Recibido, 19/09/2018 Aceptado, 30/10/2018 DOI:10.25127/aps.20183.404 Artículo original

Modelo de regresión logística para la comparación de series climatológicas registradas en la cuenca del río Torbes, Venezuela

Logistic regression model for comparison on weather series registered in river basin Torbes, Venezuela

Salli Villegas[1], Danny Villegas[1], Yary Pérez[2], Manuel Milla Pino[3*]

RESUMEN

El objetivo de esta investigación fue evaluar series de precipitación mensual mediante regresión logística multinominal para comparar la tendencia, estacionalidad y presencia de observaciones atípicas en series de precipitación mensual. Para ello se utilizaron datos de la estación meteorológica San Cristóbal del estado Táchira y series simuladas mediante modelos de eventos extremos: Pearson tipo III, Gumbel tipo I, Log-Normal y Log-Pearson tipo III. En el análisis de la tendencia y estacionalidad se utilizaron gráficos de saturación de la varianza, para ver observaciones atípicas se utilizó la distancia de Mahalanobis (D^2). Para el ajuste de modelos de eventos extremos se utilizó la estimación de máxima verosimilitud y el ajuste de densidades. Se evidenció una distribución asimétrica de las precipitaciones con una discontinuidad en el periodo 1973-1983, asociada a una alta variabilidad (75,75%) como consecuencia de la presencia de observaciones atípicas causadas por errores en los registros. También, se detectaron observaciones atípicas distribuidas en la época lluviosa, asociadas al mes de agosto de 1960, junio de 1984, julio de 1985 y de 1989. Por otro lado, la precipitación mensual se ajustó a una distribución Pearson tipo III. La regresión logística sugirió que la única variable relacionada con la distribución teórica de la serie fue la precipitación. La simulación de MonteCarlo evidenció consistencia en los estimadores de máxima verosimilitud del modelo logístico en el análisis de la precipitación mensual. Finalmente, los resultados mostraron que las metodologías consideradas son una poderosa herramienta para el estudio de la tendencia y homogeneidad de la precipitación mensual, detección de outliers multivariados y la comparación de series de precipitación mensual, respectivamente.

Palabras claves: Análisis multivariado, regresión no paramétrica, precipitación mensual.

ABSTRACT

The objective of this investigation was to evaluate series of monthly precipitation by means of multinominal logistic regression to compare the trend, seasonality and presence of atypical observations in series of monthly precipitation. For this, data from the San Cristóbal weather station of the Táchira state and series simulated by extreme event models were used: Pearson type III, Gumbel type I, Log-Normal and Log-Pearson type III. In the analysis of the trend and seasonality, saturation graphs of the variance were used. To see atypical observations, Mahalanobis distance (D^2) was used. For the adjustment of extreme event models, maximum likelihood estimation and density adjustment were used. An asymmetrical distribution of rainfall was evidenced with a discontinuity in the period 1973-1983, associated with a high variability (75.75%) as a consequence of the presence of atypical observations caused by errors in the records. Also, atypical observations were detected distributed in the rainy season, associated with the month of August 1960, June 1984, July 1985 and 1989. On the other hand, the monthly precipitation was adjusted to a Pearson type III distribution. The logistic regression suggested that the only variable related to the theoretical distribution of the series was precipitation. The MonteCarlo simulation showed consistency in the maximum likelihood estimators of the logistic model in the monthly precipitation analysis. Finally, the results showed that the methodologies considered are a powerful tool for studying the trend and homogeneity of monthly precipitation, detection of multivariate outliers and the comparison of monthly precipitation series, respectively.

Keywords: Multivariate analysis, nonparametric regression, monthly rainfall.

