Revista en Telecomunicaciones e Informática, Vol. 3, No. 5 p. 11 - 25
Medellín - Colombia. Enero - Junio de 2013, ISSN 2215-8200

Técnicas para detección de outliers multivariantes

Jeniffer Andrea Muñoz García
Estudiante Especialización en Inteligencia de Negocios UPB
jenifferandrea.munoz@alfa.upb.edu.co
Iván Amón Uribe
Coordinador Académico Especialización en Inteligencia de Negocios UPB
ivan.amon@upb.edu.co

Resumen
Este trabajo presenta un acercamiento a un tipo de error relativamente común pero de difícil detección sin las técnicas adecuadas, conocido como valores atípicos multivariantes (*outliers multivariate*), a través de una recopilación de algunos de los métodos utilizados para su detección (distancia de Mahalanobis, componentes principales, búsqueda de proyecciones y un método adaptable para identificar valores atípicos). Adicionalmente, se mencionan algunas herramientas de software utilizadas para este fin.

Abstract
This paper presents an approach to a relatively common error type but difficult to detect without the proper techniques, known as multivariate outliers, through a compilation of some of the methods used for their detection (Mahalanobis distance, main components, projections search and an adaptive method to identify outliers). Additionally, some software tools used for this purpose.

Palabras clave: Outliers multivariantes, valores atípicos.

1. Introducción

El análisis de la calidad de los datos es de gran importancia para las organizaciones, ya que datos con problemas pueden conducir a decisiones erróneas con consecuencias como pérdida de dinero, tiempo y credibilidad. Entre los posibles problemas que pueden presentar los datos, se encuentran los conocidos como valores atípicos o *"Outliers"*. Según Hawkins, un *outlier* es "una observación que se desvía mucho de otras observaciones y despierta sospechas de ser generada por un mecanismo diferente" [1]; aunque estos valores pueden aparentar ser inválidos pueden ser correctos y viceversa.

Los métodos de detección de valores atípicos se pueden dividir en univariados y multivariados. Para el caso univariado, diferentes autores han realizado múltiples investigaciones. Beckman y Cook [2] abordan temas como las técnicas de rechazo para múltiples valores atípicos así como los efectos de enmascaramiento y empantanamiento, los valores atípicos en los datos circulares, el análisis discriminante, el diseño experimental, la distribución no normal, y las series de tiempo. Barnett y Lewis [3] hacen una unificación de los métodos de análisis estadístico para la detección de valores atípicos así como los datos espaciales y los valores atípicos en las series temporales. Para el caso multivariado es mucho más complicado realizar una exploración para llegar a encontrar estos valores, debido al problema de la dimensión, por lo tanto se hace necesario conocer cuáles son los métodos existentes que permiten detectar este tipo de *outliers*.

Según Gnanadesikan y Kettenring, los *outliers* multivariantes son observaciones que se consideran extrañas no por el valor que toman en una determinada variable, sino en el conjunto de aquellas. Son más difíciles de identificar que los *outliers* unidimensionales, dado que no pueden considerarse "valores extremos", como sucede cuando se tiene una única variable bajo estudio [4]. Su presencia tiene efectos todavía más perjudiciales que en el caso unidimensional, porque distorsionan no sólo los valores de la medida de posición (media) o de dispersión (varianza), sino muy especialmente, las correlaciones entre las variables [5].

Peat y Barton, definen un valor atípico multivariante, como un caso que es un valor extremo para una combinación de variables. Por ejemplo, un niño de 8 años de edad cuya estatura sea de 155 cms y pese 45 kg es muy inusual y sería un atípico multivariante [6].

El resto del artículo está organizado como sigue: la sección 2 presenta las generalidades acerca de la detección de *outliers* multivariantes; la sección 3 presenta los métodos empleados para su detección; la sección 4 presenta algunas herramientas de software que permiten realizar esta tarea y por último se presentan las conclusiones.

