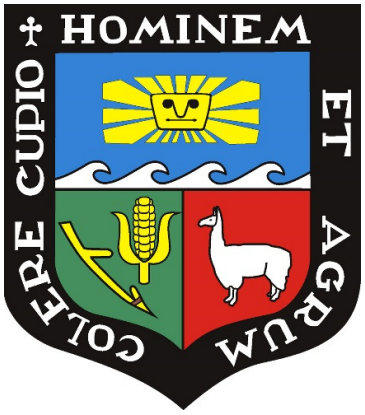
UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

ESCUELA DE POSGRADO
MAESTRÍA EN ESTADÍSTICA APLICADA



"DETECCIÓN DE DATOS MULTIVARIADOS ATÍPICOS CON SERIES FINITAS DE FOURIER"

Presentada por:
JORGE LUIS RUBIO DONET

TESIS PARA OPTAR EL GRADO DE MAGISTER SCIENTAE EN ESTADÍSTICA APLICADA

Lima – Perú
2018

UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA

ESCUELA DE POSGRADO
MAESTRÍA EN ESTADÍSTICA APLICADA

"DETECCIÓN DE DATOS MULTIVARIADOS ATÍPICOS CON SERIES FINITAS DE FOURIER"

TESIS PARA OPTAR EL GRADO DE MAGISTER SCIENTIAE EN ESTADÍSTICA APLICADA

Presentada por:

JORGE LUIS RUBIO DONET

Sustentada y aprobada ante el siguiente jurado:

Mg. Raphael Valencia Chacón	Mg. Sc. Fernando Miranda Villagómez
PRESIDENTE	PATROCINADOR
Mg. Jesús Salina Flores	Mg. Sc. Jaime Porras Cerrón
MIEMBRO	MIEMBRO

Dedicado a Charo,

Luis Fernando y

Ximena.

Los verdaderos motores de mi vida

AGRADECIMIENTOS

Mi más sincero agradecimiento a mi patrocinador de tesis y amigo Fernando Miranda Villagómez.

A mi jurado por sus valioso aportes y sugerencias que permitieron enriquecer este documento.

A la Universidad Nacional Agraria La Molina que formó mi estilo de trabajo como alumno de pregrado, alumno de posgrado, y docente. Quedo convencido que ser Molinero es ser un profesional diferente

RESUMEN

La presencia de observaciones atípicas en un conjunto de datos es una de las causas que generan distorsiones en el análisis. La detección de dichas observaciones puede ayudar a una correcta evaluación de las tendencias en el comportamiento de los datos.

Para el caso de datos multivariados se han desarrollado diversos métodos que permiten la detección de comportamientos atípicos, basados en métodos gráficos, y otros asumiendo una distribución normal multivariada. No obstante, en muchos casos el supuesto de normalidad multivariada no se cumple.

El presente trabajo propone una prueba no paramétrica basada en la aplicación del método Bootstrap, utilizando como indicador de similitud a las distancias entre las representaciones obtenidas con series finitas de Fourier, propuesta por Andrews. El método propuesto permite detectar datos multivariados atípicos, combinando la significación estadística de la prueba Bootstrap y el análisis gráfico sugerido por Andrews, y que puede ser también aplicado a datos medidos en una escala ordinal.

El método fue aplicado a cuatro conjuntos de datos, encontrando resultados satisfactorios en todos los casos.

Palabras clave: Observaciones atípicas, Gráficos de Andrews, Bootstrap. Series finitas Fourier

ABSTRACT

The presence of atypical observations in a dataset is one of the causes that generate distortions in the analysis. The detection of these observations can help to evaluate the trends in the behavior of the data.

In the case of multivariate data several methods have been developed that allow the detection of atypical behaviors, based on graphical methods, and others assuming a normal multivariate distribution. However, in many cases the assumption of multivariate normalcy is not fulfilled.

This paper proposes a non-parametric test based on the application of Bootstrap method, using as an indicator of similarity to the distances between the representations obtained with finite series of Fourier, proposed by Andrews. The proposed method allows the detection of atypical multivariate data, combining the statistical significance of the Bootstrap test and the graphical analysis suggested by Andrews, which can be applied to data measured on an ordinal scale.

The method was applied to four sets of data, finding satisfactory results in all cases.

Keywords: Multivariate outliers, Andrews plots, Bootstrap, finite series of Fourier.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

I. INTRODUCCIÓN

En todo análisis estadístico la base que lo sustenta descansa en los datos que serán utilizados en el estudio. La importancia de la calidad de los datos es fundamental en los resultados que se obtengan del estudio. Lo que usualmente un investigador se cuestiona es si sus datos confiables, de calidad, y carentes de ruido. Gran parte del tiempo que toma realizar un análisis estadístico es dedicado a la preparación de datos y a la detección de probables inconsistencias que pudieran alterar los indicadores estadísticos que serán utilizados en la investigación. En el análisis de las tendencias del comportamiento de variables es frecuente encontrar observaciones que se alejan de dichas tendencias y ocasionan que los indicadores estadísticos se vean distorsionados. Siendo, en estos casos muy importante evaluar de alguna manera la influencia que puedan tener dichas observaciones sobre los resultados obtenidos a partir de una muestra. En la actualidad existe una diversidad de métodos dirigidos a la identificación de valores atípicos. Gran parte de ellos consisten en métodos gráficos y pruebas estadísticas para evaluar si algunos datos están demasiado alejados de una tendencia con lo cual serían considerados como datos atípicos.