[1]Universidad Nacional Experimental de los Llanos Occidentales "Ezequiel Zamora", Carrera 3 Calle 17, Guanare, Portuguesa, Venezuela.
[2]Universidad Politécnica Territorial de Portuguesa "J J Montilla", Avenida Simón Bolívar, Guanare, Portuguesa, Venezuela.
[3]Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM-A), Facultad de Ingeniería Civil y Ambiental (FICIAM). Instituto de Investigación en Ingeniería Ambiental (IIIA), Calle Higos Urco N° 342-350-356, Calle Universitaria N° 304, Chachapoyas, Perú
[*]Autor de correspondencia. E-mail: manuel.milla@untrm.edu.pe

I. INTRODUCCIÓN

En Venezuela, y en cualquier parte del mundo, los estudios hidrológicos son fundamentales como fuente de datos para el diseño de obras hidráulicas y para establecer áreas vulnerables ante eventos hidrometeorológicos extremos. Según Sun et al. (2006), las lluvias han sido analizadas desde hace mucho tiempo y los estudios que se han realizado han tenido diversos objetivos. Sin embargo, en la mayoría de ellos el objetivo último es la determinación de los caudales máximos para el diseño de diferentes estructuras hidráulicas. La naturaleza de los eventos hidrológicos es probabilística, y por tanto, para efectuar tales diseños, es necesario plantearse modelos igualmente probabilísticas que representen el comportamiento de esos eventos.

En ese orden, en la realización de muchos estudios hidrológicos con fines de diseño se presentan incongruencias en la información hidrológica empleada. Razón por la cual, los modelos de la regresión logística son modelos estadísticos en los que se desea conocer la relación entre una variable dependiente cualitativa, dicotómica (regresión logística binaria o binomial) o con más de dos valores (regresión logística multinomial). En ese sentido, es importante destacar que la presente investigación se fundamenta en la búsqueda de una herramienta que de manera práctica y sencilla contribuya a mejorar el estudio de las precipitaciones. Por ello, el objetivo central de esta investigación gravita en torno al hecho de evaluar series pluviométricas mediante un modelo de regresión logística multinominal con el fin de caracterizar y comparar la información relacionada con la tendencia y estacionalidad de la precipitación mensual proveniente de estaciones meteorológicas asociadas a la cuenca del Río Torbe del estado Táchira, así como también a modelos de eventos extremos en series sintéticas.

II. MATERIAL Y MÉTODOS

Los datos de esta investigación provienen de una serie de precipitaciones mensuales registradas en la estación San Cristóbal en el periodo 1956-2000. Para la descripción estadística de las variables registradas en la estación se tuvo en consideración el tamaño de la muestra (n), años de registro, la media aritmética, la desviación estándar, la varianza, el coeficiente de variación, mínimo, cuartil (Q1), la mediana, cuartil (Q3), máxima. El análisis exploratorio de los datos (EDA) por medio gráfico se realizó con el fin de comprobar tendencias y cambios en la serie de tiempo por medio visual. Dentro del análisis exploratorio gráfico se utilizó la gráfica de serie de tiempo, el diagrama de cajas, la gráfica de doble masa y la gráfica de normalidad. Para estudiar la homogeneidad de las series de precipitación mensual se utilizó el periodo 1956 al 1991, por ser un periodo homogéneo en la serie objeto de estudio. Se realizó la detección de observaciones atípicas mediante métodos univariados, como la distancia Cook (1977), la cual mide la influencia de una observación mediante el cambio en la región elipsoidal dada en la distancia Di cuando la i-ésima observación es eliminada, así como métodos multivariados, que a menudo indican si las observaciones se encuentran relativamente lejos del centro de la distribución de los datos, específicamente la distancia de Mahalanobis, el cual es un criterio que depende de los parámetros estimados de la distribución multivariada. Se representó gráficamente las precipitaciones mensuales con el fin de ajustar modelos de eventos extremos (Log-Normal, Pearson tipo III, Log-Pearson tipo III y Gumbel tipo I) y se estimaron los parámetros para cada uno de los modelos antes mencionados mediante el método de estimación por máxima verosimilitud. Se realizó un estudio de simulación de Montecarlo con el fin de generar series de tiempo sintéticas basadas en procesos hidrológicos, los cuales son procesos estocásticos estacionarios conocidos todos los momentos de la distribución, específicamente un proceso puramente estacionario o de ruido blanco con base en tres modelos de eventos extremos (Pearson tipo III, Log-Pearson tipo III y Log-Normal). Con base en las series sintéticas de precipitación mensual generadas mediante el estudio de simulación de Montecarlo se realizó un análisis de regresión

