Revista de Climatología

Vol. 19 (2019): 67-73 ISSN 1578-8768

Evaluación de observaciones atípicas en datos climatológicos y en modelos lineales simulados

Danny Villegas[1], Yubisay Rivas[1], Yary Pérez[2], Salli Villegas[1] y Manuel Milla[3]

1. Programa de Ciencias del Agro y del Mar. Universidad Nacional Experimental de los Llanos Occidentales Ezequiel Zamora. <danny_villegas1@yahoo.com>

2. Programa Nacional de Formación Agroalimentaria. Universidad Politécnica Territorial de Portuguesa "JJ Montilla".

3. Facultad de Ingeniería Civil. Universidad Nacional de Jaén, Cajamarca, Perú.

(Recibido: 28-Jul-2019. Publicado: 15-Sep-2019)

Resumen

El objetivo de esta investigación fue identificar observaciones atípicas (*outliers*) en datos de precipitación y humedad provenientes de una estación meteorológica en Guanare, Venezuela, y en modelos lineales simulados mediante residuos Studentizados y la distancia de Cook. Se detectaron 3,54 % de *outliers* en la serie de precipitación mensual y 3,15 % en la de humedad relativa. Los años 1981 y 2005 fueron los de mayor pluviosidad, asociada a los meses de mayo, junio y julio. Los años 1988 y 1990 fueron años atípicos con una baja humedad relativa asociada al período de noviembre a marzo. Se verificó la consistencia de los estimadores $(X_{(1)}, X_{(n)})$ de los parámetros de la distribución uniforme. Los residuos studentizados y la distancia de Cook verificaron que el modelo mixto $P(x) = \pi\, P_{D(\theta)}(x) + (1-\pi)\, P_{u(X_{(1)},X_{(n)})}(x)$ fue eficiente para incorporar *outliers* en cada una de las muestras, pero no para incorporar observaciones influyentes más allá de un 6 %.

Palabras clave: *outliers*, series, precipitación, humedad, simulación.

Assessment of outliers in climatological data and simulated linear models

Abstract

The objective of this investigation was to identify outliers in precipitation and humidity data from a meteorological station in Guanare, Venezuela, and in simulated linear models using Studentized residuals and Cook's distance. A 3.54% of outliers were detected in the monthly precipitation series and 3.15% in the relative humidity series. The years 1981 and 2005 were the most rainy, associated with the months of May, June and July. The years 1988 and 1990 were atypical years with a low relative humidity associated with the period from November to March. The consistency of the estimators $(X_{(1)}, X_{(n)})$ of the parameters of the uniform distribution was verified. The studentized residuals and Cook's distance verified that the mixed model $P(x) = \pi\, P_{D(0)}(x) + (1-\pi)\, P_{u(X_{(1)},X_{(n)})}(x)$ was efficient to incorporate outliers in each of the samples, but it was not efficient to incorporate influential observations beyond 6%.

Key words: *Outliers, series, precipitation, humidity, simulation.*

1. Introducción

El clima abarca valores estadísticos sobre elementos del tiempo atmosférico en una región durante un periodo representativo: temperatura, humedad, presión, vientos y precipitaciones. Estos valores se obtienen con la recopilación de forma sistemática y homogénea de la información meteorológica, durante períodos que se consideran suficientemente representativos, de 30 años o más. Generalmente, las series climatológicas presentan datos anormalmente alejados del comportamiento de la serie, los cuales se de-

nominan observaciones atípicas (*outliers*). En ese sentido, un *outlier* es una observación "inconsistente" con el resto de los datos. En ese orden, Chacín (1998) señala que un *outlier* es una observación que tiene un valor extremo en relación a las demás observaciones. Es una particularidad e indica que el dato no es del todo típico respecto a los otros; de allí que su presencia dificulte el ajuste del modelo de regresión. Sus residuales son considerablemente mayores en valor absoluto que los otros. Estos datos atípicos han sido estudiados para explicar razones de su extraño comportamiento. Así mismo, los *outliers* pueden surgir por un variado número de motivos, entre los cuales se pueden señalar; a) medidas defectuosas del análisis; b) registro incorrecto de los datos; c) defectos en los instrumentos de medida; y d) fallas en los supuestos (una de las más comunes). Según Beckman y Cook (1983) los *outliers* pueden ser de distintos tipos: Observaciones discordantes, observaciones contaminantes y observaciones influyentes. Es por ello que, en hidrología se acostumbra, antes de los procesos de estimación de las distribuciones, detectar si la muestra disponible contiene observaciones atípicas, y así proceder a tratar adecuadamente los mismos.