Los métodos propuestos para la detección de datos atípicos son desarrollados en su gran mayoría para un análisis univariado. Es decir, dentro de los datos observados para una variable se trata de detectar aquellos valores que están muy alejados de la tendencia usual. Programas estadísticos como Minitab tienen métodos de análisis exploratorio e inferenciales que permiten detectar tendencias de datos anómalos. Destacan los gráficos de cajas y las pruebas de Dixon (1950) y Grubbs (1969). Estos métodos propuestos son simples e intuitivos que resultan eficientes para el estudio de una sola variable.

Para el caso de datos multivariados también se han elaborado métodos que permiten la detección de comportamientos atípicos, algunos basados en métodos gráficos como el propuesto por Andrews (1972) y otros sustentados mediante pruebas de hipótesis como el propuesto por Canori y Prescott (1992). Muchos de estos métodos parten del supuesto de normalidad multivariada. En estos casos se tiene que verificar primero el cumplimiento de dicho supuesto.

Un aspecto importante en un análisis exploratorio de datos consiste en decidir cuándo una observación multivariada puede ser considerada que se encuentra alejada a lo observado en la tendencia usual del comportamiento de una muestra. Varias propuestas se han realizado al respecto, siendo muchas de ellas basadas en apreciaciones subjetivas; es decir, se sustentan en la apreciación particular del investigador.

La presente investigación propuso utilizar las proyecciones de las observaciones multivariadas sugeridas por Andrews (1972) como un indicador de similitud entre dos observaciones multivariadas. Es decir, se planteó hacer la comparación de las observaciones multivariadas a través de las distancias entre las representaciones gráficas obtenidas mediante series finitas de Fourier. Para ello, se estableció alcanzar los siguientes objetivos:

- Presentar un método alternativo para comparar la similitud de observaciones multivariadas evaluando las distancias entre las representaciones gráficas de Andrews.
- Utilizar el método "Bootstrap" para evaluar si existe alguna observación multivariada que presenta un comportamiento que difiere significativamente de la tendencia usual que muestran los datos.

Para evaluar si una observación está significativamente alejada de la tendencia usual se planeó utilizar el método Bootstrap para analizar la semejanza de las distancias y de esta manera desarrollar un procedimiento no paramétrico que no requiera del supuesto de normalidad multivariada de los datos recopilados.

Para la aplicación de los procedimientos propuestos se elaboraron funciones en el lenguaje R que es parte de la presente publicación para uso de los interesados.

Para probar los procedimientos se han utilizado tres conjuntos de datos simulados, en los cuales se adicionan observaciones atípicas, y además se hace uso de un conjunto de datos tomados del Ministerio de Trabajo y Promoción del Empleo.

II. REVISIÓN DE LITERATURA

2.1 Datos atípicos

El análisis de datos está involucrado a un proceso en el cual las observaciones recopiladas son resumidas en indicadores que brindan pautas sobre el comportamiento de una realidad que supuestamente es captada en una muestra aleatoria representativa. En este proceso resulta frecuente que se tengan datos que se alejan de la tendencia usual de las observaciones de la muestra. A estos datos se les conoce como datos atípicos, datos extremos, o outliers.

Grubbs (1969) define una observación fuera de lo común, llamada más comúnmente como observación atípica, a aquellos datos que parecen desviarse demasiado de las otras observaciones de la muestra. Asimismo, Grubbs (1969) también señala que un dato extremo puede ser el resultado de una manifestación extrema de la variabilidad aleatoria inherente a los datos, y que por tanto los datos deberían mantenerse en el proceso de análisis. Hace una distinción con aquellos casos donde los datos extremos son el resultado de significativas diferencias generadas por errores experimentales, errores de cálculo, errores de registro de datos. En esto casos es conveniente investigar las razones de dichas diferencias y la posibilidad de no incluir a los valores atípicos o extremos en el análisis.

Para Ben-Gal (2005), en el análisis de un conjunto de datos una definición exacta de un valor atípico a menudo depende de suposiciones ocultas con respecto a la estructura de datos y el método de detección aplicado. Asimismo, cita las siguientes definiciones de observaciones atípicas o outliers:

- Hawkins, define un outlier como una observación que se desvía tanto de otras observaciones como para despertar la sospecha de que fue generado por un mecanismo diferente.
- Barnett & Lewis, indican que una observación atípica, o outlier, es aquella que parece desviarse demasiado de otros miembros de la muestra en la que ocurre.