logística con el fin de construir funciones logísticas para clasificar series de precipitación mensual provenientes de distancias distribuciones teorices. Los análisis antes mencionados se realizaron con la ayuda del software R 3.3.1.

III. RESULTADOS Y DISCUSIÓN

En la Tabla 1 se muestra una descripción estadística de la serie de precipitaciones mensuales registradas en la estación San Cristóbal en el periodo 1956-2000, las cuales son necesarias para la estimación de los parámetros. Allí se observa una precipitación promedio mensual de 133,3 mm ± 101,01 mm, con precipitaciones mínimas mensuales de 0 mm y máximas mensuales de 829,4 mm. Esta amplitud de valores de precipitación mensual se reflejan en la alta variabilidad presente en toda la serie (75,75%), resultando en características propias de un proceso estocástico, como lo son los procesos hidrológicos, y específicamente las series de precipitación mensual.

Tabla 1. Estadísticas para precipitaciones mensuales registradas en la estación meteorológica San Cristóbal entre 1956-2000

Estadística	Valor
N	552
Mínimo (mm)	0
Máximo (mm)	829,4
Primer cuartil Q1 (mm)	53,7
Mediana (mm)	116,5
Tercer cuartil Q3 (mm)	196,5
Media (mm)	133,3
Varianza (mm ^{2})	10202,28
Desviación estándar (mm)	101,01
Coeficiente de variación (%)	75,75

En la Figura 1 se muestra la distribución de las precipitaciones mensuales registradas en la estación San Cristóbal en el periodo 1956-2000, mediante un histograma y el ajuste de una función de densidad para dicha serie. Estos evidencian una tendencia de los datos de las precipitaciones a distribuirse de forma asimétrica, lo cual coincide con lo reportado en la literatura en relación al tipo de distribución de las precipitaciones (Beckman y Cook, 1993; Smith y Campuzano, 2000)

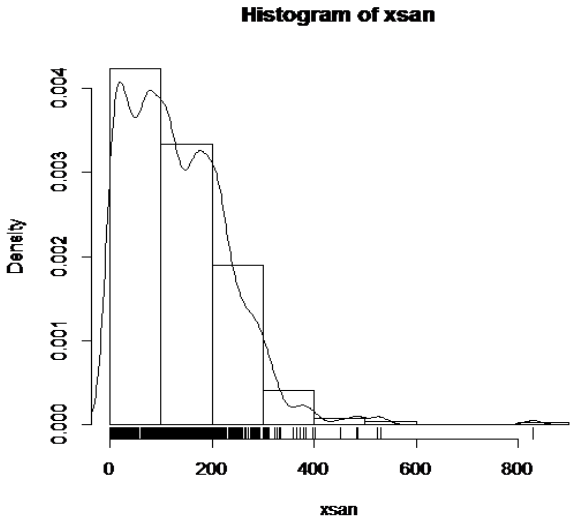

Figura 1. Distribución de las precipitaciones mensuales registradas en la estación San Cristóbal en el periodo 1956-2000.

En la Figura 2 se muestra un gráfico de la saturación de la varianza de precipitaciones mensuales registradas en la estación San Cristóbal en el periodo 1951-2000. Allí se observa que la varianza correspondiente a la serie de precipitación mensual estudiada comienza a descender a partir del año 1960 mostrando una tendencia a estabilizarse en el periodo 1960-2000, por lo que el análisis posterior debe limitarse a dicho periodo.

