técnica

Sobre el uso del dominio de atracción para la identificación de distribuciones de valores extremos para máximos

José A. Raynal Villaseñor

Universidad de las Américas-Puebla

Resumen

Usualmente, las distribuciones de valores extremos tipos I (VE-I), II (VE-II) y III (VE-II), (Gumbel, Fréchet y Weibull, respectivamente) se identifican por medio del uso de la distribución General de Valores Extremos (GVE), lo que simboliza a los tres tipos, aún cuando el I sólo puede ser representado en el momento en que se toma el límite, en caso de que $\beta \rightarrow 0$. Existen diversos procedimientos que se han propuesto en la literatura técnica para llevar a cabo la identificación de distribuciones de valores extremos sin estimar los parámetros de la distribución, por ejemplo, los criterios de Tiago de Oliveira, Van Monfort, Hosking y Hosking, Wood y Wallis. Se presenta un procedimiento basado en el Método de la Curvatura, propuesto por Castillo para la identificación de distribuciones de valores extremos para máximos, basado en muestras de datos de valores extremos. Se incluyen varios ejemplos de aplicación que ilustran el procedimiento.

Palabras Clave: valores extremos, probabilidad, dominio de atracción, gastos máximos, identificación de distribuciones, nivel de significancia, periodo de retorno

Introducción

Desde que las distribuciones de valores extremos fueron propuestas por Fréchet (1927) y Fisher-Tippett (1928), surgió el problema de identificar el tipo de distribución más adecuado para una muestra de datos en particular.

En la literatura técnica se han propuesto varios procedimientos para llevar a cabo el proceso de identificación de distribuciones de valores extremos. A continuación se presentan algunos de los procedimientos propuestos, que no requieren estimar los parámetros de la distribución de valores extremos, Raynal Villaseñor y Pérez Durán 1986).

Criterio de Tiago de Oliveira (1981)

Es una prueba del tipo de las de multiplicadores de Lagrange y consiste en la evaluación de un estadístico V_n y su comparación con otro estadístico de referencia T. El criterio de decisión es, si $|V_n| < T$ la distribución de valores extremos es tipo I, si $V_n \geq T$ es tipo III y si $V_n \geq T$ es tipo II. Los estadísticos anteriores pueden ser estimados como:

$$V_n = \sum_{i=1}^{N} V_i/(2.09797N)^{1/2} \qquad (1)$$

donde

$$V_i = [1-\exp(-Z_i)]\frac{Z_i^2}{2} - Z_i \qquad (2)$$

y:

$$Z_i = \frac{(x_i - \hat{x}_0)}{\hat{\alpha}} \qquad (3)$$

x_0 y α son los estimadores de máxima verosimilitud de los parámetros de la distribución de valores extremos tipo I. El estadístico T se calcula por medio de:

$$T = \sqrt{2Ln(N)} - \left[\frac{Ln(Ln(N)) + Ln(4\pi)}{2\sqrt{2Ln(N)}}\right] \qquad (4)$$

La inconveniencia de esta prueba es que Tiago de Oliveira (1981) recomienda se use para tamaños de muestra mayores a 400 valores y que el parámetro de forma esté en el intervalo $-1 < \beta < 0.25$. Más detalles del método se dan en el artículo referenciado.

Criterio de Van Monfort y Otten (1978)

La prueba consiste en calcular un estadístico A y compararlo con su valor crítico. Si A calculado es positivo y es menor o igual que el valor crítico, entonces la distribución de valores extremos es tipo I, en caso contrario es tipo III. Si A calculado es negativo y es mayor o igual que el valor crítico, entonces la distribución de valores extremos es tipo I, en caso contrario es tipo II. El estadístico A se obtiene de la manera siguiente:

$$A = \left[\left(\frac{\sum_{i=2}^{N} l_i \Delta_i}{\sum_{i=2}^{N} l_i} \right) - \bar{\Delta} \right] / (\sigma_{\Delta^2}/N)^{1/2} \qquad (5)$$

donde:

$$l_i = \frac{(x_i - x_{i-1})}{(m_i - m_{i-1})} \qquad (6)$$

para $i = 2, \ldots, N$

$$m_i = -Ln\left[-Ln\left(\frac{i}{N+1}\right)\right] \qquad (7)$$

para $i = 1, \ldots, N$

$$\sigma_{\Delta^2} = \sum_{i=2}^{N} (\Delta_i - \bar{\Delta})^2/(N-1) \qquad (8)$$

$$\bar{\Delta} = \sum_{i=2}^{N} \Delta_i/(N-1) \qquad (9)$$

$$\Delta_i = Ln\left[-Ln\left(\frac{i-0.5}{N+1}\right)\right] \qquad (10)$$

para $i = 2, \ldots, N$

Más detalles del método y la tabla de valores críticos se dan en el artículo referenciado.

