Saber, Universidad de Oriente, Venezuela.Vol. 25 Nº 2: 192-201. (2013)

COMPARACIÓN DE MÉTODOS DE DETECCIÓN DE DATOS ANÓMALOS MULTIVARIANTES MEDIANTE UN ESTUDIO DE SIMULACIÓN

COMPARISON OF MULTIVARIATE METHODS FOR OUTLIERS DETECTION BY SIMULATION

Luis Marcano, Wilmer Fermín

Universidad de Oriente, Núcleo de Nueva Esparta, Departamento de Estadística
E-mail: luisjmarcanoverde@hotmail.com, wilmerfermin@gmail.com

RESUMEN

Los valores anómalos son un problema omnipresente en la recolección de datos, son observaciones que se desvían en alguna dirección respecto al comportamiento general del resto del conjunto de datos y pueden afectar los resultados de aplicar métodos estadísticos univariantes o multivariantes. Es fundamental la detección de estos valores, ya sea para eliminarlos o para atenuar sus efectos en el análisis. Se han desarrollado varios métodos para la detección de valores anómalos, entre ellos están la Distancia Robusta de Mahalanobis (DRM) de Rousseeuw y Van Zomeren (1990), la Curtosis-1 de Peña y Prieto (2001) y el método FGR de Filzmoser, Garrett y Reimann (2005). En este artículo se compararon estos tres métodos, en cinco escenarios de correlación considerando variables explicativas con varios porcentajes de anómalos, mediante análisis comparativo de aplicar estos métodos en datos simulados. Los resultados evidencian que la Curtosis-1 es más eficiente que la DRM y el método FGR para la detección de valores anómalos multivariantes, independientemente de la proporción de éstos y la presencia de correlación entre las variables consideradas en el estudio.

Palabras clave: Valores anómalos multivariantes, detección, comparación, simulación.

ABSTRACT

Outliers constitute a constant problem in data collection, they are observations that deviate from the general pattern of the rest of the data and thus can affect the results that derive from the application of univariate and multivariate statistical methods. It is essential to detect these observations, either to eliminate them or to mitigate their effect on the analysis. Several outlier detection methods have been developed, including the Robust Mahalanobis Distance (DRB) by Rousseeuw and Van Zomeren (1990), the Kurtosis-1 by Peña and Prieto (2001) and the FGR method by Filzmoser, Garrett y Reimann (2005). These three methods were compared in this article, in five correlation scenarios considering explanatory variables with several percentages of outliers, by using comparative analysis of these methods in simulated data. Results show that the kurtosis-1 method is more efficient than DRM and FGR for the detection of multivariate outliers, regardless the proportion of outliers and the presence of correlation among variables in the research study.

Key words: Multivariate outliers, detection, comparison, simulation.

INTRODUCCIÓN

Gran parte del éxito del análisis estadístico de datos subyace en la recogida de la información u obtención del conjunto de datos; no obstante, por mucho cuidado que se tenga no se estará libre de errores de muestreo y de valores anómalos (valores atípicos, discrepantes, inusitados, extraños, outliers, entre otras denominaciones). Estos valores se encuentran alejados del comportamiento general del resto del conjunto de datos y no pueden ser considerados totalmente como una manifestación del proceso bajo estudio (Pérez 1987, Rousseeuw y Van Zomeren 1990). Los valores anómalos pueden generar resultados erróneos producto del análisis estadístico y, en consecuencia, es improbable obtener respuestas precisas que permitan caracterizar el proceso en estudio; en razón de ello, es fundamental detectar estos valores, en el conjunto de datos, ya sea para eliminarlos o para atenuar sus efectos en el análisis.

Entre las causas que pueden ocasionar valores atípicos, en la recolección de datos, están: (1) por variación natural, un valor discrepante de este tipo surge de una inevitable y necesaria heterogeneidad intrínseca de algunas unidades de análisis que indican un cambio natural de las partes del fenómeno bajo estudio y, por tanto, resultan de gran interés y son una pieza vital para el entendimiento de dicho fenómeno (Peña 2002); algunos autores mencionan estos valores como anómalos legítimos (Uriel y Aldás 2005) y, (2) por hechos externos al proceso, como errores en el registro de la respuesta, ya sea en el momento en el que se recoge información o en la transcripción (Hardin 2000, Peña 2002). Eliminar a priori las observaciones discrepantes del resto de los datos no es una acción prudente en general; será oportuno y necesario eliminar estas observaciones si

Recibido: julio 2012 Aprobado: marzo 2013.
Versión final: abril 2013

existe evidencia comprobada de que son producto de errores de medición o del analista; si el valor atípico proviene de variación natural, no debe removerse sino más bien resaltarse y tomarlo en cuenta de manera especial en el análisis realizado.

La detección y tratamiento de valores anómalos en el análisis de regresión lineal dispone de una amplia literatura (Cook y Critchley 2000, Jiménez 2001); además, en la última década se han desarrollado procedimientos robustos, para el análisis de datos en presencia de valores atípicos (Martínez 2010). En el ámbito del análisis de datos multivariantes los valores anómalos pueden tener un moderado o severo efecto tanto en las estadísticas descriptivas como en la modelización, reducción de dimensión, segmentación, búsqueda de variables latentes, entre otros, puesto que pueden influir moderada o severamente en la estimación de los parámetros que involucra el método aplicado. Por ejemplo, en análisis de componentes principales según lo señala Jolliffe (citado por Martínez 2010), uno o más valores atípicos pueden afectar los autovalores y/o los autovectores; así mismo en el análisis factorial, en modelos de ecuaciones estructurales, análisis clúster, análisis discriminante, regresión logística, entre otros (Filzmoser *et al*. 2005).