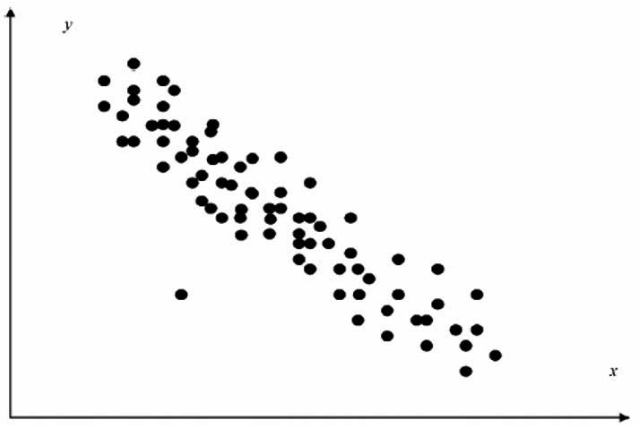
2. Detección de outliers multivariantes

2.1 Características generales

En su texto, Ben-Gal [7] afirma que:

En muchos casos las observaciones multivariantes no pueden ser detectadas como los valores extremos cuando cada variable se considera de forma independiente. La detección de *outliers* sólo es posible cuando se realiza un análisis multivariante y las interacciones entre las diferentes variables se comparan dentro de la clase de datos. Un ejemplo sencillo puede verse en la Figura 1, que presenta puntos de datos que tienen dos medidas en un espacio bidimensional. La observación de la parte inferior izquierda es claramente un caso atípico multivariado, pero no uno univariado. Al considerar cada medida por separado con respecto a la extensión de los valores a lo largo de los ejes x e y, se puede ver que caen cerca del centro de las distribuciones univariantes. Por lo tanto, la prueba para detectar los valores extremos debe tener en cuenta las relaciones entre las dos variables, identificándose así como anormal [6].

Figura 1. Espacio bidimensional con una observación *Outlier* **(esquina inferior izquierda). Tomado de Outlier detection. Capitulo 1. Irad Ben-Gal**



Los conjuntos de datos con múltiples valores atípicos están sujetos a los efectos de enmascaramiento *(masking)* y empantanamiento *(swamping)* [8]. Autores como Acuña y Rodríguez [9], dan una comprensión intuitiva de estos efectos:

Efecto de enmascaramiento. Se dice que un *outlier* enmascara a un segundo *outlier*, si el segundo *outlier* puede ser considerado como un valor extremo sólo por sí mismo, pero no en presencia del primer outlier. Así, después de la eliminación del primer *outlier*, en una segunda instancia, el otro punto se convierte en un valor atípico. El

enmascaramiento se produce cuando un grupo de observaciones extremas sesga las estimaciones de la media y de la covarianza hacia él, y la distancia resultante del valor extremo a la media es pequeña.

Efecto de empantanamiento. Se dice que un *outlier* empantana una segunda observación, si esta última puede ser considerada como un valor extremo sólo bajo la presencia de la primera. En otras palabras, después de la eliminación del primer *outlier*, la segunda observación se convierte en un no-outlier. El empantanamiento ocurre cuando un grupo de valores extremos sesga las estimaciones de la media y de la covarianza hacia él y lejos de otros valores no periféricos, y la distancia resultante de estos casos a la media es grande, haciéndolos parecer como *outliers.*

Para otras definiciones sobre estos efectos véase [1], [3], [10] y [11].

Las técnicas o métodos para detección de valores atípicos multivariantes reunidos en este trabajo y que se tratarán en las posteriores secciones son las siguientes:

- La distancia de Mahalanobis
- Método de las componentes principales
- Búsqueda de proyecciones
- Un método adaptable

3. Métodos empleados para la detecccción de outliers multivariantes

3.1 La distancia de mahalanobis

Los métodos estadísticos para la detección de valores atípicos multivariantes a menudo indican las observaciones que se encuentran relativamente lejos del centro de la distribución de datos. Se pueden implementar varias medidas de distancia para tal tarea [8].

La distancia de Mahalanobis es un criterio muy conocido que depende de los parámetros estimados de la distribución multivariada [8]. Ésta describe la distancia entre cada punto de datos y el centro de masa. Cuando un punto se encuentra en el centro de masa, la distancia de Mahalanobis es cero y cuando un punto de datos se encuentra distante del centro de masa, la distancia es mayor a cero. Por lo tanto, los puntos de datos que se encuentran lejos del centro de masa se consideran valores atípicos [12].

La distancia de Mahalanobis se calcula para cada observación en el conjunto de datos, dándosele a cada observación un peso como inverso de la distancia de Mahalanobis.

Las observaciones con valores extremos obtienen menores pesos. Finalmente, se ejecuta una regresión ponderada para minimizar el efecto de los valores extremos [13].