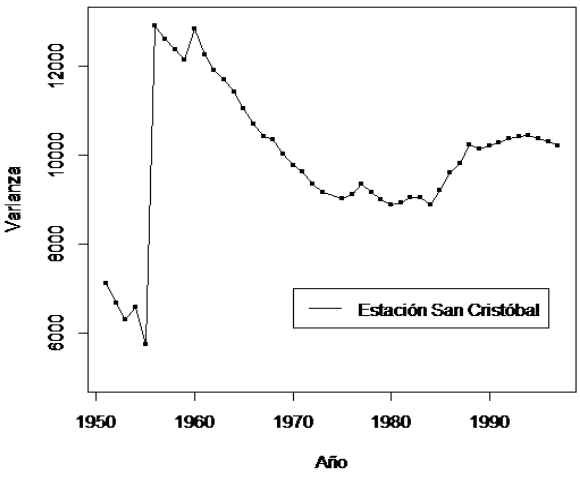

Figura 2. Saturación de la varianza de precipitaciones mensuales registradas en la estación San Cristóbal en el periodo de 1951-2000.

En la Figura 3 se muestran las precipitaciones mensuales registradas en la estación San Cristóbal en el periodo de 1951-2000. Así se observa una discontinuidad en la tendencia de la serie en el periodo 1973-1983, lo que incide en la alta variabilidad de las

precipitaciones mensuales (75,75%), con el subsecuente efecto observado en la saturación de la varianza de las mismas. Este comportamiento en la tendencia de la serie para ese período está asociado a la presencia de observaciones atípicas (outliers), causadas por errores en los registros, para lo cual se recomienda además de la detección de outlier mediante métodos multivariados, y el ajuste y estimación de parámetros mediante metodologías que consideren la presencia de verosimilitud irregulares propias de los procesos hidrológicos y distribuciones probabilísticas asimétricas, como es el caso el caso de los modelos de eventos extremos (Log-Normal, Pearson tipo III, Log-Pearson tipo III, Gumbel tipo I). También se sugiere cortar la serie y eliminar el período que presenta la discontinuidad en la tendencia (1973-1983) con el fin de reconstruir la serie de precipitaciones mensuales en todo el periodo.

En la Figura 4 se muestran los resultados de la detección multivariada de outlier mediante la Distancia de Mahalanobis (D^2) vs Quantiles de la distribución Chi- Cuadrado registradas en la estación San Cristóbal en el periodo de 1956-2000, allí se observan que existen cuatro observaciones que pudieran ser consideradas observaciones atípicas, dado que se alejan considerablemente del centro de masa (centroide o media multivariada), las cuales se describen con detalle en la Tabla 2, donde se observa que estas observaciones atípicas están asociadas a distancias de Mahalanobis relativamente grandes ($D^2 > 10$), y se distribuyen en la época de lluvia, con precipitaciones elevadas ocurridas en agosto del año 1960 (378,7 mm), junio de 1984 (484 mm), julio de 1985 (531 mm) y julio de 1989 (451,8 mm).

En la Tabla 3 se muestran los resultados del ajuste y estimación de parámetros y ajuste de modelos de eventos extremos en series de precipitación mensual de la estación meteorológica San Cristóbal en el período 1956-2000, allí se observa que el test de Kolmogorov-Smirnoff sugiere que el modelo que muestra el mejor ajuste al conjunto de datos de