Los valores atípicos en un contexto multivariante son más difíciles de detectar y visualizar gráficamente que en el caso univariado (Peña y Prieto 2001). Se han desarrollado varios métodos para la detección y tratamiento de valores atípicos multivariantes enmarcados en dos enfoques: los basados en distancias y los métodos de búsqueda de proyecciones. El primer enfoque tiene por objeto determinar valores atípicos a través de una medida de distancia al centro de los datos; algunos de los métodos basados en distancia son la Distancia Robusta de Mahalanobis (DRM) de Rousseeuw y Van Zomeren (1990); MULTOUT de Rocke y Woodruff (1999) y BACON de Billor, Hadi y Velleman (2000). El segundo enfoque está dirigido a identificar atípicos, en datos de gran extensión o dimensiones altas, mediante proyecciones en subespacios de menor dimensión (Filzmoser *et al*. 2005, López 1999, Ben-Gal 2005); algunos de los métodos bajo este enfoque son la técnica de componentes principales para la identificación de atípicos de Rao (Pérez 1987); PCOut de Filzmoser, Maronna y Werner (2008); Curtosis-1 de Peña y Prieto (2001); FGR de Filzmoser, Garrett y Reimann (2005); entre otros (Pérez 1987, Rousseeuw y Van Zomeren 1990, López 1999, Peña y Prieto 2001, Ben-Gal 2005, Filzmoser *et al*. 2005, Filzmoser *et al*. 2008).

La distancia robusta de Mahalanobis, desarrollada por Rousseeuw y Van Zomeren (1990), se considera un método clásico y de referencia en muchas de las publicaciones sobre el tema; se aplica tanto en datos de poca como de alta dimisión, aunque con esta última es mas frecuente el problema de enmascaramiento (Hardin 2000, Peña y Prieto 2001, Peña 2002, Ben-Gal 2005, Filzmoser 2004, Filzmoser *et al*. 2005, 2008). Por otro lado, la Curtosis-1 de Peña y Prieto (2001) y el método de Filzmoser *et al*. (2005), son más recientes, innovadores y prometedores de ser más eficientes en la detección de atípicos, fundamentalmente en datos de gran extensión pero siguen siendo válidos en datos de poca dimensión (Filzmoser *et al*. 2008). No obstante, en la bibliografía consultada no se ha encontrado estudio alguno que compare estos tres métodos, en cuanto a su eficiencia, en la detección de valores anómalos en datos con poca o alta dimensión, considerando independencia o correlación entre las variables y con varios porcentajes de valores anómalos presentes. En vista de lo anterior, en el presente trabajo se compararon estos tres métodos; se usó un análisis comparativo (Barrera 2007) a los resultados de aplicar cada método en datos simulados, con baja dimensiónalidad, diferentes estructuras de correlación y porcentajes de atípicos.

MATERIALES Y MÉTODOS

Se realizó un análisis comparativo entre la distancia robusta de Mahalanobis (DRM), la Curtosis-1 y el método FGR de Filzmoser, Garret y Reimann (2005) considerando los fundamentos teóricos y los resultados de aplicar cada método en datos simulados. Específicamente se establecieron semejanzas y diferencias, se verificó cuál de éstos detecta el mayor porcentaje de verdaderos atípicos y a la vez reporta el menor porcentaje de falsos atípicos, en matrices de datos simuladas. Para la simulación se usaron las funciones *rmultnorm* y round(.) del paquete "*MSBVAR*" disponibles para el software R. Se simularon matrices de datos multivariantes con poca dimensionalidad (específicamente cinco variables); estas variables se consideraron independientes y con correlación moderada, con varios porcentajes de valores anómalos; esto se escogió así para estudiar la eficiencia de los métodos en casos de poca dimensión, independencia, correlación moderada y la presencia de pocos y muchos atípicos. El estudio de la eficiencia de los métodos comparados cuando existe multicolinealidad y altas dimensiones es un problema abierto a la investigación y estará pendiente para futuros investigaciones.

Identificación de Valores Anómalos Multivariantes

Supóngase que se han medido u observado "*p*" variables en "*n*" objetos, dígase el vector Xi $X_i=(X_{i1},X_{i2},...,X_{ip})';i=1,2,...,n.$ La información obtenida estará reflejada en la matriz de casos-variables $X=(xij)_{n,p}$, la cual se puede visualizar como una nube de "*n*" puntos en el espacio *p*-dimensional, R^p. Dada la omnipresencia de los valores atípicos es posible que la nube de puntos tenga algún anómalo multivariante; es decir, alguna observación que se diferencie notablemente del resto de los datos.

Un método clásico para la detección de outliers multivariantes consiste en el cálculo de la distancia de Mahalanobis, d^i, para cada una de las observaciones con respecto a la media vectorial, $\bar{X} = (\bar{X}_1, ...\bar{X}_p)'$. Según Rousseeuw y Van Zomeren (1991), este estimador (distancia) tiene una distribución aproximadamente ji-cuadrado con "p" grados de libertad (χ^2_p). Un valor atípico multivariante se corresponderá con una d_i^2 grande; esto es $d_i^2 > X^2_{p,1-\alpha}$ para un nivel α, usualmente igual a 0,05; 0,01 o 0,025. Esta distancia viene dada por:

$$d_i^2 = (X_i,\bar{X}) = (X_i - \bar{X})'S_x^{-1}(X_i,-\bar{X}) \quad (1)$$

siendo *i*=1,2,…,*n* y *Sx* la matriz de varianzas-covarianzas

Esta forma de detectar valores anómalos multivariantes resulta poco eficiente, debido a que las estimaciones del vector de medias y la matriz de covarianza se ven distorsionadas por la presencia de los valores atípicos; en consecuencia, el cálculo de la distancia de Mahalanobis produce el efecto enmascaramiento (*masking effect*) (Chiang 2007), que consiste en dar distancias pequeñas para algunos anómalos, reportándose falsamente como observaciones del grueso general de los datos (observaciones típicas o que siguen la distribución generadora de los datos) y, el efecto inundación (*swamping effect*) (Chiang 2007), que consiste en dar grandes distancias a observaciones típicas, reportándose éstas como falsos atípicos. Es decir, un valor atípico afecta la distancia de Mahalanobis de manera tal que se puede encubrir a otro o identificar, erróneamente, a una observación normal como atípica (Chiang 2007, Ben-Gal 2005, Filzmoser 2004, Peña 2002).

Distancia Robusta de Mahalanobis (DRM)

La DRM de Rousseeuw y Van Zomeren (1990) se basa en el uso de la distancia de Mahalanobis, pero sustituyendo los estimadores de los parámetros de localización (vector de media) y escala (matriz de covarianza), en la expresión (1), por estimaciones robustas; acostumbrándose a estimar esta dupla a través del método Mínimo Determinante de Covarianza (MDC). La distancia robusta de Mahalanobis viene dada por:

$$d_i = (X_i - \bar{X}_{MDC})' S^{-1}_{MDC}(X_i - \bar{X}_{MDC}) \quad (2)$$

siendo $\bar{X}_{MDC}$ y S^{-1}_{MDC} los estimadores robustos para la media y matriz de varianza-covarianza obtenidos mediante el MDC

Rousseeuw y Van Zomeren (1990, 1991), demostraron que d_i tiene distribución ji-cuadrado con "p" grados de libertad (χ^2_p). La d_i para el i-ésimo sujeto, respecto a $\bar{X}_{MCD}$, que exceda el cuantil $\chi^2_{p,1-\alpha}$, para algún α pequeño (por ejemplo 0,1; 0,05; 0,01; 0,025), conduce a identificar dicha observación como un valor anómalo multivariante. Usando la DRM para la detección de valores anómalos se resuelven los problemas de enmascaramiento e inundación que afectan a la distancia de Mahalanobis tradicional (Calvo 2010, Rousseeuw y Van Zomeren 1990, 1991).

La Curtosis-1

La Curtosis-1 (Peña y Prieto 2001) se basa en proyectar la nube de "n" puntos en R^p sobre dos nuevos espacios p-dimensionales: el primero obtenido con las direcciones ortogonales de máxima curtosis, y el segundo obtenido de las *p* direcciones ortogonales de mínima curtosis; coeficientes de curtosis muy altos o muy bajos, sugieren la presencia de valores atípicos; se identifican como posibles valores anómalos a aquellas observaciones que son extremas en tales direcciones (Hernández 2005). El método consta de los siguientes pasos:

1- Con los datos estandarizados se calculan *p* direcciones ortogonales de máxima curtosis (y *p* direcciones ortogonales de mínima curtosis), resolviendo:

$$d_j = argmax_d \frac{1}{n}\sum_{1=1}^{n}\left(d'y_i^{(j)}\right)^4; \quad (3)$$

tal que $d'\,d = 1$ y arg max(.) argumento a maximizar.

2- Proyectar los datos de forma univariante, en cada una de las $j = 1,..,p$ direcciones, $z_i^{(j)} = d'y_i^{(j)}$.

3- Determinar $r_i = \underset{1 \le j \le 2p}{max} \frac{\left|z_i^{(j)} - mediana\left(z^{(j)}\right)\right|}{DAM\left(z^{(j)}\right)}$, (4)

siendo *DAM* ($z^{(j)}$) la desviación absoluta respecto a la mediana de los datos proyectados, $z^{(j)}$. Si $r > \beta_p$, entonces la *i*-ésima observación es sospechosa de ser atípica y es etiquetada como tal; las que no son sospechosas forman un conjunto U; el valor crítico β_p es escogido para asegurar un nivel razonable de error Tipo I; Peña y Prieto (2001), presentan una tabla con diversos valores para β_p.

4- Se calcula la distancia de Mahalanobis, denotada d_i^R, de cada una de las observaciones, etiquetadas como posibles outliers, con respecto a la media de las observaciones no sospechosas (esas que forman el conjunto U del paso anterior).

5- Aquellas observaciones $i \notin U$, tal que, $d_i^R < X^2_{p,1-\alpha}$ no son consideradas como outliers y se incluyen en el conjunto *U*. Este proceso se repite hasta que no existan observaciones candidatas a pertenecer al conjunto *U*, o hasta que *U* venga a ser el conjunto de todas las observaciones originales (Hernández 2005). Los atípicos serán aquellas observaciones $i \notin U$, tal que, $d_i^R < X^2_{p,1-\alpha}$.