Supóngase que se tienen dos grupos distintos (poblaciones) que se etiquetan como G1 y G2. Por ejemplo, G1 y G2 pueden representar a las niñas y niños respectivamente o en una situación de diagnóstico médico, las personas sanas y las enfermas respectivamente. Considere la posibilidad de un número (por ejemplo, p) de las características relevantes de los individuos en estos grupos. Estas características o mediciones, pueden estar en algunas características físicas tales como la altura o el peso, o en algunas de las características médicas, tales como la presión sanguínea o el ritmo cardíaco. La variable X denota un vector (al azar) que contiene las mediciones efectuadas en un individuo determinado o entidad objeto de estudio. A menudo, en la práctica, se tiene interés en medir y resumir las diferencias entre los grupos, en este caso G1 y G2. Una suposición común es tomar el vector aleatorio p-dimensional "X" como teniendo la misma variación sobre su media dentro de cualquiera de los grupos. Entonces la diferencia entre los grupos se puede considerar en términos de la diferencia entre los vectores medios de X, en cada grupo con respecto a la variación común dentro de los grupos [14].

Una medida de este tipo es la distancia de Mahalanobis definida de la siguiente manera:

$$MSD_i = \sqrt{(x_i - \bar{x})^T - S_n^{-1}\,(x_i - \bar{x})} \qquad (1)$$

El súper índice T, denota la matriz transpuesta, $\bar{x}$ expresa la media del vector muestral y Sn la matriz de covarianza muestral, donde

$$Sn = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)(x_i - \bar{x}_n)^{\mathsf{T}}.$$

Para datos multivariantes distribuidos normalmente, los valores de la distancia de Mahalanobis tienen aproximadamente una distribución chi-cuadrado con p grados de libertad. En consecuencia, aquellas observaciones con una distancia de Mahalanobis grande se indican como valores atípicos.

Los efectos de enmascaramiento y empantanamiento juegan un rol importante en la adecuación de la distancia de Mahalanobis como criterio para la detección de valores atípicos. Es decir, los efectos de enmascaramiento podrían disminuir la distancia de Mahalanobis de un valor atípico. Esto puede ocurrir, por ejemplo, cuando un pequeño grupo de *outliers* atrae X_n a e infla S_n hacia su dirección. Por otra parte, los efectos de empantanamiento podrían aumentar la distancia de Mahalanobis de las observaciones que no son *outliers*. Por ejemplo, cuando un pequeño grupo de valores atípicos atrae $\bar{x}_n$ e infla S_n lejos del patrón de la mayoría de las observaciones [8].

Los problemas de enmascaramiento y empantanamiento pueden resolverse usando estimaciones robustas, como el estimador M multivariado, el estimador S bicuadrático, el estimador de covarianza de mínimo determinante (MCD), entre otros, los cuales por definición son menos afectados por *outliers*, siendo menos probable que influencien los parámetros usados en la MSD Los puntos que no son atípicos, determinarán completamente la estimación de la forma y posición de los datos. Muchos de los métodos de estimación, incluyendo el método robusto de los M estimadores, fallan si la fracción de *outliers* es mayor que $1/(p+1)$ donde p es la dimensión del conjunto de datos o número de variables, indicando que en dimensiones grandes, una pequeña cantidad de valores atípicos puede producir estimaciones deficientes. Por lo tanto, las distancias de Mahalanobis deben ser estimadas por un procedimiento robusto a fin de proporcionar medidas fiables para el reconocimiento de los valores extremos [15].

3.2 Componentes principales

El método de las componentes principales es una de las técnicas de mayor difusión y utilidad entre las técnicas multivariantes utilizadas para la detección de valores atípicos.

Sea $\mathbf{z} = [x_1 x_2 \ldots x_p]$ un vector aleatorio, con matriz de varianzas y covarianzas $\sum$, definida positiva. Las componentes principales son variables aleatorias Y_1, Y_2, ... Y_p, que cumplen:

- $Y_i = \sum_{j=1}^{p} l_{ij} X_j = l_i' X$ y $l_i' l_i = 1$ para $i = 1,2, \ldots,p$
- $Cov(Y_i, Y_j) = l_i' \sum l_j = 0, i \neq j$
- $Y_i = l_i' X$ es la combinación lineal de máxima varianza.

El siguiente teorema da la relación que existe entre las componentes principales y los valores y vectores propios de $\sum$. Además exhibe el hecho de que la traza de $\sum$ es igual a la suma de las varianzas de las componentes principales.