De igual manera, los modelos lineales se usan frecuentemente en diversas áreas del conocimiento, por ejemplo, en el área agronómica, donde por muchos años el estudio de las relaciones entre los cultivos y los factores ambientales ha estado dominado por el empirismo. La predicción del comportamiento de variables como rendimiento, frente a distintas condiciones de suelo, clima, manejo o variedad está también sujeta a interacciones complejas, que desde el punto de vista estadístico presumen la existencia de multicolinealidad y a la presencia de observaciones que tengan una gran influencia sobre un modelo lineal, denominadas *outliers*, cuyo efecto puede observarse en el estimador de mínimos cuadrados ordinarios, en el vector de predicciones y en la matriz de dispersión. En tal sentido, el propósito de esta investigación es evaluar la presencia de observaciones atípicas (*outliers*) en modelos de regresión lineal múltiple cuando existe la presencia de multicolinealidad, además de comparar el efecto que estos puedan tener sobre la estabilidad de los estimadores de mínimos cuadrados ordinarios del modelo y la bondad de ajuste del mismo, para lo cual se utilizaron datos provenientes de un estudio de simulación con base en un modelo con regresores colineales, considerando diferentes tamaños de muestra y distribuciones teóricas continuas de los regresores, así como observaciones atípicas en series climatológicas.

2. Área de estudio

El área de estudio está localizada en el municipio Guanare, Venezuela, específicamente en la estación meteorológica del Ministerio del Poder Popular Para el Ambiente (MPPPA), ubicada en el campus de la Universidad Nacional experimental de los Llanos Occidentales Ezequiel Zamora a una altura de 263 msnm entre las coordenadas geográficas 09°03'54"N y 69°48'23"O, la cual se encuentra en el sector de la Mesa de Cavacas, San Juan de Guanaguanare, al norte del municipio Guanare, estado Portuguesa, Venezuela. El área de estudio se encuentra dentro de la zona de vida de Bosque Seco Tropical (Ewel *et al.*, 1976; Holdridge, 1967). Según análisis de los datos de MPPPA (2009), el área presenta una precipitación media anual de 1769,7 mm y se definen dos periodos bien marcados con respecto a la distribución de las lluvias a lo largo del año, uno lluvioso que va desde abril hasta noviembre y uno seco de diciembre a marzo. La temperatura media anual es de 26,4 °C, el mes de marzo registra las temperaturas medias más altas del año (28,1 °C) y el valor máximo de evaporación (211,1 mm), y el mes de julio las temperaturas medias más bajas (25,3 °C). La humedad relativa tiene un promedio anual de 71 %, varía entre 61 %, en los meses de febrero y marzo y 78 % en los meses de junio y agosto. Geológicamente el área de estudio se ubica sobre las formaciones Río Yuca y Guanapa, sobre materiales recientes (aluviones). Geomorfológicamente esta área se encuentra ubicada sobre el valle del río Guanare, principalmente en la terraza alta nivel superior (Rengel *et al.*, 1983).

Los suelos que predominan en los glacis de las terrazas altas de nivel superior y los vallecitos entre terrazas, son del orden Alfisoles, y se clasifican como AquulticHaplustalfs y Aquultic Paleustalfs, que se caracterizan por tener condiciones ácuicas (saturación y reducción química) en algún tiempo del año, un horizonte argílico que tiene una saturación con bases de menos de 75 % y un 75 % de suelo mineral y arcilloso en los primeros subhorizontes (Larreal *et al.*, 1979).