Método de Filzmoser, Garrett y Reimann (FGR)

El método FGR se basa en la comparación de la distribución empírica de la distancia robusta de Mahalanobis y la distribución teórica de la misma (la cual es la distribución X^2_p). Para una descripción de los argumentos del método, considérese que $G_N(u)$ es la distribución empírica de la distancia robusta de Mahalanobis y que G (*u)* es la función de distribución teórica X^2_p. Sea, además, $p(\delta)$ la medida de la desviación de las distribuciones empíricas y teóricas sólo en las colas, definida por el valor $\delta = X^2_{p,1-\alpha}$, es decir:

$$p(\delta) = \sup_{u \geq \delta} \{G(u) - G_n(u)\}^+ \qquad (5)$$

donde $\{\cdot\}^+$ indica la parte positiva. Si $p(\delta)$ es más grande que un valor crítico, dígase $p_{crit}(\delta,n,p)$, puede considerarse como una medida de anómalos en la muestra, en otro caso la medida es cero (Filzmoser 2004; Filzmoser *et al*. 2005). Esto es:

$$\alpha(\delta) = \begin{cases} 0 & si\ p(\delta) \leq p_{crit}(\delta,n,p) \\ p_n(\delta) & si\ p(\delta) \geq p_{crit}(\delta,n,p) \end{cases} \qquad (6)$$

El valor crítico, también llamado cuantil ajustado, empleado para identificar los valores anómalos en una muestra es $C = G_N^{-1}(1-\alpha(\delta))$. Una observación multivariante será identificada como atípica si la distancia robusta de Mahalanobis, con respecto a la media vectorial, es mayor que el cuantil ajustado, *C*.

Comparación de los Métodos Mediante Estudio por Simulación

Los fundamentos teóricos demarcan una conceptualización diferente de cada método para la detección de valores atípicos multivariantes; no obstante, tienen en común el concepto de distancia y la distribución ji-cuadrada asociada a ésta. La DRM se basa en estudiar la separación del anómalo del centro de masa de los datos considerando estimaciones robustas; mientras que la curtosis-1 se basa en analizar lo aplastado o puntiagudo que los anómalos pueden deformar la distribución de los datos, es por ello que se trabaja con las distancias de las proyecciones de mínima y máxima curtosis; por otro lado, el método FGR detecta anómalos a través de las máximas o mínimas discrepancias calculadas de las distribuciones teóricas y observadas en los datos, las cuales tienen distribución χ^2_p.

Desde un punto de vista práctico la comparación de los métodos DRM, Curtosis-1 y FGR se basó en aplicar éstos a matrices de datos simuladas. El proceso de simulación se realizó generando "mediciones de p = 5 variables estandarizadas en n = 200 sujetos" en cinco escenarios; en cada uno (excepto en el primero que requirió sólo 100 matrices), se simularon quinientas matrices de orden n.p, de las cuales 100 matrices fueron contaminadas con un anómalo cada una (α = 0,01), 100 matrices con 5 anómalos cada una (α = 0,05), 100 matrices con 10 anómalos cada una (α = 0,1), 100 con 15 (α = 0,15) y 100 con 20 (α = 0,2) anómalos. Posteriormente se aplicó cada método a las 500 matrices en cada escenario y se procedió a cuantificar tanto los verdaderos anómalos detectados así como las observaciones típicas detectadas y como anómalas. Específicamente los escenarios son siguientes:

(1) Simulación desde la distribución $N_5(0, I_5)$. En este caso no se impuso atípicos a las matrices de datos.

(2) Simulación de datos con la distribución $N_5(0,I_5)$ y contaminados con valores anómalos de la distribución $N(3\varepsilon, I_5)$, con $\varepsilon = (1,1,1,1,1)'$. La rutina usada en el software R fue round(rmultnorm(a,mu,s),rmultnorm(b,3*e,s)),2); donde: a=200(1-α), b=200α, mu=0 , s=I_5, e= ε y siendo α la proporción de anómalos. Se obtuvieron 100 matrices de datos de orden 200 x 5 cada una con un anómalo; 100 con 5 anómalos, 100 con 10, 100 con 15 y 100 matrices con 20 cada una. Este escenario es igual al anterior

pero se impone anómalo a las matrices.

(3) Simulación de datos con la distribución $N_5(0,I_5)$ y contaminados con anómalos desde la distribución $N_5(3\varepsilon, 0{,}01 \cdot I_5)$. La diferencia de este caso con el anterior que los anómalos están mas cercanos entre sí.

(4) Simulación de datos con la distribución $N_5(0,S_1)$ y contaminados con valores anómalos de la distribución $N(3\varepsilon, S_1)$. S_1 se estableció considerando correlaciones moderadas (entre 0,5 y 0,7) y arbitrarias entre las variables simuladas. Explícitamente:

$$S_1 = \begin{pmatrix} 1 & 0{,}55 & 0{,}70 & 0{,}68 & 0{,}06 \\ 0{,}55 & 1 & 0{,}68 & 0{,}67 & 0{,}49 \\ 0{,}70 & 0{,}68 & 1 & 0{,}66 & 0{,}62 \\ 0{,}68 & 0{,}67 & 0{,}66 & 1 & 0{,}66 \\ 0{,}60 & 0{,}49 & 0{,}62 & 0{,}66 & 1 \end{pmatrix}$$

(5) Simulación de datos con la distribución N_5 (μ_1,S_1) y contaminados con outliers provenientes de una distribución N_5 (μ_2,I_5), donde μ_1=(2,4,1,6,10)´ y μ_2=(5,6,3,8,5)´. En este escenario no se consideró correlación entre las variables para la distribución generadora de los anómalos.