Teorema 2.1 *Si* $(P_j, \lambda_j), j = 1,2, \ldots, p$ *son los pares de vectores y valores propios de* $\sum$, *y* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$

1. *La j-ésima componente principal está dada por:* $Y_j = P_j' X \quad j = 1,2, \ldots, p.$
2. Si $\sigma_{jj} = var(X_j)$ entonces: $\sum_{j=1}^{p} \sigma_{jj} = \sum_{j=1}^{p} var(Y_j)$

Para la demostración ver Johnson [16].

Ahora se considera el caso de las principales componentes muestrales. Sean $X_{p\times n} = [X_1, X_2, \dots X_n]$, una muestra aleatoria del vector aleatorio $X_{p\times 1}$ y S la matriz de varianzas y covarianzas muestral, con valores y vectores propios asociados: $(P_j, \lambda_j), j = 1,2, \dots, \mathrm{p}$. Entonces las respectivas componentes principales muestrales corresponden a las p-filas de:

$$Y = PZ \qquad (2)$$

donde:

P: Es una matriz ortogonal, $P'P = I_p$, cuyas filas, P_i', son los vectores propios de S con sus respectivos valores propios asociados λ_i, expresados en orden decreciente de magnitud, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$

Z: Es una matriz $p \times n$ y su j-ésima columna contiene las observaciones centradas. $X_j - \overline{X}$.

El método de componentes principales [4], es un método para el ajuste de subespacios lineales o una técnica estadística útil para detectar y describir posibles singularidades en los datos. El interés estará especialmente en las proyecciones de los datos en las coordenadas de las componentes principales correspondientes a los valores propios más pequeños (las últimas filas de Y). En general, con datos p-dimensionales la proyección en la componente principal de menor varianza será pertinente para el estudio de la desviación de una observación a partir del hiperplano ajustado, mientras proyecciones en las q coordenadas $(q < p)$ de las componentes principales con varianza mínima serán de utilidad para el estudio de la desviación de una observación a partir del subespacio lineal ajustado de dimensión $p - q$, permitiendo la detección de observaciones atípicas u *outliers* en la muestra.

Para la detección de observaciones *outliers*, Rao [17] sugirió un método que consiste en calcular la longitud de la perpendicular desde cada punto X_i sobre el mejor subespacio ajustado por medio de las últimas q-componentes principales. El cuadrado de la perpendicular de cada punto $\mathrm{X_i}$, está dada por la expresión:

$$\mathrm{d}_i^2 = \sum_{j=p-q-1}^{p}[P_j'\,(X_i - \bar{X})]^2 = \sum_{j=1}^{p}[P_j'\,[(X_i - \bar{X}]^2 - \sum_{j=1}^{q}[P_j'\,[(X_i - \bar{X}]^2$$

La expresión anterior se simplifica en:

$$d_i^2 = [P(X_i - \bar{X})]'[P(X_i - \bar{X})] - \sum_{j=1}^{q}[P_j'\,[(X_i - \bar{X}]^2$$

y se obtiene finalmente:

$$d_i^2 = (X_i - \bar{X})'(X_i - \bar{X}) - \sum_{i=1}^{q}[P_i'\,[(X_i - \bar{X}]^2 \qquad (3)$$

La expresión anterior es válida para cada punto $X_i. i = 1,2,\dots,n$. Rao propone, como un segundo paso, considerar los valores más grandes de los $\{d_i^2\}\ {}_i^n= 1$ como indicativos de posibles candidatos a *outliers* ya que representan un ajuste pobre en el espacio de dimensión $(p-q)$.

Varios autores sugieren la realización de un análisis preliminar de las componentes principales asociadas a los datos y observar los valores muestrales de la proyección de las observaciones sobre las componentes principales de diferente orden, ya que las primeras componentes principales son las más sensibles a *outliers* afectando las varianzas y las covarianzas (o las correlaciones si el análisis de componentes principales se ha realizado utilizando la matriz de correlación, en vez de la matriz de covarianza muestral), además las últimas componentes son sensibles a *outliers* al adicionar dimensiones espurias a los datos u ocultar singularidades [18].

3.3 Búsqueda de proyecciones

Para el manejo de datos cuya dimensionalidad es alta ha sido implementada *"Projection Pursuit" (P.P),* en español "búsqueda de proyecciones" para atacar diversos problemas donde la dimensión se vuelve crítica a la hora de aplicar técnicas corrientes [18].

El término "búsqueda de proyecciones" es una técnica para el análisis exploratorio de conjuntos de datos multivariables de considerable extensión. Consiste esencialmente en la búsqueda de estructuras interesantes en el conjunto de datos al realizar proyecciones en subespacios de dimensión inferior. Se encarga de seleccionar las proyecciones de mayor interés mediante la optimización de algún índice de proyección (I.P). El índice de proyección se construye de tal forma que cuantifique lo interesante de la proyección [19] [20].

