ANÁLISIS DE LAS PRUEBAS DE DETECCIÓN DE VALORES ANORMALMENTE EXTREMOS (OUTLIERS) EN SERIES HIDROLÓGICAS

Ricardo A. Smith y Claudia P. Campuzano
Posgrado en Aprovechamiento de los Recursos Hidráulicos
Facultad de Minas, Universidad Nacional de Colombia, Medellín
cpcampuz@andromeda.unalmed.edu.co

RESUMEN

Se presentan un conjunto de pruebas para detección de valores anormalmente extremos (outliers) en series hidrológicas. Se hace un análisis de la potencia de estas pruebas utilizando series sintéticamente generadas con las distribuciones Normal, Gumbel y Weibull, ubicando aleatoria y sistemáticamente los outliers, para diferentes números de outliers, y tanto para outliers por encima como por debajo de la media. Igualmente se presentan algunos resultados utilizando información histórica, tratando de analizar la relación de ocurrencia de outliers con la ocurrencia del fenómeno de El Niño – Oscilación del Sur (ENSO) en sus dos fases (El Niño y La Niña). Se presentan finalmente algunas conclusiones y recomendaciones.

PALABRAS CLAVES: Pruebas estadísticas, Puntos anormalmente extremos, series hidrológicas

ABSTRACT

Several statistical tests for detection of outliers in hydrological series are presented. A power analysis of the tests is made using synthetic generated series. The probability distribution functions used for the power analysis were the Normal, Gumbel and Weibull distributions. Several numbers of outliers were used for both: outliers above and below the mean. Additionally, some results using historical data are also presented, trying to analyze the relationship between outliers occurrence and the occurrence of the ENSO phenomena. Some conclusions and recommendations are finally presented.

KEY WORDS: Statistical tests, outliers, hydrologic time series, power analysis

1. INTRODUCCION

Un punto anormalmente extremo u outlier es una observación alejada anormalmente del comportamiento general de las observaciones en una serie. Los outliers puede ser causados por errores de medida, errores de transcripción, averías de los instrumentos, y problemas de calibración. También pueden indicar mayor variabilidad espacial o temporal que la esperada. Los outliers dan la impresión de ser muestras de una distribución distinta al resto de las observaciones, parecen no ser representativos de la muestra. En los procesos de estimación estadística los outliers a menudo llevan a resultados sesgados y escogencia de distribuciones inadecuadas. Se acostumbra entonces en

hidrología, antes de los procesos de estimación de las distribuciones, detectar si la muestra disponible tiene outliers, y si así es, proceder a la remoción de los mismos.

Diferentes pruebas han sido propuestas para detectar outliers y se encuentran referenciadas en algunos libros de estadística aplicada a la hidrología (Gilbert, 1987; Gibbons, 1994; Kottegoda y Rosso, 1997; Helsel y Hirsch, 1992) y otros en la literatura sobre el tema (Tietjen y Moore, 1972). Todas esas pruebas han sido desarrolladas básicamente para probar la hipótesis nula de que todas las observaciones fueron tomadas de poblaciones igualmente distribuidas. Si no se rechaza la hipótesis nula, entonces se asume que la muestra no tiene outliers. Las pruebas propuestas en general se pueden clasificar en dos grupos:

- Basados en distribuciones
- Basados en cercanía

Las pruebas basadas en distribuciones prueban si la distribución de la muestra con y sin el punto o los puntos supuestamente outliers es la misma. Si es la misma se acepta la hipótesis de que la muestra no tiene outliers. Las pruebas basadas en cercanías prueban si el punto supuestamente outlier esta lo suficientemente alejado de la siguiente observación más cercana a ese punto. Si no lo esta se acepta la hipótesis de que la muestra no tiene outliers.

Las pruebas de detección de outliers comúnmente utilizadas en hidrología se aplican para rangos de tamaños de muestras diferentes, lo cual hace difícil su comparación. Se espera que estas pruebas tengan diferente potencia (en el sentido estadístico) y sería conveniente analizarla para concluir sobre su confiabilidad en aplicaciones hidrológicas.