Con el método FGR se consideró una observación como anómala si la distancia robusta de Mahalanobis con respecto al vector de medias es mayor que el cuantil ajustado; en este caso se definió $\delta = X^2_{5;\,0{,}975}$. Con la DRM y la Curtosis-1 se utilizó el cuantil $\delta = X^2_{5;\,0{,}975}$ para diferenciar los valores atípicos de las observaciones no contaminadas. Los cálculos de aplicar DRM y FGR se realizaron mediante las funciones *uni.plot(.)* y *aq.plot(x*) respectivamente, del paquete *MSBVAR*, disponible en el software R; mientras que los cálculos para la Curtosis-1 se realizaron mediante la función *kur_rce(.)*, integrada en un programa bajo el software Matlab, desarrollado por Peña y Prieto (2001) y disponible en *http://halweb.uc3m.es/fip/download3.html.*

Es importante destacar que la simulación permitió comparar la eficiencia de los métodos en baja dimensionalidad (en particular, p = 5 variables); en estas circunstancias la DRM maximiza su eficiencia en la detección de atípicos; no obstante, la Curtosis-1 como FGR están concebidos para maximizar su eficiencia en datos de alta dimensionalidad. En ese sentido, esta investigación da respuesta ante la incertidumbre de los métodos en la detección de anómalos en baja dimensionalidad.

RESULTADOS

En la Tabla 1 se presentan los resultados de aplicar cada método, en cada escenario y con las diferentes proporciones (α) de valores anómalos del proceso simulado. Los resultados son el porcentaje de verdaderos anómalos observados correctamente y porcentaje de falsos anómalos reportados (esas observaciones del comportamiento general de los datos detectadas como valores atípicos). En el escenario (1) no se considera la detección de verdaderos anómalos ya que a las matrices simuladas no fueron contaminadas con atípicos; no obstante, los métodos sí reportaron falsos anómalos. Como se puede ver en la tabla, la DRM detectó un 2,86% falsos anómalos; la curtosis-1 detectó 1,17% mientras que FGR detectó un 0,78%. Se puede apreciar que el método FGR es más eficiente en cuanto a que no detecta falsos anómalos.

En el escenario (2) se puede apreciar, en la Tabla 1, que los tres métodos son igual de eficientes en la detección de verdaderos anómalos, excepto FGR que tuvo un porcentaje bajo de detección (69%) en matrices con un anómalo; pero en el caso de detección de falsos anómalos la eficiencia de FGR fue mejor, excepto cuando las matrices tenían un anómalo, pues la curtosis-1 se desenvuelve mejor. La Figura 1 presenta el comportamiento de cada método; en la parte superior de la figura se ve el comportamiento en la detección de verdaderos anómalos; tanto la curtosis-1 como la DRM se comportan igual para los diferentes porcentajes de anómalos presentes, mientras que FGR es poco eficiente cuando la proporción de anómalos en los datos es baja pero mejora su detección cuando el número de valores atípicos incrementa. En la parte inferior de la figura se ve el comportamiento de los métodos en cuanto a la detección de falsos anómalos; en este caso la curtosis-1 es más eficiente porque no detecta falsos anómalos independientemente del porcentaje presente en los datos.

Tabla 1. Resultados de aplicar los métodos DRM, Curtosis-1 y FGR a los datos simulados considerando correlación entre variables y diferente proporción (α) de anómalos.

Método		Porcentaje de anómalos								
		Correctamente detectados				Incorrectamente identificados				
		Escenario Reproducido				Escenario Reproducido				
		(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
DRM						2,86				
	0,01	100	100	71,5	95		11,7	2,7	2,73	3
	0,05	99,5	100	60,3	100		1,9	1,8	2,19	2,2
	0,1	99,8	100	54,25	100		1,4	1,5	1,5	1,35
	0,15	97,9	61	41,53	100		0,9	3,6	1,01	0,83
	0,2	96,4	0	26,52	100		0,7	15,1	0,79	0,72
FGR						0,78				
	0,01	69	60	32,5	60		5,7	1	0,96	1,15
	0,05	99,7	100	48,5	100		0,7	0,5	1,21	0,7
	0,1	99,3	100	43,3	100		0,3	0,4	0,99	0,35
	0,15	99,2	61	33,60	100		0,1	2,9	0,75	0,11
	0,2	99,1	0	20,15	100		0,1	14,2	0,38	0,17
Curtosis-1						1,17				
	0,01	100	100	59,5	98		1,4	1,2	1,02	1,15
	0,05	99,5	100	43,1	100		1,1	1	0,92	0,75
	0,1	99,8	100	23,7	100		1,1	1,2	0,74	1,15
	0,15	97,9	97,9	4,9	100		1	0,9	0,59	1
	0,2	96,4	96	3,58	100		1	1	0,42	1,44

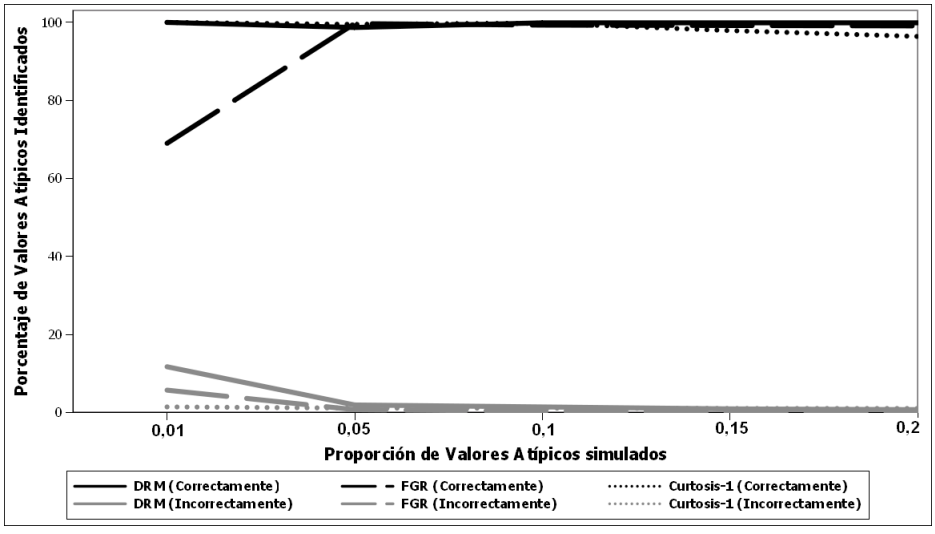

Figura 1. Comportamiento de DRM, Curtosis-1 y FGR en el escenario 2.