Criterio de Hosking, Wallis y Wood (1985)

Esta prueba consiste en calcular un estadístico Z y compararlo con su valor crítico y de aquí se establece el tipo de distribución de valores extremos. El estadístico Z tiene la forma:

$$Z = \beta \left(\frac{N}{0.5633}\right)^{1/2} \qquad (11)$$

donde

$$\beta = 7.859c + 29554c^2 \qquad (12)$$

$$c = \left(\frac{2b_1 - b_0}{3b_2 - b_0}\right) - \frac{Ln(2)}{Ln(3)} \qquad (13)$$

$$b_0 = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad (14)$$

$$b_1 = \frac{1}{N(N-1)} \sum_{i=1}^{N} (i-1)\, x_i \qquad (15)$$

$$b_2 = \frac{1}{N(N-1)(N-2)} \sum_{i=1}^{N} (i-1)(i-2)\, x_i \qquad (16)$$

En un análisis previo, Raynal Villaseñor y Pérez Durán (1986), concluyeron que "ninguno de los tres criterios analizados es muy eficiente en detectar el tipo correcto de distribución de valores extremos para tamaños de muestra menores de 50" y que "los tres criterios son mucho más efectivos cuando la distribución es verdaderamente de tipo I o Gumbel, mostrando debilidad en detectar los otros dos tipos".

En el artículo propuesto, se presenta un procedimiento que se fundamenta en el Método de la Curvatura, propuesto por Castillo (1988) para la identificación de distribuciones de valores extremos para máximos, basado en muestras de datos de valores extremos que representan una excelente opción para el proceso de identificación de distribuciones de valores extremos para máximos sin estimar los parámetros de la distribución.

Identificación de distribuciones de valores extremos: método de la curvatura

El método desarrollado por Castillo y Galambos (1986), Castillo (1988) finalmente fue mejorado por Castillo (1989). El punto clave en este método es medir la curvatura en la cola de la distribución que interesa, por medio de la relación entre las pendientes promedio en dos zonas vecinas de la cola de la distribución. Para calcular las pendientes medias en la zona de la cola de la distribución, el método de la curvatura propone ajustar dos líneas rectas por medio del método de mínimos cuadrados, y usar, como medida de la curvatura, el cociente de las pendientes de esas dos líneas.

La cola derecha de la distribución utiliza el estadístico, Castillo (1993):

$$S = \frac{S_{n1,\ n2}}{S_{n3,\ n4}} \tag{17}$$

donde $S_{i,j}$ son las pendientes inversas de las líneas de mínimos cuadrados ajustadas en papel de probabilidad Gumbel para máximos a los estadísticos de orden $i \leq r \leq j$. De aquí se tiene que, Castillo (1993):

$$S_{ni,\ nj} = \frac{m\Phi_{11} - \Phi_{10}\Phi_{01}}{m\Phi_{20} - \Phi_{10}^2} \tag{18}$$

donde:

$$m = n_j - n_i + 1 \tag{19}$$

y:

$$\Phi_{01} = \sum_{k=ni}^{nj} X_k \tag{20}$$

$$\Phi_{10} = \sum_{k=ni}^{nj} - Ln\left[-Ln\left(\frac{k-0.5}{N}\right)\right] \tag{21}$$

$$\Phi_{20} = \sum_{k=ni}^{nj} \left\{- Ln\left[-Ln\left(\frac{k-0.5}{N}\right)\right]\right\}^2 \tag{22}$$

$$\Phi_{11} = \sum_{k=ni}^{nj} - X_k\left[-Ln\left[-Ln\left(\frac{k-0.5}{N}\right)\right]\right] \tag{23}$$

donde N es el tamaño de muestra y:

$$n_1 = N - [2\sqrt{N}] + 1 \tag{24}$$

$$n_2 = n_3 = N - \left[\frac{(2\sqrt{N})}{2}\right] + 1 \tag{25}$$

$$n_4 = N \tag{26}$$

donde [.] es tan sólo la parte entera en (24) y (25).