Friedman y Stuetzle [21], suministran una revisión de los antecedentes, la implementación y la aplicación de P.P a datos de gran dimensión. Ver también Correa y Salazar [18] [22].

El algoritmo de búsqueda de proyecciones se puede resumir en seis pasos:

1. Centre y esfere los datos
2. Escoja un índice
3. Encuentre una dirección
4. Proyecte los datos y evalúe el índice
5. Si el índice no es un máximo (o mínimo) regrese a (3)
6. Analice los datos proyectados

1) Centrar y esferar los datos originales: este paso se hace comúnmente antes de proyectar los datos. Los argumentos a favor de este paso son dos. Primero, se presenta

una economía computacional ya que en cualquier dirección que proyectemos los datos las proyecciones tendrán media cero y varianza uno. Segundo, no interesa la información suministrada por la localización o la estructura de covarianzas de los datos, sino de estructuras más complejas, tales como clústers o no linealidades.

Centrando y esferando los datos. Sea Y la matriz de información muestral, efectuemos una descomposición en valores propios de la matriz de covarianzas: z $\sum = E[(Y - EY)(Y - EY)'] = UDU'$

Donde U es una matriz ortonormal y D es una matriz diagonal $p \times p$. Puede definirse:

$$Z = \sum^{-1/2}(Y - EY)$$

Donde $E[Z] = 0$ y $E[ZZ'] = 1$, la matriz identidad. Esta transformación centra y esfera los datos además elimina la estructura de correlación.

2) Escoger un índice: La selección del índice es un paso crítico en el método. El índice indicará si una proyección es interesante o no y si vale la pena estudiar la estructura planteada. Se debe recordar que el procedimiento es automático y solo se tendrá el resultado al final de la búsqueda [18]. La distribución normal está caracterizada completamente por su vector de medias y su matriz de covarianzas. Además, si se tiene una distribución multinormal, cualquier proyección será normal ya que combinaciones lineales de variables aleatorias normales independientes se distribuyen normalmente. Si la proyección menos posible de ser normal es normal, entonces no se necesita buscar más proyecciones. Si se tiene un conjunto de datos con muy alta dimensión, la mayoría de las proyecciones a bajas dimensiones tienden a ser normales. Como argumento final se tiene que entre todas las distribuciones continuas con varianza fija la normal maximiza la entropía. Entonces como índice para medir qué tan atractiva es una proyección se puede escoger una función que mida la entropía de los datos, entre más alejada de la normal sea la proyección, mejor.

3) Seleccionar una dirección: La selección de las direcciones plantea un problema de optimización. El método de optimización dependerá del índice escogido.

4) Proyectar los datos: Este paso se realiza con la dirección seleccionada en el paso 3.

5) Evaluar el índice: Si éste es un máximo, entonces se puede seguir al próximo paso, en caso contrario se debe retornar al paso 3.

6) Analizar la proyección: Eliminar la estructura hallada de los datos y repetir pasos 3 al 6 nuevamente. Luego de haber encontrado algunas proyecciones interesantes, ¿qué se debe hacer?

1. Identifique clústers, aíslelos e investigue cada uno de ellos separadamente.
2. Identifique clústers y localícelos (i.e. reemplácelos por ejemplo por su centro y clasifique los puntos de acuerdo con su pertenencia a un clúster).
3. Busque una descripción tranquila (separe estructura de ruido en una forma no paramétrica).

En la literatura existen una gran diversidad de índices de P.P., tanto en una como en dos dimensiones, propuestos por: Friedman y Tukey [16], Jones y Sibson [23], Yenyukov [24], [25], entre otros [26], [18].

3.4 Un método adaptable

El siguiente es un método desarrollado para identificar valores atípicos en el espacio multivariado desarrollado por P. Filzmoser del Departamento de Estadística y Teoría de la Probabilidad Viena, Austria. El método compara la diferencia entre la distribución empírica de los cuadrados de las distancias robustas y la función de distribución chi-cuadrado. El método propuesto admite diferentes dimensiones de los datos, y también diferentes tamaños de muestra.