Se presenta a continuación un análisis de la potencia de las pruebas de detección de outliers comúnmente utilizadas en hidrología basada en generación de series sintéticas. Igualmente se presentan algunas conclusiones y recomendaciones basadas en estos experimentos.

2. PRUEBAS DE DETECCION DE OUTLIERS

Las diferentes pruebas disponibles se aplican para diferentes rangos de tamaño de la muestra. Algunas de ellas solo pueden detectar un outlier en la muestra y otras pueden detectar varios. A continuación se describen brevemente estas pruebas.

2.1. Pruebas basadas en Cercanías

- **Prueba de Dixon**

Esta es una prueba propuesta para un grupo de observaciones con un número pequeño de observaciones. Se recomienda que primero se identifique el grupo de observaciones que se sospecha pueden ser outliers. Si M representa el número de observaciones que se sospecha pueden ser outliers, entonces la prueba de Dixon podría ser aplicada para todos los M outliers. Una vez se identifica un outlier usando la prueba este es removido del grupo de observaciones y se puede proceder a probar si hay otro outlier. Esta prueba puede usarse tanto para detectar outliers en la dirección de los máximos como en la dirección de los mínimos. La metodología para esta prueba puede encontrarse en Salas et al, 1992 y Gibbons, 1994.

- **Pruebas Estadísticas Em y Lm**

Las pruebas de los estadísticos Em y Lm pueden usarse para probar la existencia de *M* outliers en un grupo de observaciones que se asume fue tomado de una población normalmente distribuida con media y varianza desconocidas. Se recomienda primero identificar el grupo de *M* observaciones sospechosas de ser outliers, pudiendo aplicarse la prueba para diferentes valores de *M*. La prueba del estadístico Lm es una prueba de hipótesis de dos lados y se puede usar para identificar outliers por encima o por debajo de la media. La prueba del estadístico Em puede usarse para detectar si las *M* observaciones más extremas (por encima o por debajo de la media) en una muestra son outliers. (Salas et al, 1992 y Tietjen y Moore, 1972).

2.2. Pruebas basadas en Distribuciones

- **Prueba de la Distribución Normal**

Pruebas de normalidad realizadas utilizando grupos de observaciones con outliers en general rechazan la hipótesis de que los datos fueron tomados de una población normalmente distribuida. Una prueba para detectar outliers puede ser entonces probar repetidamente la hipótesis de normalidad cuando las observaciones que se sospechan son outliers se remueven de manera secuencial del grupo de observaciones. Cualquiera de las pruebas de normalidad existentes en la literatura se puede usar con este propósito. (Salas et al, 1992).

- **Prueba Studentized Deviates**

La prueba Studentized Deviates es una prueba de detección de multioutliers. Se asume que el grupo de observaciones se tomó de una población normalmente distribuida. La hipótesis nula es que la distribución de la población de donde se tomaron las observaciones que se sospecha son outliers es la misma distribución normal de la población de donde se obtuvo el resto de las observaciones de la muestra. La prueba requiere que se especifique, antes de realizar la misma, un número superior M de outliers potenciales en el grupo de observaciones. (Salas et al, 1992 y Kottegoda y Rosso, 1997)

- **La Prueba de Roesner**

Esta es una prueba de las llamadas mutioutliers ya que permite detectar hasta diez outliers en un grupo de datos. La prueba es válida para tamaños de muestra mayores o iguales a 25 observaciones. El procedimiento asume que las observaciones se tomaron de una población normalmente distribuida. La prueba trata de evitar el enmascaramiento de un outlier por otro cuando están relativamente cercanos.

Antes de realizar la prueba se necesita especificar un número M de outliers potenciales presentes en el grupo de observaciones. La hipótesis nula es que todo el grupo de observaciones representa una muestra tomada de una distribución normal y la hipótesis alternativa es que hay M outliers, o M - *1* outliers, o....., o *1* outlier (Gilbert, 1987, p .188). La prueba de Roesner es una prueba de hipótesis de dos lados e identifica outliers por encima o por debajo de la media. (Salas et al, 1992; Gilbert, 1987 y Gibbon, 1994).