3. Materiales y Métodos

Se seleccionaron registros de precipitación y humedad de la estación meteorológica Mesa de Cavacas en el periodo (1978-2010) para identificar valores atípicos (*outliers*). De la misma manera, se realizó un análisis exploratorio de los datos; prueba de Anderson-Darling para inferir sobre la simetría de la distribución y verificar el cumplimiento del supuesto de normalidad que requiere el método de detección de *outliers* (Residuos Studentizados). Para la detección de *outliers* se utilizaron los residuos studentizados, los cuales se calculan mediante la siguiente ecuación:

$$t_i = e_i \sqrt{\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}}$$

De esta manera, el residuo studentizado t_i sigue una distribución t de Student y, en este caso, se puede usar un valor crítico basado en la corrección de Bonferroni con $(n-p-1)$ grados de libertad. En ese orden, según Birkes y Dodge (1993), sugieren que un residual studentizado mayor a 2 en valor absoluto puede catalogarse como *outlier*. De igual forma, se utilizó la distancia de Cook (1977) para la detección de *outliers* mediante la ecuación:

$$D_i = \frac{1}{p}\, r_i^2\, \frac{p_{ii}}{1-p_{ii}}$$

siendo r_i el i-ésimo residuo internamente studentizado, donde n representa el número de observaciones, p es el número de predictores del modelo, *SSE* es la suma de cuadrados del error, h_{ii} es el i-ésimo elemento de la diagonal de la matriz de varianza-covarianzas y e_i es el i-ésimo residual. En el resultado estadístico propuesto por Cook, la influencia de una observación es medida por el cambio en el centro de la región elipsoidal dada en D_i cuando la i-ésima observación es eliminada. Así pues, en Cook y Weisberg (1980) se sugiere que cada D_i sea comparada con el percentil de una F con r y $n-r$ grados de libertad; en otras palabras, grandes valores de D_i indican que la observación es influencial. De igual manera, se realizó una Simulación de Monte Carlo (1000 simulaciones) considerando cuatro tamaños de muestra: n=10, n=20, n=30 y n=50, y un modelo lineal, tal como se presenta a continuación: $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e$, con β_0=0, β_1=β_2=β_3=1.

Las variables X_1 y X_2 fueron generadas partiendo de tres distribuciones teóricas continuas (uniforme, normal, y exponencial) y X_3 fue establecida como una combinación lineal del tipo $X_3 = X_1 + 2X_2 + e$. El componente aleatorio de esta combinación fue generado de una distribución normal $\mathcal{N}(0,2)$. Los errores aleatorios del modelo se generaron de una distribución normal $\mathcal{N}(0,1)$, utilizando un modelo contaminado con un porcentaje dado de *outliers*, en este caso el modelo: $P(x) = \pi\, P_{D(\theta)}(x) + (1-\pi)\, P_{u(X_{(1)},X_{(n)})}(x)$. Los valores de contaminación π que se utilizaron son: 0,05, 0,10, 0,15 y 0,20, lo que equivale a contaminar la distribución $\mathcal{N}(0,1)$ con los porcentajes de *outliers* 5 %, 10 %, 15 % y 20 %, generados de una distribución uniforme $\mathcal{U}(a,b)$. La simulación antes descrita se realizó mediante la ayuda de algoritmos en el entorno de programación R.

4. Resultados y Discusión.

En la tabla 1 se muestran los resultados de la prueba de Anderson-Darling para la precipitación y humedad, donde se observa que ambas variables muestran una tendencia a distribuirse normalmente ($p < 0{,}05$), lo que verifica el cumplimiento del supuesto de normalidad, que es un requisito previo para el uso de los Residuos Studentizados. Estos resultados coinciden con los reportados por Guenni *et al.* (2008) en un estudio sobre la tendencia y estacionalidad de la precipitación mensual en Venezuela.

Tabla 1: Prueba de Normalidad (Anderson-Darling) para la precipitación mensual (mm) y humedad (%) medida en la estación meteorológica Mesa de Cavacas del municipio Guanare, estado Portuguesa en el periodo (1978-2010). ($p < 0{,}05$: Cumplimiento del supuesto de normalidad).