La trama de chi-cuadrado es útil para visualizar la desviación de la distribución de los datos de normalidad multivariante en las colas. Este principio se utiliza a continuación. $G_n\ (u)$ denota la función de distribución empírica del cuadrado de las distancias robustas, RD_i^2 y sea $G(u)$ la función de distribución de X_p^2. Para muestras multivariantes distribuidas normalmente, G_n converge a G. Por lo tanto, las colas de G_n y G pueden ser comparados para detectar valores atípicos.

Las colas se definen por $\delta = X_{p;1-\alpha}^2$ para un $\propto$ pequeño (por ejemplo, $\propto = 0{:}025$),), y

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

en donde " + " indica las diferencias positivas. De esta manera, $p_n(\delta)$ mide la salida empírica a partir de la distribución teórica sólo en las colas, definido por el valor de δ. $p_n(\delta)$ se puede considerar como una medida de los valores atípicos en la muestra. Gervini [27] utilizó esta idea como una etapa de ponderación para la estimación robusta de la ubicación y de dispersión multivariante. De esta manera, la eficiencia (en términos de precisión estadística) del estimador podría mejorarse considerablemente.

$p_n(\delta)$ no se utiliza directamente como una medida de los valores atípicos. El umbral debe ser infinito en el caso de una distribución normal multivariante de datos, i.e. valores extremos o valores de la misma distribución no deben ser declarados como valores atípicos. Por lo tanto un valor crítico P_{crit} se introduce, lo que ayuda a distinguir entre los valores atípicos y los extremos. La medida de los valores extremos de la muestra se define como

$$\propto_n \delta) = \begin{cases} o & si\ P_n(\delta) \leq P_{crit}(\delta, n, p) \\ P_n(\delta) & si\ P_n(\delta) > P_{crit}(\delta, n, p) \end{cases}$$

El valor de umbral que se llamará *cuantil ajustado* se determina entonces como $c_n(\delta) = G_n^{-1}(1 - \alpha_n(\delta))$. El valor crítico para distinguir entre los valores atípicos y los extremos se puede derivar por simulación, y el resultado es aproximadamente

$$p_{crit}(\delta, n, p) = \frac{0.24 - 0.003\,.p}{\sqrt{n}} \quad para \quad \delta = X^2_{p,0.975} \quad y \quad p \leq 10$$

Y

$$p_{crit}(\delta, n, p) = \frac{0.252 - 0.0018\,.p}{\sqrt{n}} \ para\ \delta = X^2_{p,0.975} \quad y \quad p > 10$$

(Ver Filzmonser, Reinann, and Garrett [28]).

4. Herramientas para la detección de outliers multivariantes

4.1 R

R es un lenguaje y un entorno para el cálculo estadístico y el dibujo de gráficas. Es un proyecto GNU desarrollado por los laboratorios Bell. Está disponible como software libre bajo el contrato GNU de la *Free Software Foundation*. Se puede compilar y ejecutar en una gran variedad de sistemas UNIX (como FreeBSD y Linux). También en Windows y MacOS siendo una solución integrada para la manipulación de datos, cálculo y generación de gráficas [29] [30].

Cuenta con el paquete "mvoutlier" (Detección de valores atípicos multivariados basados en métodos robustos) el cual fue creado para la detección de valores atípicos multivariados.

Algunas de las funciones más sobresalientes con que cuenta este paquete pueden observarse en la Tabla 1:

Tabla 1. Funciones del paquete "mvoutlier"

Función	Descripción
aq.plot	Gráfico del cuantil ajustado
Arw	Estimador adaptativo reponderado para la localización y dispersión multivariable
Bhorizon	Horizonte B de los Datos de Kola
bss.background	Mapa de fondo para el proyecto BSS
chisq.plot	Gráfico Chi-Cuadrado
color.plot	Color del gráfico
corr.plot	Gráfico de correlación: correlación robusta bivariada versus clásico
dd.plot	Distancia - Gráfico de la distancia
locoutNeighbor	Gráfico de diagnóstico para identificar valores atípicos locales con diferentes tamaños de vecindad
locoutPercent	Gráfico de diagnóstico para identificar valores atípicos locales con un tamaño fijo de vecindad
map.plot	Gráfico de valores atípicos multivariados en un mapa
plot.mvoutlierCoDa	Grafico para la interpretación de los valores atípicos multivariados de Coda

Para obtener mayor información y acceder a la descarga directa del software puede referirse a su sitio web oficial: http://www.r-project.org/