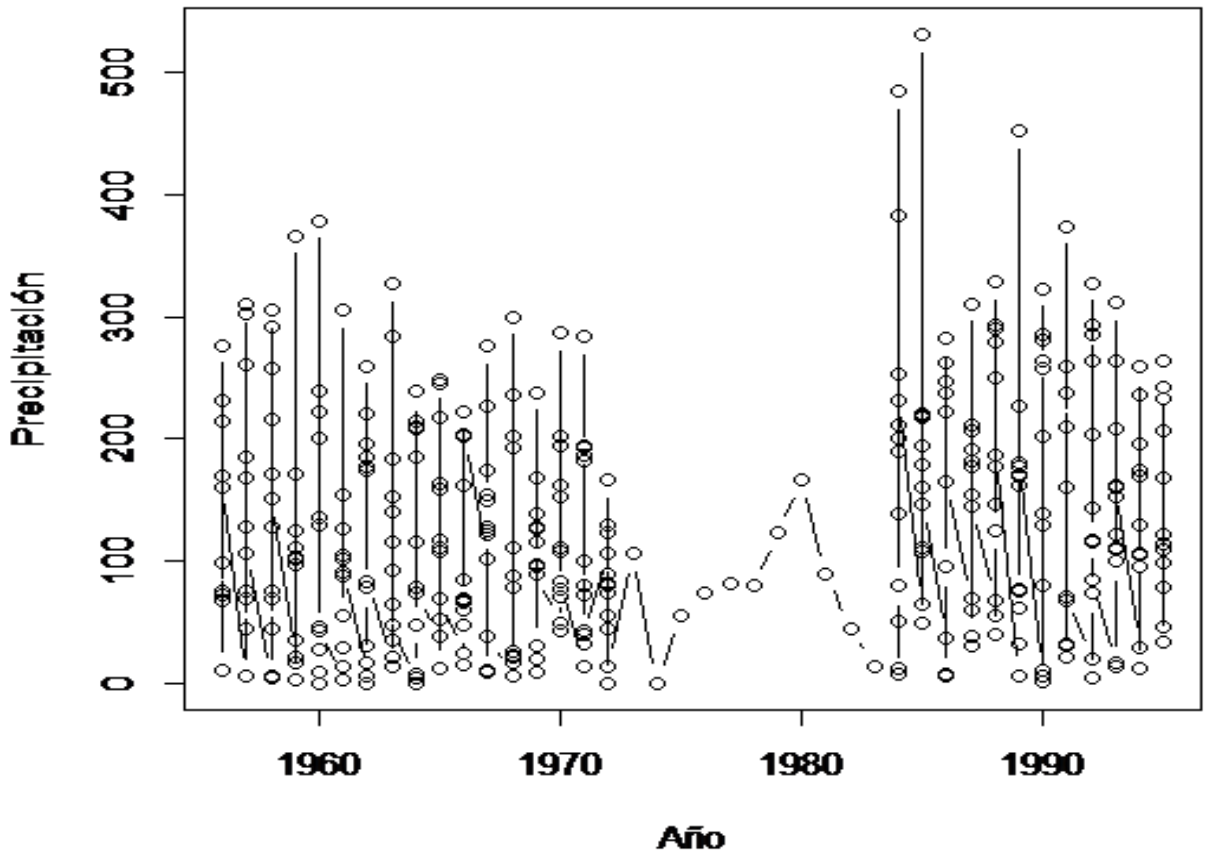

Figura 3. Precipitaciones mensuales registradas en la estación San Cristóbal en el periodo de 1951-2000.

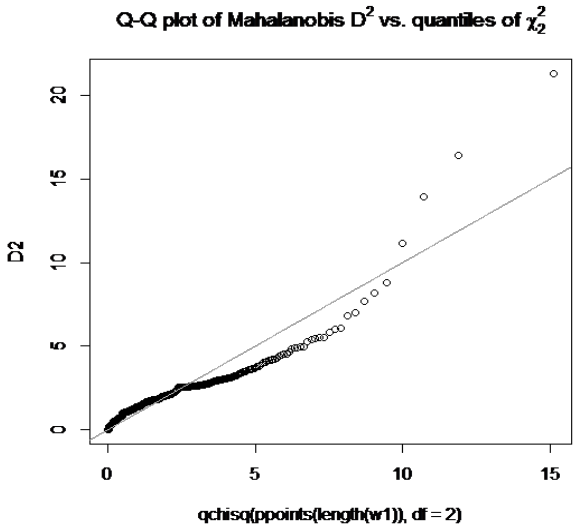

Figura 4. Distancia de Mahalanobis vs Quantiles de la distribución Chi-Cuadrado en series de precipitaciones mensuales registradas en la estación San Cristóbal en el periodo de 1951-2000.