3. ANÁLISIS DE LA POTENCIA DE LAS PRUEBAS

Para realizar el análisis de la potencia de las pruebas, se generaron 100 series sintéticas con las distribuciones Normal, Weibull y Gumbel de 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 500, 1000 datos. Las series normales se generaron utilizando un método basado en la aproximación del teorema del límite central (Salas, et al, 1992). Las series Weibull se generaron utilizando la ecuación (Salas et al, 1992):

$$X = x_o + \alpha \left[-\ln (1.0 - U) \right]^{1/\beta} \qquad (10)$$

Donde x_o es el parámetro de ubicación, α es el parámetro de escala y β es el parámetro de forma de la distribución de Weibull y U es un número aleatorio uniforme (0,1). Las series Gumbel se generaron utilizando la inversa de la CDF de la distribución Gumbel con parámetros (Salas, et al, 1992).

Una vez generadas las series sintéticas se procedió a incluir los outliers en las mismas. Con este objetivo se introdujeron outliers en esas series de 1 hasta 10 outliers dependiendo del tamaño de las series. Para determinar el tamaño de los outliers se usó la expresión $\mu + \sigma K$, variando el valor de K hasta 10, y utilizando la media y la desviación típica estimadas con los datos de la serie. La ubicación de los outliers dentro de las series se realizó mediante la generación de números uniformes (0,1), de acuerdo con el número de outliers. De acuerdo con las pruebas realizadas se concluyó inicialmente que se acepta la presencia de outliers cuando $K \geq 4$. En este análisis se hicieron extensivas pruebas con las series generadas variando el tamaño de la muestra N, el valor de K, el número de outliers M y la distribución de la serie sintética.

A manera de ejemplo, a continuación se presentan solo algunos resultados, ya que no es posible presentarlos todos acá. Los resultados se presentan en forma de tablas en donde se indica el porcentaje de aciertos de las pruebas. En las tablas que se presentan, R corresponde a la prueba de Roesner, SD a la prueba Studentized Deviates, N a la prueba Normal, D a la prueba Dixon y Em y Lm a las pruebas de los estadísticos Em y Lm. Por ejemplo, en la Tabla 2 se muestra para K=5 en el cuadro de 0 outliers detectados, la columna de Em y para 10 datos un valor de 21. Esto significa que con la prueba del estadístico Em para las series de 10 observaciones no se detectaron outliers en el 21% de los casos y si se detectaron (en el cuadro de 1 outlier detectado) en el 79% de los casos.

En la Tabla 1 se presentan los resultados para series de una distribución Normal con M=1, analizando outliers por encima de la media y K=5 y K=10. La Tabla 2 es para el mismo caso pero para outliers por debajo de la media. La Tabla 3 presenta los resultados para series sacadas de una distribución Gumbel con M=4, analizando outliers por encima de la media y K=6, 7, 9 y 10 (tamaño de los cuatro outliers). En este caso se presentan los resultados solo para las detecciones de 2, 3 y 4 outliers. La Tabla 5 presenta los resultados para series tomadas de la distribución Weibull con M=10, analizando outliers por encima de la media y K=5, 5, 6, 6, 6, 9, 9, 9, 10 y 10 (tamaño de los outliers). En este caso se presentan los resultados solo para las detecciones de 0,1 y 10 outliers.

TABLA 1. Porcentaje de aciertos, distribución Normal, con M=1 y outliers por encima de la media

K=5

	0 outliers detectados						1 outlier detectado					
Datos	**N**	**D**	**Em**	**Lm**	**R**	**SD**	**N**	**D**	**Em**	**Lm**	**R**	**SD**
5	0	43	64	41			100	57	36	59		
10	0	19	21	7			100	81	79	93		
15	0	6	5	1			100	94	95	99		