4.2 Ibm spss statistics

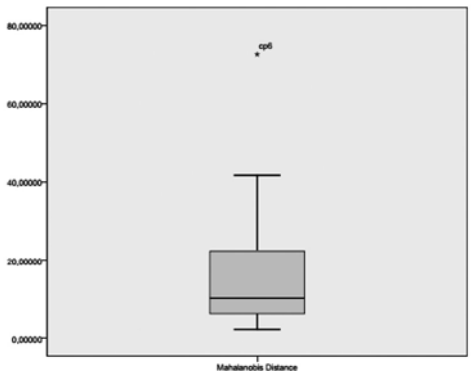
IBM® SPSS® Statistics es un sistema global para el análisis de datos. SPSS Statistics puede adquirir datos de casi cualquier tipo de archivo y utilizarlos para generar informes tabulares, gráficos y diagramas de distribuciones y tendencias, estadísticos descriptivos y análisis estadísticos complejos [31].

La página Web de IBM SPSS Inc (http://www.spss.com) proporciona acceso a algunos archivos de datos y otras informaciones de utilidad [32]. Este software es licenciado y cuenta con la opción que permite identificar valores atípicos multivariados.

Para obtener la distancia de Mahalanobis en un conjunto de datos multivariantes se debe seleccionar la opción de regresiones lineales indicando una variable como dependiente y las demás variables cuantitativas como independientes y marcar la opción Mahalanobis.

Como resultado se obtiene una nueva variable llamada mah_1 con el total de distancias de Mahalanobis de los datos sin valores faltantes al vector media $\bar{x}$ que se encuentre en la vista de datos. También es posible obtener el diagrama de cajas y bigotes de la variable mah_1. La figura 2 muestra un ejemplo de este diagrama donde se puede observar un posible valor atípico multivariante (el dato etiquetado como cp6 que se encuentra fuera de la caja) [33].

Figura 2. Diagrama de caja de la variable mah_1. Tomado de: Análisis de datos, Universidad Autónoma de Madrid. [En línea] Disponible en: http://www.uam.es/personal_pdi/ciencias/dfaraco/docencia/AD/Practica1AD2011.pdf



5. Conclusiones

Se recopiló información en cuanto a valores atípicos multivariantes, dejando plasmado las características generales más sobresalientes como su definición, los problemas o efectos que se presentan en estos tipos de datos (enmascaramiento y empantanamiento) y algunos métodos tradicionales para su detección (distancia de Mahalanobis, componentes principales, búsqueda de proyecciones) y un método adaptable para identificar valores atípicos).

En la actualidad existen herramientas informáticas que permiten realizar el análisis de grandes cantidades de información facilitando para el caso de los valores atípicos multivariantes identificar de manera más sencilla las correlaciones entre las variables.

R y SPSS Statistics son dos de las herramientas más utilizadas que ofrecen un completo análisis multivariante de los datos, bien sea el primero de estos a través de su consola de comandos o con SPSS de una manera más gráfica y visual para el usuario.

6. Referencias

[1] D.M. Hawkins, *"Identification of Outliers"*. London, Chapman & Hall. 1980.

[2] R. J Beckamn. R. D Cook, *"Outlier.....s"*, Technometrics Vol 25, No. 2. pp 119-149. 1983.

[3] V. Barnett, T. Lewis, *"Outliers in statistical data"*, 3rd edition. Chichester, John Wiley & Sons, 1994, 584 pp.

[4] R. Gnanadesikan, J. R. Kettenring, "*Robust Estimates Residuals and Outlier Detection with Multiresponse Data"*. Biometrics. Vol 28, pp 81-124. 1972.

[5] M. Antonio, D. Bárbara. "*El problema de los outliers multivariantes en el análisis de sectores clave y cluster industrial*". [En línea]. Disponible en: http://www.unizar.es/jornadasio-zaragoza/archivos/pdf/Ponencia_Morillas_Antonio.pdf.

[6] J. Peat, B. Barton, Medical Statistics: *"A guide to data analysis and critical appraisal"*. Blackwell Publishing. 2005

[7]] I. Ben-Gal, "*Outlier detection*", In: Maimon O. and Rockach L. (Eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005

[8] M. Quaglino, J. Merello, Métodos multivariados en estudios de vulnerabilidad social en la provincia de santa fe. Noviembre del 2012. [En línea]. Disponible en: http://www.fcecon.unr.edu.ar/web-nueva/sites/default/files/u16/Decimocuartas/quaglino_merello_metodo_multivariados_en_estudios.pdf

[9] E. Acuna, C. A. Rodriguez, (2004) *"Meta analysis study of outlier detection methods in classification*". [En línea] Disponible en: http://academic.uprm.edu/eacuna/paperout.pdf

[10] B. Iglewics, J. Martinez, *"Outlier Detection using robust measures of scale"*, Journal of Sattistical Computation and Simulation, Vol. 15, No 4. pp. 285-293, 1982.