Tabla 2. Observaciones atípicas en series de precipitaciones mensuales registradas en la estación San Cristóbal en el periodo de 1951-2000

Año	Mes	Precipitación (mm)	Distancia de Mahalanobis (D^2)
1960	Agosto	378,7	11,4
1984	Junio	484	16,39
1985	Julio	531	21,29
1989	Julio	451,8	13,91

precipitación mensual es el Pearson tipo III, estos resultados son verificados al observar la Figura 5, donde se muestran las densidades para los cuatro modelos de eventos extremos, allí se observa que el modelo Pearson tipo III es el que mejor se ajusta al histograma de precipitaciones mensuales de la estación San Cristóbal. Estos resultados verifican lo señalado por algunos autores, quienes afirman que el modelo Pearson tipo III o Gamma de tres parámetros

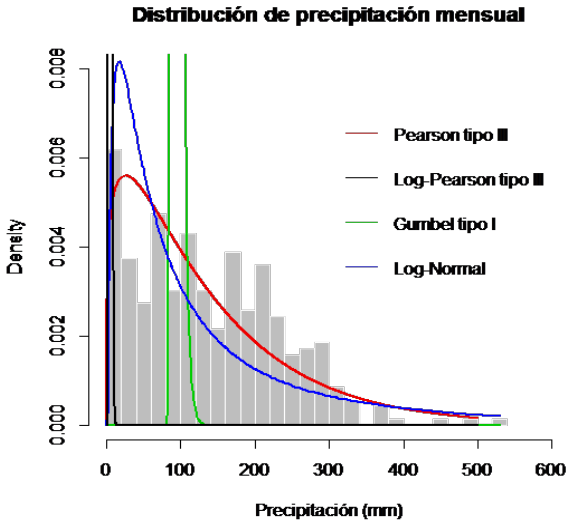

Figura 5. Ajuste de modelos de eventos extremos en una serie de precipitaciones mensuales registradas en la estación meteorológica San Cristóbal en el periodo 1951-2000.

es el que mejor se ajusta a la distribución de las precipitaciones mensuales (Alexanderson, 1986; Martelo, 2004; Searcy et al., 1963).

En la Tabla 4 se muestran los resultados de la regresión logística sobre tres series de precipitaciones mensuales simuladas provenientes de una distribución Pearson tipo III, Log-Pearson y Gumbel con precipitación, temperatura y humedad como regresoras como las que se muestran en la Figura 6, allí se observa que el estadístico Z de Wald para el análisis individual de las variables regresoras sugiere que la única variable relacionada significativamente con el tipo de distribución teórica de la serie es la precipitación.

Tabla 3. Ajuste y estimación por máxima verosimilitud de parámetros de modelos de eventos extremos en series de precipitación mensual de la estación meteorológica San Cristóbal en el período 1956-2000

Modelo	Parámetro estimado	Bondad de ajuste (Test de Kolmogorov -Smirnoff)	
		Estadístico de prueba (D)	Significación (P-valor)
Log-Normal	μx= 4,459 σx= 1,275	0,15533	1,019e -7
Pearson tipo III	α = 1,231 x_O= 0,009	0,089845	0,007263
Log-Pearso n tipo III	α = 10,751 x_O= 2,339	0,19097	1,894e -11
Gumbel tipo I	α = 91,016 β =4,361	0,53869	2,2e -16

Tabla 4. Regresión logística con tres series de precipitaciones mensuales simuladas provenientes de una distribución Pearson tipo III, Log-Pearson y Gumbel con precipitación, temperatura y humedad como regresoras