Variable	n	Estadístico de prueba (Anderson-Darling)	Significación asintótica (p valor)
Precipitación (mm)	396	7,515	$< 0,005$
Humedad (%)	381	3,664	$< 0,005$

En la tabla 2 se muestran los *outliers* detectados en los registros de precipitación mensual mediante los Residuos Studentizados, allí se observa que se detectaron 3,54 % de *outliers* en toda la serie, los cuales tienden a distribuirse hacia los meses de mayo, junio y julio, con algunos casos en los meses de agosto, septiembre y octubre. Estos *outliers* están asociados a valores de precipitación bastante altos (358,7 mm hasta 479,1 mm), acompañados de residuos studentizados que van desde 2,03 hasta 2,7, los cuales los identifican como *outliers*. En tal sentido, los mismos se corresponden a precipitaciones mensuales registradas en los años 1979, 1981, 1982, 1983, 1990, 1992, 1993 y 1994. En tal sentido, el año 1981 es el que muestra la mayor cantidad de *outliers*, por lo que se puede decir que este fue un año bastante atípico en relación a la incidencia de lluvias en los meses de mayo, junio y septiembre. De igual manera, el año 2005 se ubica en segundo lugar en relación a la presencia de *outliers*, lo que lo ubica como un año con precipitaciones elevadas, ocurridas en los meses de mayo y junio. Estos resultados coinciden con lo reportado por Guenni *et al.* (2008) en un estudio sobre la tendencia y estacionalidad de la precipitación mensual en Venezuela, quienes señalan la presencia de valores extremos que indican niveles de precipitación elevados, referidos a las estaciones con ciclo unimodal con niveles pluviométricos más altos que ocurren en junio, julio y agosto, y a las estaciones con ciclo bimodal, en donde los valores máximos ocurren de abril a mayo o de septiembre a octubre.

Tabla 2: *Outliers* detectados (3,54 %) en los registros de precipitación mensual (mm) medida en la estación meteorológica Mesa de Cavacas del municipio Guanare, estado Portuguesa en el periodo (1978-2010).

Año	Mes	Precipitación (mm)	Residuo (mm)	Residuo Studentizado
1979	Julio	396,9	233,12	2,03
1981	Mayo	439,8	295,54	2,57
1981	Junio	467,0	313,79	2,73
1981	Septiembre	479,1	299,05	2,6
1982	Mayo	449,0	305,55	2,66
1983	Junio	418,0	266,42	2,32
1990	Octubre	423,8	242,11	2,1
1992	Junio	441,6	297,32	2,58
1993	Agosto	420,1	258,74	2,25
1994	Julio	406,6	255,00	2,21
1998	Mayo	368,9	238,44	2,07
2004	Mayo	358,7	233,12	2,03
2004	Junio	385,7	251,17	2,18
2005	Mayo	375,8	251,03	2,18

En la tabla 3 se muestran los *outliers* detectados en los registros de humedad mediante los Residuos Studentizados. Allí se observa que se detectaron 3,15 % de *outliers* en toda la serie, los cuales tienden a distribuirse hacia los meses de diciembre y marzo. Estos *outliers* están asociados a valores de humedad

que van desde 44 % hasta 87 %, acompañados de residuos negativos relativamente altos, que van desde -15,89 hasta -29,52 y de residuos studentizados grandes con valores desde -2,00 hasta -3,71, con la excepción del último *outlier* correspondiente al mes de marzo del año 1993, el cual precisamente muestra el valor más alto de humedad en el grupo de valores atípicos detectados. De esta manera, estos valores atípicos se corresponden a valores de humedad registrados en los años 1986, 1987, 1988, 1989, 1990, 1991, 1993 y 2008. En tal sentido, los años 1988 y 1990 son los que muestran la mayor cantidad de *outliers*, por lo que se puede afirmar que estos fueron dos años atípicos en relación a la humedad, fundamentalmente años caracterizados por una baja humedad asociada al período de noviembre a marzo. Estos resultados coinciden con lo señalado por Quiroz *et al.*, (2017) en un estudio de la tendencia de la precipitación y las sequías en Venezuela.