[11] L. Davies, U. Gather, *"The identification of multiple outliers,"* Journal of the American Statistical Association, Vol. 88. No 423, pp 782-792, 1993.

[12] S. Matsumoto et al.*" Comparison of Outlier Detection Methods in Faultproneness Models"*. En: Proceedings of the First international Symposium on Empirical Software Engineering and Measurement ESEM 2007 (Madrid, España, Septiembre 20 - 21, 2007). IEEE Computer Society, Washington, DC, 461-463.

[13] K. Tiwary et. al. "*Selecting the Appropriate Outlier Treatment for Common Industry"*. SAS Conference Proceedings: NESUG 2007. Noviembre 11-14, 2007, Baltimore, Maryland.

[14] G.J. McLachlan, Mahalanobis Distance. Resonace. June 1999

[15] Quaglino, Marta, Merello, Juliana*, "Métodos multivariados en estudios de vulnerabilidad social en la provincia de santa fe"* 2012

[16] R.A. Johnson, D. W. Wichern, "*Applied Multivariate Statistical Analysis"*. Prentice Hall International. Inc. 2-Edicion. 1998

[17] C.R. Rao, *"The use and interpretation of principal components analysis in applied research"*. Sankhya. A 26, 1964. pp. 329-358.

[18] L. Victor, "Detección de outliers multivariables mediante projection pursuit," M.S. tesis, Universidad Nacional de Colombia, Seccional Medellin, 1999. [En línea]. Disponible en: http://www.bdigital.unal.edu.co/1495/1/15383124.1999.pdf

[19] J.H. Friedman, J. W. Turkey, *"A Projection Pursuit Algorithm for Exploratory Data Analysis"*. IEEE Transactions on Computers. Vol. C-23, No. 9, pp.881-890. 1974

[20] J.H. Friedman,*" Exploratory Projection Pursuit"*. Journal of the American Statistical Association. Vol. 82, No. 397, pp. 249-266. 1987

[21] J. H. Friedman, W. Stuetzle, *"Projection Pursuit for Data Analysis"*. Modern Data Analysis. Pp. 123-147. 1982

[22] J. C. Correa, J. C. Salazar. *"¿Qué es Projection Pursuit?"*.Enviado a publicación revista Estadística de Colombia. 1997

[23] 1M. C. Jones, R. Sibson, *"What is Projection Pursuit?"* Journal Royal Statist, Soc. Series A, Part 1, pp. 1-36. 1987

[24] I.S. Yenyukov, *"Detecting Structures by Means of Projection Pursuit"*. International Association for Statistical Computing. pp 47-58. 1989

[25] I.S. Yenyukov, "*Indeces For projection Pursuit Data Analysis"*. Learning Symbolic and Numerical Knowlege. E. Diday, ed. IMRIA. New York. pp. 181-188. 1989

[26] V.I. López, (1999) *"Deteccion de Outliers Multivariables medinate Projection Pursuit"*. Universidad Nacional de Colombia- Seccional Medellin. [En línea]. Disponible en: http://www.bdigital.unal.edu.co/1495/1/15383124.1999.pdf

[27] D. Gervini, *"A robust and efficient adaptive reweighted estimator of multivariate location and scater"*. Journal of multivariate analysis. Vol. 84, pp. 116-144. 2003

[28] P. Filzmoser, C.Reimann, R.G. Garrett. *"Multivariate outlier detection in exploration geochemistry"*. Computers & Geosciences. Vol. 31. pp 579–587. 2005

[29] Centro informático cientifico de Andalucia. R. [Em línea]. Disponible en: http://www.cica.es/Software/r.html

[30] The R Project for Statistical Computing. [En línea]. Disponible en: http://www.r-project.org/

[31] IBM, *Manual del usuario del sistema básico de IBM SPSS Statistics 20.* [En línea]. Disponible en: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Core_System_Users_Guide.pdf

[32] *IBM, SPSS software.* [En línea]. Disponible en: http://www-01.ibm.com/software/analytics/spss/

[33] *Análisis de Datos.* Universidad Autónoma de Madrid. [En línea]. Disponible en: http://www.uam.es/personal_pdi/ciencias/dfaraco/docencia/AD/Practica1AD2011.pdf