	Coeficiente	Desv. Estándar coeficientes	Z	P	Odds Ratio	IC 95% L. I.	IC 95% L. S.
Función logística 1 Log-normal / Pearson III							
Intercepto	-0,61638	4,70181	-0,13	0,896			
Precipitación	-0,596177	0,08172	-7,30	0,000	0,55	0,47	0,65
Temperatura	0,107136	0,09822	1,09	0,275	1,11	0,92	1,35
Humedad	0,056729	0,04637	1,22	0,221	1,06	0,97	1,16
Función logística 2 Gumbel I / Pearson III							
Intercepto	0,889606	1,7004	0,52	0,601			
Precipitación	-0,007238	0,001172	-6,18	0,000	0,99	0,99	1,00
Temperatura	-0,000882	0,035366	-0,02	0,980	1,00	0,93	1,07
Humedad	-0,002259	0,017184	-0,13	0,895	1,00	0,96	1,03

L. I.: Límite inferior, L.S.: Límite superior

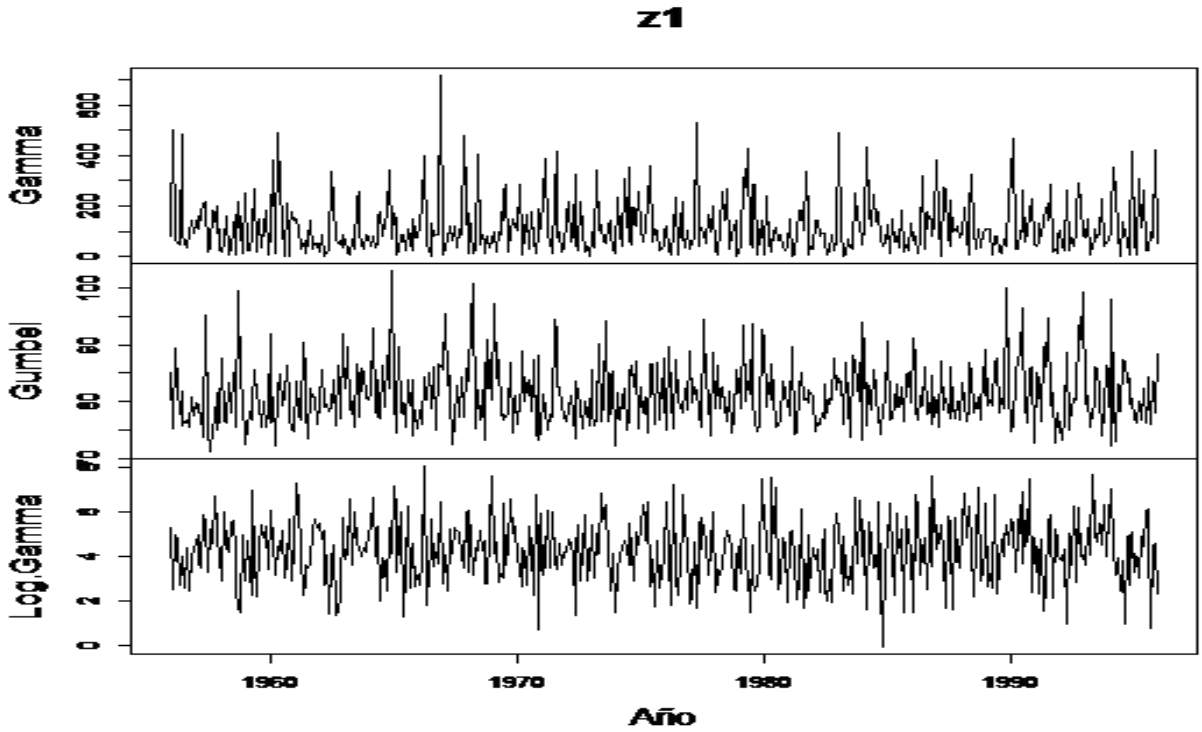

Figura 6. Series de precipitaciones mensuales reconstruida (simulada mediante una distribución Gamma o Pearson tipo III, Log-Pearson y Gumbel) en la estación meteorológica San Cristóbal en el periodo 1956-2000.