Los valores de n_1, n_2, n_3 y n_4 están justificados dado que es un problema de máximos el que se pretende resolver (cola derecha de la distribución). Hay que notar que n_1, n_2, n_3 y n_4 definen el límite de dos zonas de la cola de la distribución donde las líneas de mínimos cuadrados son ajustadas.

Debe notarse también que cuando el valor del estadístico S sea cercano a la unidad se tendrá una cola de la distribución que es una línea recta y los valores que están por arriba o por abajo de este valor, dependiendo de la curvatura de la cola de la distribución, corresponderán para valores grandes o valores pequeños, a los dominios de atracción del tipo Weibull o Fréchet, respectivamente.

En el cuadro 1 (Castillo, 1993), se dan los valores críticos y sus niveles de significancia para diferentes tamaños de muestra, cuando se prueban los dominios de atracción siguientes: Gumbel contra Weibull y Gumbel contra Fréchet.

1. Valores críticos y niveles de significancia asociados con las pruebas Gumbel contra Weibull y Gumbel contra Fréchet (Castillo, 1993)

Tamaño de muestra	Nivel de significancia	Valores Weibull	Críticos Fréchet
10	0.01	15.723	0.120
10	0.02	11.211	0.155
10	0.05	6.477	0.238
10	0.10	4.333	0.340
10	0.20	2.643	0.497
10	0.50	1.135	1.135
20	0.01	9.363	0.159
20	0.02	6.739	0.209
20	0.05	4.770	0.294
20	0.10	3.401	0.387
20	0.20	2.302	0.543
20	0.50	1.133	1.133
40	0.01	5.814	0.216
40	0.02	4.673	0.264
40	0.05	3.392	0.362
40	0.10	2.662	0.465
40	0.20	1.955	0.622
40	0.50	1.083	1.083
60	0.01	4.990	0.265
60	0.02	4.045	0.303
60	0.05	3.090	0.396
60	0.10	2.427	0.485
60	020	1.836	0.651
60	0.50	1.085	1.085
80	0.01	4.673	0.285
80	0.02	3.720	0.336
80	0.05	2.877	0.418
80	0.10	2.264	0.518
80	0.20	1.758	0.658
80	0.50	1.073	1.073
100	0.01	3.613	0.297
100	0.02	3.153	0.345
100	0.05	2.588	0.433
100	0.10	2.135	0.533
100	0.20	1.657	0.678
100	0.50	1.055	1.055
200	0.01	2.969	0.360
200	0.02	2.732	0.410
200	0.05	2.250	0.506
200	0.10	1.895	0.592
200	0.20	1.530	0.713
200	0.50	1.046	1.046

Ejemplos de aplicación

Con base en la capacidad de análisis de valores extremos anuales que tiene el paquete FLODRO 2.0, Raynal-Villaseñor (1996), y en particular para estimar los parámetros de la distribución GVE y para identificar los tipos de distribuciones de valores extremos por medio de los métodos contenidos en este artículo, se han analizado los registros históricos anuales (cuadro 1) en las siguientes estaciones hidrométricas.

Datos hidrométricos (m^3/s):

(1) Villalba, Chih. (1939-1981)
(2) Jaina, Sin. (1942-1980)
(3) St. Mary's River at Stillwater (1915-1986), Nueva Escocia, Canadá, Kite (1988)
(4) El Orégano, Son. (1941-1981)
(5) Edmonton, North Sakatchewan River, Canadá, (40 valores), Van Monfort (1970)

Datos de lluvia (cm):

(6) Filadelfia, Pensilvania, EUA (40 valores), Castillo (1988)

Duración de periodos de calma (días):

(7) Duración de periodos de calma (40 valores), Castillo (1988)

Los resultados del proceso de identificación propuesto están contenidos en el cuadro 2. En el cuadro 3, se muestran los valores de los parámetros de la distribución general de valores extremos obtenidos por medio del método de máxima verosimilitud para cada estación seleccionada que pueden considerarse como los más acertados en cuanto a identificación del tipo de distribución de valores extremos, ya que al usar la distribución GVE no se supone ningún tipo en particular de antemano, comparados contra el parámetro de forma obtenido por el proceso de estimación de parámetros y que será considerado como el verdadero, el método de Tiago de Oliveira acertó en tres de siete opciones aún cuando este no debe ser usado para muetras menores a 400 datos.

El método de Van Monfort acertó en cuatro de siete opciones, el de Hosking acertó en cuatro de siete opciones, el de dominio de atracción por el método de la curvatura acertó en cuatro de siete opciones, de aquí se desprende que el procedimiento funciona adecuadamente, ya que sólo en el caso de la estación Jaina el procedimiento propuesto falló en identificar la distribución como EV-II en lugar de EV-I, como se confirmó con la fase de estimación de parámetros de la distribución general de valores extremos.