En el escenario (3) se puede apreciar, en la Tabla 1, que la curtosis-1 detecta muy bien los verdaderos anómalos, independientemente del porcentaje de datos atípicos presentes en la matriz; los métodos DRM y FGR no fueron capaces de detectar anómalos cuando las matrices tenían veinte de éstos y detectaron sólo el 61% de las veces para quince anómalos en los datos. Por otra parte, la curtosis-1 es la que reporta el mínimo número de falsos anómalos a través de las diferentes proporciones; la DRM y FGR llegan a reportar 15 y 14% de falsos anómalos. La Figura 2 presenta el comportamiento de cada método; en la parte superior de la figura se ve el comportamiento en la detección de verdaderos anómalos; la curtosis-1 detecta bien a través del porcentaje de anómalos; mientras que FGR es poco eficiente, con pocos (1%) y muchos (15 y 20%) anómalos en los datos; la DRM se comporta igual que la FGR con 15 y 20%. En la parte inferior de la figura se evidencia el comportamiento de los métodos en cuanto a la detección de falsos anómalos; en este caso la curtosis-1 es más eficiente porque no detecta falsos anómalos independientemente del porcentaje de éstos presentes en los datos, mientras que la DRM y FGR detectan mayor número de falsos anómalos a medida que las matrices de datos tengan mayor proporción.

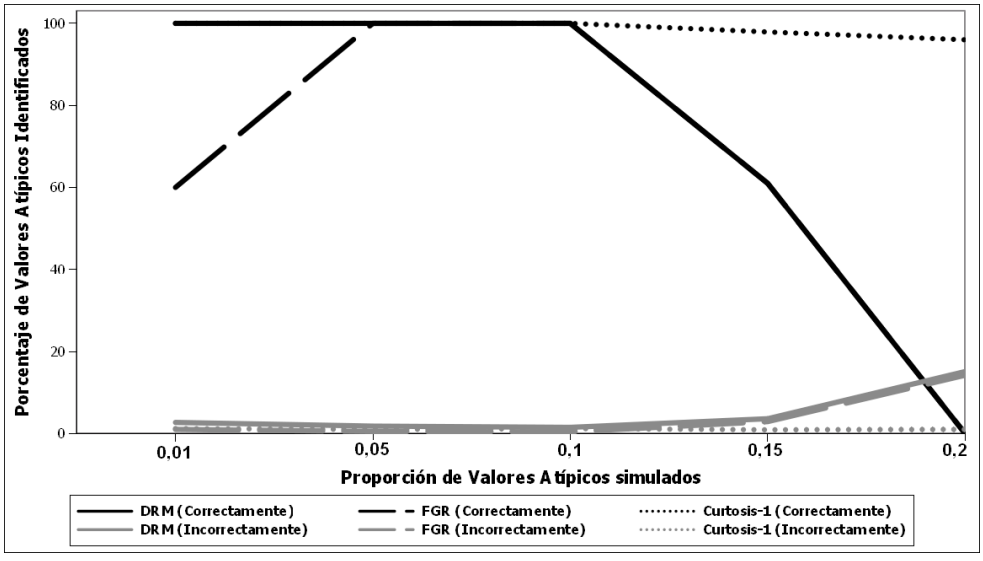

Figura 2. Comportamiento de DRM, Curtosis-1 y FGR en el escenario 3.

En el escenario (4) se puede apreciar, en la Tabla 1, que ninguno de los métodos mostró un buen desempeño en la detección de verdaderos anómalos, independientemente del porcentaje presente en los datos; se puede ver que con un anómalo la DRM lo detectó el 71,5% de la veces seguido de la curtosis-1 con 59,5% y FGR con 32,5%; con 20 anómalos en las matrices la DRM detectó el 26,5% seguido de FGR y la curtosis-1 con el peor desenvolvimiento con 3,58%. En cuanto a la detección de falsos anómalos, la curtosis-1 reportó el mínimo número de falsos anómalos, seguido de FGR y la DRM. La Figura 3 presenta el comportamiento de cada método en este escenario; en la parte superior de la figura se ve el comportamiento en la detección de verdaderos anómalos; la DRM es más eficiente a través de todo el rango de anómalos en los datos; FGR es el que peor se desenvuelve con pocos anómalos pero mejora, sin llegar a superar la DRM cuando hay mayor porcentaje de datos en la muestra. En la parte inferior de la figura se ve el comportamiento de los métodos en cuanto a la detección de falsos anómalos, la eficiencia es similar pero ligeramente superior en la curtosis-1.