20	0	1	2	0			100	99	98	100		
25	0	2	1	0	1	0	100	98	99	100	99	100
30	0		0	0	0	0	100		100	100	100	100
35	0		0	0	0	0	100		100	100	100	100
40	0		0	0	0	0	100		100	100	100	100
45	0		0	0	0	0	100		100	100	100	100
50	0		0	0	0	0	100		100	100	100	100
100	0				0	0	100				100	100
500	0				0		100				100	
1000	0				0		100				100	

K=10

	0 outliers detectados						1 outlier detectado					
Datos	**N**	**D**	**Em**	**Lm**	**R**	**SD**	**N**	**D**	**Em**	**Lm**	**R**	**SD**
5	0	0	6	0			100	100	94	100		
10	0	0	0	0			100	100	100	100		
15	0	0	0	0			100	100	100	100		
20	0	0	0	0			100	100	100	100		
25	0	0	0	0	0	0	100	100	100	100	100	100
30	0		0	0	0	0	100		100	100	100	100
35	0		0	0	0	0	100		100	100	100	100
40	0		0	0	0	0	100		100	100	100	100
45	0		0	0	0	0	100		100	100	100	100
50	0		0	0	0	0	100		100	100	100	100
100	0				0	0	100				100	100
500	0				0		100				100	
1000	0				0		100				100	

TABLA 2. Porcentaje de aciertos, distribución Normal, con M=1 y outliers por debajo de la media

K=5

	0 outliers detectados						1 outlier detectado					
Datos	**N**	**D**	**Em**	**LmS**	**R**	**SD**	**N**	**D**	**Em**	**LmS**	**R**	**SD**
5	0	43	64	41			100	57	36	59		
10	0	19	21	7			100	81	79	93		
15	0	6	5	1			100	94	95	99		
20	0	1	2	0			100	99	98	100		
25	0	2	1	0	1	0	100	98	99	100	99	100
30	0		0	0	0	0	100		100	100	100	100
35	0		0	0	0	0	100		100	100	100	100
40	0		0	0	0	0	100		100	100	100	100
45	0		0	0	0	0	100		100	100	100	100
50	0		0	0	0	0	100		100	100	100	100
100	0				0	0	100				100	100
500	0				0		100				100	
1000	0				0		100				100	

K=10

	0 outliers detectados						1 outlier detectado					
Datos	**N**	**D**	**Em**	**LmS**	**R**	**SD**	**N**	**D**	**Em**	**LmS**	**R**	**SD**
5	0	0	6	0			100	100	94	100		
10	0	0	0	0			100	100	100	100		
15	0	0	0	0			100	100	100	100		
20	0	0	0	0			100	100	100	100		

25	0	0	0	0	0	0	100	100	100	100	100	100
30	0		0	0	0	0	100		100	100	100	100
35	0		0	0	0	0	100		100	100	100	100
40	0		0	0	0	0	100		100	100	100	100
45	0		0	0	0	0	100		100	100	100	100
50	0		0	0	0	0	100		100	100	100	100
100	0				0	0	100				100	100
500	0				0		100				100	
1000	0				0		100				100	

TABLA 3. Porcentaje de aciertos, Gumbel, con M=4 y outliers por encima de la media

K = (6,7,9,10)

	2 outliers detectados						3 outliers detectados						4 outliers detectados					
N	**R**	**SD**	**N**	**D**	**Em**	**Lm**	**R**	**SD**	**N**	**D**	**Em**	**Lm**	**R**	**SD**	**N**	**D**	**Em**	**Lm**
10			0	0	0	0			0	0	0	0			0	0	0	0
15			1	0	0	0			0	5	0	0			83	70	42	83
20	0	0	6	0	0	5	0	0	4	5	4	4	100	100	90	77	88	90
25	0	0	2	0	1	2	0	0	5	8	5	5	100	100	92	78	92	92
30	0	0	1		1	1	0	0	2		2	2	0	0	97		97	97
35	0	0	1		1	1	0	0	3		3	3	0	0	96		96	96
40	0	0	0		0	0	0	0	3		3	3	0	0	97		97	97
45	0	0	0		0	0	0	0	2		2	2	0	0	98		98	98
50	0	0	0		0	0	0	0	1		1	1	0	0	99		99	99
100	0	0	0				0	0	1				0	0	99			
500	0	0	0				0	0	0				0	0	100			
1000	0	0	0				0	0	0				100	0	100			