Tabla 3: *Outliers* detectados (3,15 %) en los registros de humedad relativa (%) en la estación meteorológica Mesa de Cavacas del municipio Guanare, estado Portuguesa en el periodo (1978-2010).

Año	Mes	Humedad (%)	Residuo	Residuo Studentizado
1986	Diciembre	62	-15,89	-2
1987	Diciembre	55	-23,05	-2,9
1988	Enero	47	-17,43	-2,59
1988	Marzo	48	-18,94	-2,38
1988	Diciembre	62	-16,21	-2,04
1989	Abril	49	-19,35	-2,43
1990	Febrero	44	-21,99	-2,76
1990	Noviembre	58	-19,27	-2,42
1990	Diciembre	49	-29,52	-3,71
1991	Diciembre	59	-19,68	-2,47
1993	Marzo	87	19,28	2,42
2008	Marzo	52	-18,07	-2,27

En la tabla 4 se muestran los parámetros $(X_{(1)}, X_{(n)})$ de la distribución uniforme utilizada para contaminar muestras de tamaño n=10, 20, 30 y 50 en modelos lineales considerando tres distribuciones de los regresores (uniforme, normal y exponencial), donde se observa que los parámetros mostraron un comportamiento similar, tanto en los valores de estos, como en la tendencia a disminuir conforme se incrementa el tamaño de la muestra y la proporción de *outliers*. Estos resultados verifican la consistencia de los estimadores $(X_{(1)}, X_{(n)})$ de los parámetros de la distribución uniforme $P_{u(X_{(1)},X_{(n)})}(x)$. Se observaron valores de $X_{(1)}$ que van desde -31 hasta -11 y de $X_{(n)}$ desde 11 hasta 31. Es importante resaltar que los residuos studentizados permitieron verificar que el modelo mixto $P(x) = \pi\, P_{D(0)}(x) + (1-\pi)\, P_{u(X_{(1)},X_{(n)})}(x)$ utilizado para simular errores contaminados con observaciones atípicas fue eficiente para incorporar los porcentajes de *outliers* (5 %, 10 %, 15 % y 20 %) tomando como referencia los valores de $(X_{(1)}, X_{(n)})$ de la distribución uniforme. No obstante, cuando se realizó el análisis de observaciones influyentes con el método de la distancia de Cook solo se logró detectar hasta un 6 % de observaciones influyentes en cada una de las muestras consideradas.

5. Conclusiones

Se detectaron 3,54 % de *outliers* en la serie de precipitación, específicamente en mayo, junio y julio, asociados a grandes residuos studentizados. Así mismo, las precipitaciones más altas se registraron en los años 1979, 1981, 1982, 1983, 1990, 1992, 1993 y 1994. El año 1981 reportó la mayor cantidad de *outliers*, por lo que se concluye que fue un año bastante atípico en relación a la incidencia de lluvias. El año 2005 fue el segundo en relación a la presencia de *outliers*, lo que lo convierte en un año con precipitaciones elevadas en los meses de mayo y junio. Se detectaron 3,15 % de *outliers* en la serie de humedad, los cuales se distribuyeron entre diciembre y marzo, con valores que van de relativamente

bajos a altos, residuos negativos altos y residuos studentizados grandes. Los años 1988 y 1990 fueron los de mayor cantidad de *outliers* en relación a la humedad, considerados años atípicos con una baja humedad en el período de noviembre a marzo. Se recomienda realizar estudios de simulación con modelos climatológicos tales como Normal, Gumbel y Weibull para evaluar el efecto de los *outliers* sobre estos modelos. Se observó que los parámetros $(X_{(1)}, X_{(n)})$ de la distribución uniforme utilizada para contaminar con *outliers* modelos lineales en cada una de las distribuciones mostraron un comportamiento similar, tanto en los valores de estos, como en la tendencia a disminuir conforme aumenta el tamaño de la muestra y la proporción de *outliers*, lo que verifica la consistencia de los estimadores $(X_{(1)}, X_{(n)})$ de los parámetros de la distribución uniforme . Los residuos studentizados permitieron verificar que el modelo mixto $P(x) = \pi\, P_{D(\theta)}(x) + (1-\pi)\, P_{u(X_{(1)},X_{(n)})}(x)$ fue eficiente para incorporar porcentajes de *outliers* en la muestra, mas no así para incorporar observaciones influyentes por encima de 6 %, lo que sugiere considerar otras distribuciones simétricas en el modelo mixto para incrementar el porcentaje de estas en la muestra.