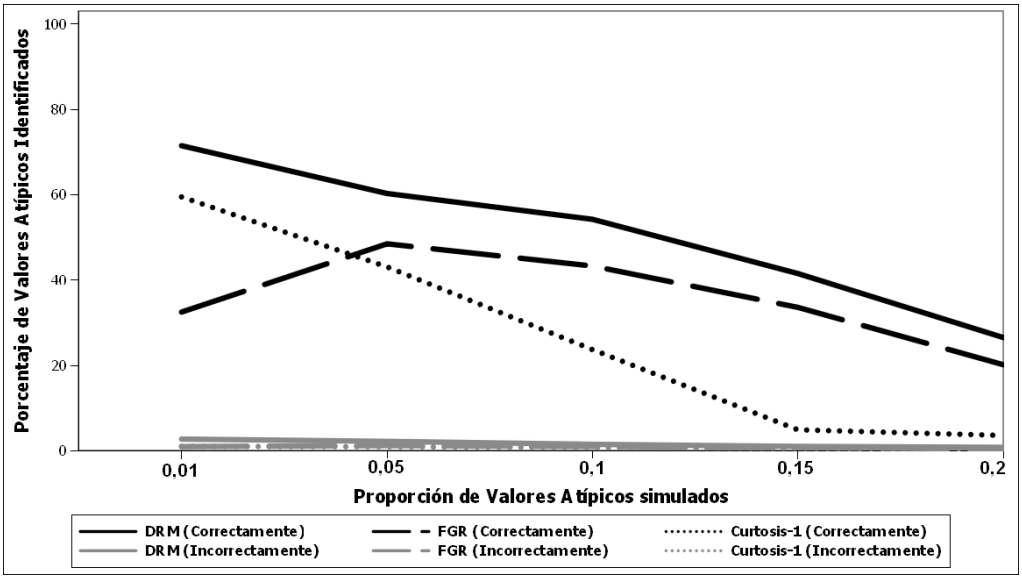

Figura 3. Comportamiento de DRM, Curtosis-1 y FGR en el escenario 4.

Finalmente, en el escenario (5) se puede apreciar, en la Tabla 1, que los tres métodos son igual de eficientes en la detección de verdaderos anómalos en los diferentes porcentajes, excepto el de FGR que tuvo un porcentaje de detección bajo (60%) en el caso de matrices con un anómalo; en la detección de falsos anómalos de igual forma la eficiencia fue igual para los tres método. La Figura 4 presenta el comportamiento de cada método; la detección de verdaderos anómalos con la curtosis-1 es ligeramente mejor cuando se dispone de pocos anómalos en la data. No obstante, en cuanto a la detección de falsos anómalos, la eficiencia es similar en los tres métodos.

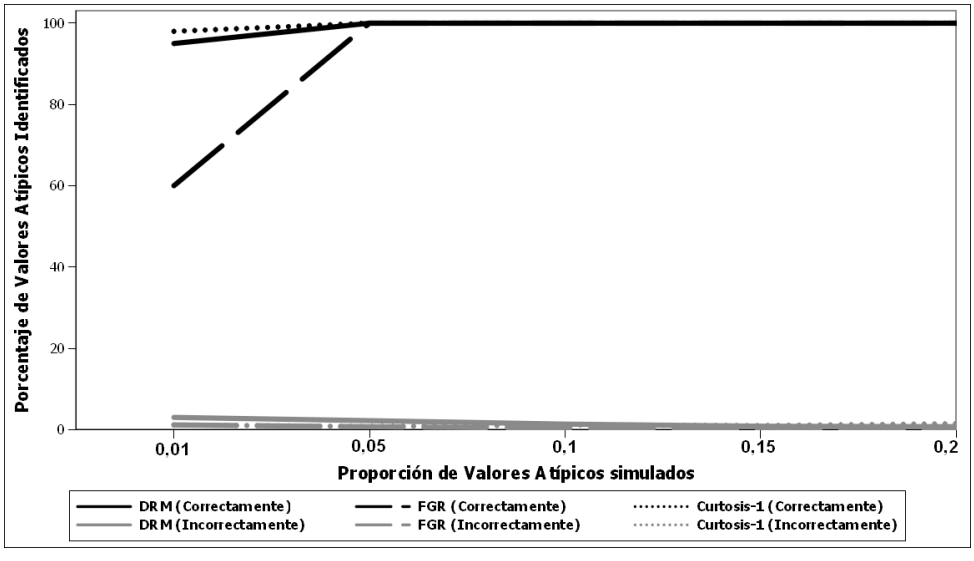

Figura 4. Comportamiento de DRM, Curtosis-1 y FGR en el escenario 5.

DISCUSIÓN

Los valores anómalos afectan los resultados del análisis estadístico cuando se aplican métodos uni o multivariantes, en consecuencia es fundamental para el analista de datos detectar estos valores para eliminarlos o atenuar sus efectos del análisis. Entre los diferentes métodos de detección de anómalos se han comparado la DRM, la curtosis-1 y el método FGR mediante análisis comparativo de los fundamentos teóricos y aplicación práctica en datos simulados. La simulación, arbitraria, de 2100 matrices de datos, con varias proporciones de

anómalos, permitió estudiar la eficiencia de cada método, en baja dimensión (5 variables), en cuanto a si, por una parte, detectaba la proporción exacta de verdaderos anómalos y por otra parte, no reportaba falsos anómalos.

Los resultados permiten constatar, a través de los cinco escenarios estudiados, en cuanto a la detección de verdaderos anómalos que los tres métodos detectan un alto porcentaje en los escenarios (1), (2), (3) y (5); incluso en (5) los tres métodos detectan el 100% de los valores atípicos, excepto en las matrices con un anómalo donde FGR lo detecta el 60% de las veces, mientras que la DRM lo hace el 95% y la curtosis-1, con el mejor desempeño, 98% de las veces. Sin embargo, en el escenario (4), donde se introduce correlación moderada entre las variables, tanto en la distribución para el grueso general de los datos como para la distribución de los anómalos, los tres métodos se hacen ineficientes en la detección del porcentaje de verdaderos anómalos; el peor desempeño se le atribuye a la curtosis-1 que, cuando la verdadera proporción de anómalos es 20%, llega a detectar sólo 3,58%; el mejor desempeño en esta proporción se le atribuye a la DRM que llega a detectar el 26,52% de las veces la verdadera proporción; cuando se tiene un anómalo en las matrices de datos el mejor desempeño se le atribuye a la DRM que lo detecta el 71,5% de las veces, seguido de la curtosis-1 con 59,5% de las veces. En cuanto al reporte de falsos anómalos, los resultados muestran que la curtosis-1 y FGR reportan el mínimo porcentaje en cada escenario; también se puede ver que en (3) y con 20 anómalos en cada matriz, la curtosis-1 detecta hasta 1% de las veces falsos anómalos, mientras que FGR 14,2 % y DRM 15,1% de las veces.