TABLA 4. Porcentaje de aciertos, Weibull, con M=10 y outliers por encima de la media

K=(5,5,6,6,6,9,9,9,10,10)

	0 outliers detectados					1 outlier detectado					10 outliers detectados				
N	**R**	**N**	**D**	**Em**	**Lm**	**R**	**N**	**D**	**Em**	**Lm**	**R**	**N**	**D**	**Em**	**Lm**
15		0	100	100	100		100	0	0	0		0	0	0	0
20		0	100	100	100		100	0	0	0		0	0	0	0
25	0	0	100	100	100	0	100	0	0	0		0	0	0	0
30	100	0		4	0	0	0		0	0	0	100		96	100
35	100	0		0	0	0	0		0	0	0	100		100	100
40	100	0		0	0	0	0		0	0	0	100		100	100
45	100	0		0	0	0	0		0	0	0	100		100	100
50	100	0		0	0	0	0		0	0	0	100		100	100
100	100	0				0	0				0	100			
500	100	0				0	0				0	100			
1000	0	0				0	0				100	100			

A partir de los experimentos realizados, y analizando todos los resultados obtenidos, se pueden hacer las siguientes conclusiones:

a. Para las series generadas con la Distribución Normal, y para las pruebas de outliers por encima de la media se observó:

- Cuando el número de datos es muy pequeño (<10), los resultados de todas las pruebas no son buenos.
- En la prueba de Distribución Normal, mientras más grande la serie, se van volviendo peores los resultados. Es una prueba buena para N < 50.
- La prueba Studentized Deviates es la que mejores resultados da para el rango de aplicación ($25 \leq N \leq 100$ y $M \leq 5$).
- Para M=1, en el esquema de $\mu + \sigma K$, K debe ser ≥ 4 para que las pruebas den resultados satisfactorios.
- Con M=1, mientras menor sea el número de datos en el análisis, mayor debe ser K para que se obtengan resultados satisfactorios. Por ejemplo para:

  $N = 15 \quad K \geq 5$
  $N = 10 \quad K \geq 6$
  $N = 5 \quad K \geq 7$

- Cuando $K \geq 10$ y M =1 todas las pruebas detectan siempre los outliers sin importar el valor de N
- La prueba de la distribución normal, cuando los números bases son normales y M =1, es una prueba muy acertiva, detectando siempre todos los outliers.
- En los análisis para M =1, las pruebas que se van deteriorando a medida que K decrece son las siguientes:

Prueba	K	N a partir del cual se obtienen resultados satisfactorios	Prueba	K	N a partir del cual se obtienen resultados satisfactorios	Prueba	K	N a partir del cual se obtienen resultados satisfactorios
Em	10	5	Lm	10	5	Dixon	10	5
	9	5		9	5		9	5
	8	10		8	5		8	5
	7	10		7	10		7	10
	6	10		6	10		6	10
	5	15		5	10		5	15
	4	35		4	25		4	----

- La prueba Dixon solo debe usarse para:

  $K \geq 5 \quad N \geq 15$
  $6 \leq K \leq 7 \quad N \geq 10$
  $K \geq 8 \quad N \geq 5$

- La prueba Roesner es muy acertiva para $K \geq 5$, identificando siempre el outlier.
- La prueba Studentized Deviates es muy acertiva para $K \geq 5$, identificando siempre los outliers.
- En general, la ubicación del outlier en la serie, no afecta los resultados.
- Para valores de $M \geq 1$, las pruebas de Roesner, Em y Lm, presentan mejores resultados cuando la combinación de valores de $K \geq 5$.
- En la prueba de Roesner, a medida que aumenta el número de outliers, el número de datos de la serie debe aumentar para poder detectarlos.

b. Para las series generadas con la Distribución Normal, realizando las pruebas para los outliers por debajo de la media se observó:

- Para este caso las pruebas de Roesner, Studentized Deviates, Dixon, y Lm no dan buenos resultados.
- Las pruebas Normal y Em dan buenos resultados, en todos los casos detectan outliers, aunque a medida que aumenta el número de outliers a detectar, aumenta la exigencia en el número de datos que debe tener la serie.

c. Para las series generadas con la Distribución Gumbel, realizando las pruebas para los outliers por encima de la media se observó:

- La prueba Normal es muy eficiente en el caso de M =1, siempre detecta el outlier sin importar ni N, ni K. Para el caso de $M \geq 2$, da buenos resultados para combinaciones de $K \geq 5$ y $N \geq 40$.
- La prueba de Roesner, en general, no presenta resultados satisfactorios, solo detecta outliers para $N \geq 1000$. Solamente para el caso de M =1 y $K \geq 6$ presenta resultados satisfactorios.
- La prueba de Studentized Deviates no presenta resultados satisfactorios. Al igual que la prueba de Roesner solo para el caso de M =1 y $K \geq 6$ presenta resultados buenos.
- Las pruebas de Em y Lm presentan buenos resultados en el caso de combinaciones de $K \geq 6$ y $N \geq 40$.

d. Para las series generadas con la Distribución Gumbel, realizando las pruebas para los outliers por debajo de la media se observó:

- Las pruebas de Roesner, Studentized Deviates, Dixon y Lm no presentan resultados satisfactorios en todos los casos.
- La prueba Em da buenos resultados para $N \geq 20$.
- La prueba Normal da buenos resultados para $25 \leq N \leq 100$.

e. Para las series generadas con la Distribución Weibulls, realizando las pruebas para los outliers por encima de la media se observó:

- Para M = 1, todas las pruebas detectan el outlier para $K \geq 6$. La prueba Roesner detecta el outlier para $K \geq 5$ y la prueba Normal lo detecta en todos los casos.
- Para $M \geq 2$, las pruebas Roesner y Studentized Deviates no detectan outliers en ninguno de los casos, Roesner tiene una excepción detectando outliers solamente para $N \geq 1000$.
- Las pruebas Em y Lm presentan resultados satisfactorios, sin embargo, a medida que aumenta el número de outliers, disminuye su capacidad de detectarlos para valores de N pequeños.
- La prueba Normal es la que presenta los mejores resultados, detectando outliers en todos los casos.

f. Para las series generadas con la Distribución Weibull, realizando las pruebas para los outliers por debajo de la media se observó:

- Las pruebas Normal y Em dan buenos resultados para valores de $N \geq 25$.
- Las otras pruebas presentan malos resultados en todos los casos.

4. ANÁLISIS DE SERIES HISTÓRICAS

Además de los experimentos anteriores con series sintéticas, las pruebas de outliers se utilizaron para detectar outliers en algunas series históricas de caudales máximos instantáneos en estaciones ubicadas en diferentes Departamentos, y en las cuales se sospecha la presencia de outliers. En la Tabla 5 se presentan los resultados para los diferentes departamentos. En esta Tabla se usa la misma nomenclatura que en las Tablas anteriores para diferenciar las pruebas. Se añade la columna outliers esperados para indicar el número de outliers que visualmente identificaron los analistas. Solo se presentan algunos de los resultados obtenidos a manera de ejemplo.

TABLA 5. Outliers detectados en las series de Caudales Máximos Instantáneos.

ANTIOQUIA

SERIE	Datos	R	SD	N	D	Em	Lm	Outliers
		>=25	20-100	Todos	3-25	3-50	3-50	Esperados
1	12	-	-	1	1	1	1	1
2	21	-	0	1	1	1	1	2
3	16	-	-	1	0	0	0	2
4	11	-	-	8	1	0	0	1
5	18	-	-	1	1	1	1	1

CUNDINAMARCA

SERIE	Datos	R	SD	N	D	Em	Lm	Outliers
		>=25	20-100	Todos	3-25	3-50	3-50	Esperados
1	37	0	0	1	-	0	0	5
2	41	0	0	3	-	3	3	3
3	48	0	0	1	-	0	0	2
4	34	0	0	10	-	10	10	10
5	37	1	0	1	-	1	1	3