Este caso parece ser típico cuando el valor de parámetro de forma tiene un valor menor a −0.40 como ha sido observado en otras estaciones no reportadas en este artículo.

2. Resumen de resultados del proceso de identificación basado en el método de la curvatura de las estaciones hidrométricas analizadas

Est. Param.	(1)	(2)	(3)	(4)	(5)	(6)	(7)
n_1	31	29	57	30	35	29	29
$n_2 = n_3$	38	35	65	36	42	35	35
n_4	43	40	72	41	47	40	40
S_{n1+n2}	111.28	961.46	69.81	181.66	18.41	7.04	1.60
Φ_{01}	3490.50	10273.00	4927.00	5066.17	528.68	828.36	50.19
Φ_{10}	11.86	10.02	15.97	10.42	12.66	10.22	10.22
Φ_{20}	18.30	14.88	28.92	16.02	20.73	15.45	15.45
Φ_{11}	5253.94	15217.57	8783.90	7633.92	849.38	1213.20	74.13
m_{n1+n2}	8	7	9	7	8	7	7
S_{n3+n4}	464.27	1419.28	194.11	255.26	39.34	29.36	2.76
Φ_{01}	6205.70	18241.00	5522.00	5506.17	683.73	752.45	56.55
Φ_{10}	17.26	16.65	25.00	16.96	17.81	16.80	16.80
Φ_{20}	53.74	50.32	84.09	52.06	56.96	51.19	51.19
Φ_{11}	21053.4	58602.93	18179.19	16139.88	2202.94	2125.05	165.75
m_{n3+n4}	8	7	9	7	8	7	7
S	0.24	0.68	0.36	0.71	0.47	0.24	0.58
N.S.*	0.01	0.24	0.029	0.26	0.09	0.02	0.17
Decisión	VE-II	VE-I	VE-II	VE-I	VE-II	VE-II	VE-I

* N.S. = Nivel de Significancia

3. Valores de los parámetros de la distribución general de valores extremos

Est. Param.	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ubicac.	201.13	653.58	348.81	347.44	34.98	100.65	3.66
Escala	145.70	333.50	107.75	224.64	14.13	15.29	2.54
Forma	-0.3154	-0.5175	-0.0151	0.0633	-0.3569	0.3910	0.1369
Decisión	VE-II	VE-II	VE-II	VE-III	VE-II	VE-III	VE-III

4. Estadísticos de prueba para los criterios de Hosking *et al.*, Tiago de Oliveira y Monfort y Otten

Estación método	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Tiago	7.5080	7.6289	0.3974	-0.0483	3.8042	-1.7654	-0.5825
Van Monfort	-4.8148	-5.5018	-0.7482	-0.0054	-3.6833	2.2992	0.4002
Hosking	-3.4349	-3.7763	-0.2402	0.5523	-2.7714	3.5836	1.2331

5. Decisiones sobre las muestras de datos de los ejemplos de aplicación

Estación método	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Estimac.**	VE-II	VE-II	VE-II	VE-III	VE-II	-VE-III	VE-III
Tiago	VE-II*	VE-II*	VE-I	VE-I	VE-II*	-VE-I	VE-I
Van Monfort	VE-II*	VE-II*	VE-I	VE-I	VE-II*	VE-I	VE-III*
Hosking	VE-II*	VE-II*	VE-I	VE-I	VE-II*	VE-III*	VE-I
Dom. atrac.	VE-II*	VE-I	VE-II*	VE-I	VE-II*	VE-III*	VE-I

* Aciertos con respecto al criterio de estimación ** Criterio que se considera más axacto

En el caso de la estación el Orégano, Sin. la falla en la identificación no es tal ya que la distribución puede ser EV-I o EV-III indistintamente dada la magnitud tan pequeña del parámetro de forma obtenido. Todos los métodos considerados en este artículo presentan fallas de una u otra manera al identificar distribuciones cuyos parámetros de forma están en las cercanías de cero. Ver cuadros 4 y 5.