El modelo sería:

$$P(\text{Log} - \text{normal/precipitación (mm)}) = \frac{(e^{6,53351 - 0,585475 * \text{precipitación}})}{(1 + e^{6,53351 - 0,585475 * \text{precipitación}})}$$

$$P(\text{Gumbel I/precipitación (mm)}) = \frac{(e^{0,681652 - 0,0072367 * \text{precipitación}})}{(1 + e^{0,681652 - 0,0072367 * \text{precipitación}})}$$

En la Tabla 5 se muestran los resultados de las pruebas de bondad de ajuste de un modelo de regresión logística con tres series de precipitaciones mensuales simuladas provenientes de una distribución Pearson tipo III, Log-Pearson y Gumbel con precipitación, temperatura y humedad como regresoras, allí se observa que los resultados de los estadísticos asociados a la razón de verosimilitud mejoran conforme se incrementa el tamaño de la muestra. Estoevidencia la consistencia de los estimadores de máxima verosimilitud del modelo logístico en el análisis de series de precipitación mensual.

Tabla 5. Bondad de ajuste de un modelo de regresión logística con tres series de precipitaciones mensuales simuladas provenientes de una distribución Pearson tipo III, Log-Pearson y Gumbel con precipitación, temperatura y humedad como regresoras

Muestral (n)	Log-verosimilitud	G	Pvalor
10	-12,806	40,306	0,000
30	-38,558	120,634	0,000
50	-71.89	186,405	0,000
100	-164,551	330,065	0,000
200	-317,966	682,403	0,000
300	-473,993	1029,517	0,000
400	-636,465	1363,739	0,000
480	-761,46	1641,083	0,000

IV. CONCLUSIONES Y RECOMENDACIONES

La regresión logística sobre tres series de precipitaciones mensuales simuladas provenientes de una distribución Pearson tipo III, Log-Pearson y Gumbel con precipitación, temperatura y humedad como regresoras sugirió que la única variable relacionada significativamente con el tipo de distribución teórica de la serie fue la precipitación. Los resultados de la regresión logística con tres series de precipitaciones mensuales simuladas provenientes de una distribución Pearson tipo III, Log-Pearson y Gumbel con precipitación, temperatura y humedad como regresoras, mostraron como los estadísticos asociados a la razón de verosimilitud mejoraron conforme se incrementó el tamaño de la muestra, lo que evidenció la consistencia de los estimadores de máxima verosimilitud del modelo logístico en el análisis de series de precipitación mensual. Finalmente, en virtud de los resultados obtenidos en esta investigación se recomienda considerar las metodologías presentadas, como es el caso de la saturación de la varianza como alternativa para el estudio de la tendencia y homogeneidad en series de precipitación mensual, así como el uso de las distancias de Mahalanobis (D^2) para la detección de outliers multivariados y la regresión logística como una poderosa herramienta para la comparación de series de precipitación mensual.

V. REFERENCIAS BIBLIOGRÁFICAS

Alexanderson, H. 1986. "A Homogeneity Test to Precipitation Data." *International Journal of Climatology* 6 (6): 661–675.

Beckman, R., y R. D. Cook. 1983. "Outlier…….s." *Technometrics* 25 (2): 119-149.

Martelo, M. T. 2004. *Consecuencias Ambientales Generales del Cambio Climático en Venezuela*. Trabajo de ascenso. Universidad Central de Venezuela. Maracay (Venezuela).

Searcy, J. K., y C. H. Hardison. 1960. *Double Mass Curves. Manual of hydrology: Part 1. General Surface Water Techniques*. Washington D. C. (EEUU): Department of Agriculture.

Smith, R., y C. Campuzano. 2000. "Análisis exploratorio para la detección de cambios y tendencias en series hidrológicas." En *XIV Seminario Nacional de Hidráulica e Hidrología*. Bogotá (Colombia).

Sun, Y., S. Solomon, A. Dai, y R. W. Portmann. 2006. "How Often Does It Rain?" *Journal of Climate* 19 (6): 916-934.