RISARALDA

SERIE	Datos	R	SD	N	D	Em	Lm	Outliers
		>=25	20-100	Todos	3-25	3-50	3-50	Esperados
1	23	-	0	1	0	0	0	1
2	7	-	-	1	1	0	1	1
3	49	0	0	1	-	0	0	3
4	15	-	-	1	0	0	0	3
5	24	-	0	1	0	0	0	2

VALLE

SERIE	Datos	R	SD	N	D	Em	Lm	Outliers
		>=25	20-100	Todos	3-25	3-50	3-50	Esperados
1	31	0	0	1	-	0	0	1
2	46	0	0	1	-	0	0	4
3	26	0	0	1	-	0	0	0
4	47	0	0	1	-	0	0	1
5	27	1	1	1	-	1	1	2

6	42	0	0	1	-	0	0	2
7	18	-	-	1	0	0	0	2
8	35	0	0	1	-	0	0	1
9	15	-	-	1	0	0	0	2
10	34	0	4	2	-	2	2	2
11	21	-	0	1	0	0	0	3
12	39	1	1	1	-	1	1	2
13	21	-	0	1	0	0	0	1
14	16	-	-	1	0	0	0	2
15	21	-	1	1	1	1	1	2

Analizando estos resultados no se encontraron diferencias regionales. Aunque las pruebas detectaron outliers, no fueron muy asertivas en la detección del número de outliers esperados.

También se utilizaron las pruebas para detectar outliers en series históricas de caudales, con el fin de determinar si los outliers detectados correspondían a períodos pertenecientes a El Niño o a La Niña. En las Tablas 6 y 7 se presentan los resultados obtenidos para este análisis. Solo se presentan algunas de las series utilizadas a manera de ejemplo.

En estas Tablas se usaron series de caudales medios diarios con registros muy extensos (N≥3000) y los análisis se hicieron para los 100 valores más extremos. En la Tablas que hay a continuación, la columna outliers en año Niño o Niña indica el número de outliers detectados en años Niño o Niña, en la siguiente columna se presenta el mismo resultado en porcentaje, la columna outliers en otro, indica el número de outliers en año no-Niño o no-Niña, en la siguiente columna se presenta el mismo resultado en porcentaje, y la columna posición valor máximo año Niño o Niña, indica la posición del máximo outlier detectado en año Niño o Niña en toda la serie de outliers. Las series que se utilizaron corresponden a estaciones ubicadas en todo el país, las cuales tenían registros de 10 o más años de caudales medios diarios completos (sin información faltante).

TABLA 6. Outliers detectados en las series de caudales históricos para períodos de El Niño.

SERIE	OUTLIERS DETECTADOS	OUTLIERS EN AÑO NIÑO	OUTLIERS EN AÑO NIÑO (%)	OUTLIERS EN OTRO (%)	POSICIÓN VALOR MAX AÑO NIÑO
1	100	70	70	30	1
2	100	37	37	63	4
3	100	28	28	72	10
4	100	15	15	85	29
5	100	9	9	91	41

TABLA 7. Outliers detectados en las series de caudales históricos para períodos de La Niña.

SERIE	OUTLIERS DETECTADOS	OUTLIERS EN AÑO NIÑA	OUTLIERS EN AÑO NIÑA (%)	OUTLIERS EN OTRO (%)	POSICIÓN VALOR MAX AÑO NIÑA
1	100	13	13	87	29
2	100	42	42	58	1
3	100	55	55	45	9
4	100	2	2	98	40

5	100	36	36	64	2

En el análisis de las series de caudales medios diarios históricos se encontró que las series 7 y 8, correspondientes a la estación La Virginia en el departamento de Risaralda y La Pintada en el departamento de Caldas, poseen los porcentajes más altos de determinación de outliers en año Niña, cuando se realizan los análisis por encima de la media, como se mostró en la Tabla 8. Mientras que las series 1, 7 y 11, correspondientes a las estaciones Puente Santander en el departamento del Huila, La Virginia en el departamento de Risaralda y Piedras Blancas en el departamento de Antioquia, poseen los porcentajes más altos de determinación de outliers en año Niño, cuando se realizan los análisis por debajo de la media, como se presentó en la Tabla 6.