Conclusiones y Recomendaciones

Se ha presentado el método de la curvatura para la identificación de distribuciones de valores extremos para máximos. De análisis realizados con los demás métodos descritos en este trabajo, Raynal Villaseñor y Pérez Durán (1986), se concluye que la opción presentada es tan buena como las metodologías existentes en la literatura, cuando no se requiere la estimación de los parámetros por medio de la distribución general de valores extremos, es necesario tener en cuenta que si el parámetro de forma fuera menor a –0.40, el método no será tan eficaz para identificar la distribución como en los otros casos. El autor recomienda la metodología presentada para la identificación de distribuciones de valores extremos para máximos con la precaución ya descrita.

Recibido: febrero, 1995
Aprobado: julio, 1996

Agradecimiento

A la Universidad de las Américas-Puebla (UDLA-P). Este trabajo forma parte del proyecto "Aplicación de la Distribución General de Valores Extremos al Análisis de Gastos Máximos y Mínimos", patrocinado por el Instituto de Investigación y Posgrado de la UDLA-P.

Referencias

Castillo, E. y Galambos, J. 1986. *Determining the domain of attraction of an extreme value distribution from a set of data*, Philadelphia: Temple University. 12 p.

Castillo, E. 1988. *Extreme value theory in engineering*. San Diego: Academic Press. 389 p.

Castillo, E.; Galambos, J. y Sarabia, J.M. 1989. The selection of the domain of attraction of an extreme value distribution from a set of data. *Lecture Notes in Statistics*, Philadelphia: Temple University. 51:181-190.

Castillo, E. 1993. *Engineering analysis of extreme value data.* Maryland: National Institute of Standards and Technology. 35 p.

Fisher, R.A. y Tippett, L.H.C. 1928. Limiting forms of the frequency distributions of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society.* Cambridge: Cambridge Philosophical Society. 24:180-190.

Fréchet, M. 1927. Sur la Loi de probabilité de l'ecart maximum. *Annals de la Societé Polonaise de Mathematique*, Cracovia, Polonia. Soc. Polonaise de Math. 6 :93.

Gumbel, E. J. y Goldstein, N. 1964. Analysis of empirical bivariate extremal distributions. *Journal of the American Statistical Association* 59:794-816.

Hosking, J.R.M. 1984. Testing whether the shape parameter is zero in the generalized extreme value distribution. *Biometrika* 71(2):367-374.

Hosking, J.R.M.; Wallis, J.R. y Wood, E.F. 1985. Estimation of the generalized extreme value distribution by the method of probability weighted moments. *Technometrics* 27(3): 251-262.

Kite, G.W. 1988. *Frequency and risk analyses in hydrology.* Colorado: Water Resources Publications. 257 p.

Otten, A. y Van Monfort, M.A.J. 1978. The power of two tests on the type of distribution of extremes. *Journal of Hydrology* 37:195-199.

Raynal Villaseñor, J. A. y Pérez Durán, J. F. 1986. Identificación de distribuciones de valores extremos. *Memorias del XII Congreso de la Academia Nacional de Ingeniería, A. C.*, México, D.F.: Academia Nacional de Ingeniería, A. C. pp. 529-533.

Raynal-Villasenor, J. A. 1996. FLODRO 2.0: A user-friendly computer package for flood and drought frequency analyses. *Proceedings of the North American Water and Environment Congress'96*, New York: ASCE. Aceptado para publicación.

Tiago de Oliveira, J. 1981. Statistical choice of univariate extreme value models. En: *Statistical Distribution in scientific work*, G. Taillie *et al* (eds.), Vol. 6 (pp. 653-670) Holanda: D. Reidel Pub. Co.

Van Monfort, M.A.J. 1970. On testing that the distribution of extremes is of type I when type II is the alternative. *Journal of Hydrology* 11: 421-427.

Abstract

Raynal, J.A. "On the Use of Domain of Attraction to Identify Extreme Value Distributions for Maximum Flows". Hudraulic Engineering in Mexico (in Spanish). Vol XII. Num 2, pages 57-62 May-August, 1997.

The extreme value distributions types I (EV-I), II (EV-II) y III (EV-III), (Gumbel, Fréchet and Weibull, respectively) are more often identified through the use of the general extreme value distribution, which can represent types II and III directly, and distribution type I only when the limit $\beta \rightarrow 0$ is taken. There are several other procedures that have been proposed in the literature to identify the extreme value distributions without estimating their parameters. The criteria has been proposed recently in the literature to address such problems. In this paper, a procedure is presented, based on the Curvature Method proposed by Castillo, to identify the extreme value distributions for the maximum flow, based on a sample of extreme value data. Several examples of application are included to illustrate the proposed procedure.

Keywords: *Extreme values, probability, domain of attraction, floods, identification of distributions, significance level, return period*