La curtosis-1, en general, es más eficiente en cuanto a la detección de verdaderos anómalos y reporte de falsos anómalos; no obstante, ha mostrado el peor desempeño cuando los datos multivariantes tienen correlación entre las variables, en este caso FGR ha mostrado ser superior a los demás pero sigue siendo muy ineficiente. Por otra parte, la detección de falsos anómalos debe preocupar cuando el analista pretende eliminar estos valores, pues estaría eliminando observaciones del grueso general de los datos cuando no debería; sin embargo, si la matriz tiene valores atípicos, se deben eliminar previa comprobación de que son producto de errores de medición, caso contario deben usarse métodos de análisis robustos.

CONCLUSIÓN

La Curtosis-1 De Peña y Prieto (2001) es más eficiente que la distancia Robusta de Mahalanobis (DRM) y el método FGR de Filzmoser *et al*. (2005), para la detección de valores atípicos multivariantes, independientemente de la proporción de anómalos existentes; no obstante, si la estructura de correlación entre el grueso general de los datos se exhibe como similar a la estructura de correlación entre el grupo de valores atípicos se recomienda usar la DRM.

REFERENCIAS BIBLIOGRÁFICAS

Barrera M. 2007. Análisis en Investigación. Caracas, Venezuela. Ediciones Quirón S.A.

Ben-Gal I. 2005. Outlier detection. Data Mining and Knowledge Discovery Handbook, pp.131-146.

Billor N, Hadi AS, Velleman PF. 2000. BACON: blocked adaptive computationally efficient outlier nominators. Comput. Stat. Data Anal. 4(34):279-298.

Calvo J. 2010. Funções de Autocorrelação robustos, Teste estacionária e Teste de raiz unitária. Tesis de Master no publicada. Universidade Aberta de Portugal, Portugal.

Chiang J. 2007. The Masking and Swamping Effects Using the Planted Mean-Shift Outliers Models. Int. J. Contemp. Math. Sciences. 2(7):297-307.

Cook RD, Critchley F. 2000. Identifying Outliers and Regression Mixtures Graphically. J. Am. Stat. Assoc. 95(451):781-794.

Filzmoser P. 2004. A Multivariate Outlier Detection Method. Disponible en línea en: http://tinyurl.com/7sypyak (Acceso 11.12.2011).

Filzmoser P, Garrett R, Reimann C. 2005. Multivariate Outlier Detection in Exploration Geochemistry. Comput. Geosci. 31(5):579-587.

Filzmoser P, Maronna R, Werner M. 2008. Outlier Identification in High Dimensions. Comput. Stat. Data Anal. 52:1694-1711.

Hardin J. 2000. Multivariate Outlier Detection and Robust Clustering with Minimum Covariance Determinant Estimation and S-Estimation. Universidad de California, EEUU. Disponible en línea en: http://pages.pomona.edu/jsh04747/Research/t.pdf. (Acceso 01.11.2011).

Hernández S. 2005. Biplots Robustos. Tesis doctoral no publicada. Universidad de Salamanca, España.

Jiménez MJ. 2001. Una Generalización de la Estadística de Cook. Rev. Colombiana de Estadística. 24(2):111-120.

Jolliffe IT. 1986. Principal component Analysis: Springer-Verlag.

López V. 1999. Detección de Outliers Multivariables Mediante Projection Pursuit. Universidad Nacional de Colombia, Seccional Medellín, Colombia. Disponible en línea en: http://tinyurl.com/82o9d8s. (Acceso 11.12.2011).

Martínez J. 2010. Una Extensión de la Distancia de Cook a la Regresión de Mínimos Cuadrados Parciales. Tesis doctoral no publicada. Universidad Central de Venezuela, Caracas.

Peña D, Prieto F. 2001. Multivariate outlier detection and robust covariance matrix estimation. Technometrics. 43(3):286-310.

Peña D. 2002. Análisis de Datos Multivariante. Madrid, España: Editorial McGraw-Hill Interamericana de España, S.A; Madrid, España. pp. 120-125.

Pérez J. 1987. Identificación de Outliers en Muestras Multivariantes. Universidad de Sevilla, España. Disponible en línea en: http://tinyurl.com/7zk4rwk. (Acceso 01.11.2011).

Rocke DM, Woodruff DL. 1999. A synthesis of outlier detection and cluster identification.Technical report, University of California, Davis, Davis CA, 95616. Disponible en línea en: http://handel.cipic.ucdavis.edu/ dmrocke/Synth5.pdf. (Acceso: 15.11.2011).

Rousseeuw P, van Zomeren B. 1990. Unmasking Multivariate Outliers and Leverage Points. J. Am. Stat. Assoc. 85(411):633-651.

Rousseeuw P J, van Zomeren BC. 1991. Robust distances: Simulation and cutoff values. Directions in Robust Statistics and Diagnostics, Part II," Springer-Verlag, New York.

Uriel E, Aldas J. 2005. Análisis Multivariante Aplicado. Madrid, Thomson S.A.