Con información encontrada en estudios recientes sobre hidrología Colombiana como el de Rendón (2001), se determinó que estas estaciones pueden responder a la influencia que ejerce la corriente del chorro del Chocó sobre estas zonas, correspondiendo al debilitamiento de la corriente de chorro del occidente Colombiano o corriente de chorro del Chocó, como respuesta a la ocurrencia de la fase cálida del Niño y al aumento de la corriente de chorro del Chocó, como respuesta a la ocurrencia de la fase fría de La Niña. Estos resultados confirman entonces la relación entre la ocurrencia del fenómeno ENSO en sus dos fases y la hidrología colombiana.

5. CONCLUSIONES Y RECOMENDACIONES

De acuerdo con las características de las pruebas presentadas y con los resultados obtenidos se pueden hacer las siguientes observaciones y conclusiones:

a. Existen varias pruebas para detección de outliers con diferentes exigencias de información para su aplicación, y diferentes suposiciones sobre la distribución de donde se asume se tomó la muestra (paramétricas y no paramétricas).
b. En el caso de Colombia las series son de relativa poca longitud y por lo tanto esta situación tiende a favorecer las pruebas que son poco exigentes en el tamaño de la muestra, que generalmente son las pruebas de menor potencia estadística.
c. Las pruebas que mejores resultados arrojaron fueron las de la Distribución Normal y la de Studentized Deviates, seguidas por las pruebas de Roesner y del estadístico Em.
d. En general, las pruebas tienden a dar resultados satisfactorios para $K \geq 6$ (tamaño del outlier), $N \geq 30$ (número de observaciones) y M (número de outliers) pequeño.
e. Se confirmó la relación entre la hidrología colombiana y la ocurrencia del fenómeno ENSO en sus dos fases. Las estaciones en donde más claramente se detectaron los outliers están influenciadas por la corriente del chorro del Chocó, el cual se ve claramente afectado por la ocurrencia del fenómeno ENSO.
f. Las pruebas detectaron outliers, pero no fueron muy asertivas en la detección del número de outliers esperados cuando se usaron las series de caudales medios diarios históricos. Las pruebas que mejores resultados dieron fueron las mismas que en el caso de series sintéticas.
g. Como recomendación se propone el uso extensivo de las pruebas de detección de puntos anormales, de tal manera que se puedan identificar y ser removidos de los análisis posteriores que se hagan con las series. La no remoción de estos puntos significa trabajar con muestras que vienen de poblaciones distintas, lo cual puede tener consecuencias no apropiadas sobre los análisis que se hagan.

6. REFERENCIAS

Gibbon, R.D., 1994. Statistical Methods for Groundwater Monitoring. John Wiley and Sons Inc., New York.

Gilbert, R.O., 1987. Statistical Methods for Environmental Pollution Monitoring. Van Nostrand Reinhold Company, New York.

Helsel, D.R. and Hirsch, R.M., 1992. Statistical Methods in Water Resources. Elsevier, Amsterdam, p. 522.

Kottegoda, N.T. and Rosso, R., 1997. Probability, Statistics and Reliability for Civil and Environmental Engineers. McGraw Hill Book Co., New York.

Rendón, Angela. Influencia de tres corrientes de chorro sobre la hidrología colombiana. Tesis de pregrado. Facultad de Minas. Universidad Nacional de Colombia. Medellín, 2001.

Salas, J; Smith, R; Tabios, G and Heo, J. Statistical computer techniques in water resources and environmental engineering. Department of civil engineering. Colorado State University. January, 1992.

Tietjen, G.L and Moore, R.H, 1972. Some Grubbs-Type statistics for the detection of several outliers. Technometrics, 14(3): 583-597.