Tabla 4: Parámetros de la distribución uniforme utilizada para contaminar con % de *outliers* muestras provenientes de modelos lineales simulados.

n	% teórico de *outliers*	Parámetros de la distribución uniforme $P_{u(X_{(1)},X_{(n)})}(x)$ utilizados para contaminar la muestra con % outliers					
		Uniforme		Normal		Exponencial	
		$X_{(1)}$	$X_{(n)}$	$X_{(1)}$	$X_{(n)}$	$X_{(1)}$	$X_{(n)}$
10	5	-31	31	-31	31	-31	31
	10	-22	22	-22	22	-21	21
	15	-18	18	-18	18	-18	18
	20	-16	16	-14	14	-11	11
20	5	-22	22	-23	23	-22	22
	10	-18	18	-17	17	-18	18
	15	-15	15	-15	15	-15	15
	20	-13	13	-13	13	-13	13
30	5	-20	20	-19	19	-20	20
	10	-17	17	-16	16	-17	17
	15	-15	15	-14	14	-14	14
	20	-13	13	-12	12	-12	12
50	5	-17	17	-18	18	-18	18
	10	-15	15	-16	16	-16	16
	15	-13	13	-13	13	-14	14
	20	-11	11	-12	12	-12	12

6. Bibliografía

Beckman R, Cook RD (1983): Outlier ... ..., s. *Technometrics*, 25:119-149.

Birkes D, Dodge Y (1993): *Alternative methods of regression*. John Wiley and Sons, New York.

Chacín F (1998): *Análisis de regresión y superficies de respuesta*. Facultad de Agronomía. Universidad Central de Venezuela,Venezuela.

Cook R (1977): Detection of influential observations in linear regression. *Technometrics*, 19:15-18.

Cook R, Weisberg S (1980): Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22:495-508.

Ewel J, Madriz A, Tosi J (1976): *Zonas de vida de Venezuela*. Memoria explicativa sobre el mapa ecológico. 2ª edición. MAC-FONAIAP, Caracas.

Guenni L, Degryze E, Alvarado K (2008): Análisis de la tendencia y la estacionalidad de la precipitación mensual en Venezuela. *Revista Colombiana de Estadística*. 31:41-65.

Holdridge LR (1967): *Life zone ecology*. Tropical Science Center, San José, Costa Rica.

Larreal M, Schargel R, Salazar E, Chacón E, Díaz M, Jiménez A, Sánchez T, Hernández J (1979): *Metodologías y algunos resultados de los estudios de los suelos de las cuencas entre los ríos Guanare y Masparro*. MARNR, Zona 8, Guanare estado Portuguesa.

MPPPA (2009): *Record climatológico Estación Mesa de Cavacas. Serial 2281. Periodo (1978-2009)*. Dirección Estadal Ambiental Portuguesa.

Quiroz I, Paredes F, Guevara E (2017): *Análisis de la tendencia de la precipitación y las sequías en Venezuela*. Disertaciones Doctorales en Ambiente y Desarrollo. Coordinación de Postgrado del Vicerrectorado de Infraestructura y Procesos Industriales. Cojedes, Venezuela, pp. 31-51.

Rengel A, Ortega F, Aymard G (1983): Dinámica de las variaciones de la cobertura vegetal y la erosión en el piedemonte de Guanare. *Boletín Técnico Programa R.N.R. UNELLEZ-Guanare*. 8:1-94.