- Johnson, define un valor atípico como una observación en un conjunto de datos que parece ser inconsistente con el resto de ese conjunto de datos.

Para el caso multivariado, Muñoz García & Amón Uribe (2013) citan las siguientes definiciones propuestas por diversos investigadores:

- Gnanadesikan y Kettenring, los outliers multivariados son observaciones que se consideran extrañas no por el valor que toman en una determinada variable, sino en el conjunto de aquellas.
- Peat y Barton, definen un valor atípico multivariado, como un caso que es un valor extremo para una combinación de variables.

Ben-Gal (2005) afirma que en muchos casos las observaciones multivariadas no pueden ser detectadas como valores extremos cuando cada variable se considera de forma independiente. La detección de outliers sólo es posible cuando se realiza un análisis multivariado y las interacciones entre las diferentes variables se comparan dentro de la clase de datos.

Ahora, en el análisis de la presencia de observaciones multivariadas atípicas es necesario tener presente varios aspectos, tales como los problemas de la dimensión, y sobre todo, como lo señalan Acuña & Rodriguez (2004), Barnett & Lewis (1994), Ben-Gal (2005), Beckman & Cook (1983), y Muñoz García & Amón Uribe (2013) es conveniente tener en cuenta los siguientes 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.

2.2 Representación gráfica de observaciones multivariadas

En un análisis multivariado el gran reto es entender el comportamiento conjunto de las variables incluidas en una investigación, para decidir si una observación puede o no ser considerada como atípica. Una de las formas de realizar el análisis es mediante gráficos que representen dicha información. Buja (1996) señala que la representación gráfica de datos multivariados tiene el compromiso de producir una imagen estática que resuma sus características, y que la decisión principal consiste en elegir el tipo de gráfico que puede servir como punto de partida. Para esto se debe tener las siguientes consideraciones:

- *Diagramas de dispersión*, donde los casos son representados por la localización de puntos.
- *Trazas*, donde los casos son representados como funciones de un parámetro real, tal como los gráficos de coordenadas paralelas, y las curvas de Andrews (1972), y
- *Figuras*, donde los casos son representados como símbolos complejos cuyas características son funciones de los datos, tales como árboles, castillos, estrellas, rostros de Chernoff, etc.

Asimismo, como lo indica Moustafa (2009), el descubrimiento de patrones visuales en grandes conjuntos de datos multivariantes es un problema difícil en los campos de la minería de datos y el análisis exploratorio de datos. Esto es debido, en parte, al problema aglomeración que resulta de poner en un solo gráfico muchas representaciones. Esto desafía a la mayoría de las técnicas de visualización de la información en general, y en particular a las técnicas de coordenadas paralelas, y además, la aglomeración gráfica aumenta con el número de observaciones muestrales haciendo más difícil la detección de grupos, tendencias, correlaciones, periodicidad, y anomalías. Moustafa (2009) hace mención a las técnicas de coordenadas paralelas desarrolladas por Inselberg (1985), Wegman (1990), y Andrews (1972), en las cuales la representación de un número elevado de datos hace que el problema de aglomeración se acentúe en gran medida.

Moustafa (2009) define el gráfico de coordenadas paralelas generalizado como una transformación de un espacio p-dimensional, producto de la observación de "p" variables, a un espacio bidimensional mediante funciones básicas que permitan obtener una representación propia para cada observación multivariada., que se describe como sigue:

Sea $\boldsymbol{x}_i = \left(x_{i1}, x_{i2}, \cdots, x_{ip}\right)$ una observación multivariada de "p" variables, y sea $\boldsymbol{\phi}_t = \left(\phi_{1t}, \phi_{2t}, \cdots, \phi_{pt}\right)$ un vector básico, para i=1,2,. . ., m, donde "m" corresponde al número de observaciones por analizar. Si se calcula el producto interno de éstos vectores se obtiene la función:

$$\rho_{it} = \sum_{j=1}^{p} x_{ij}\phi_{jt} = \boldsymbol{x}_i\boldsymbol{\phi}_t^T, \qquad i = 1, 2, \cdots, m$$

La cual puede ser también escrita como: $\rho_t = \boldsymbol{X}\boldsymbol{\phi}_t^T$, donde $\rho_t = \left(\rho_{1t}, \rho_{2t}, \cdots, \rho_{mt}\right)^T$. Esto viene a ser el producto interno de cada observación con el vector rotado $\boldsymbol{\phi}_t^T$. (un vector que cambia de orientación con cada valor de "t"). Andrews (1972) propuso una función $\boldsymbol{\phi}_t^T$ basada en series finitas de Fourier.

Muchos métodos han sido propuestos para obtener representaciones que intentan sintetizar en un gráfico los datos multivariados correspondientes a un individuo. Fienberg (1979) hace mención a varios métodos gráficos alternativos. Andrews (1972) publicó su artículo: "Plots of high-dimensional data", en el cual propuso representar cada dato k-variado $\boldsymbol{x} = (x_{1,}x_{2,}\cdots, x_{k,})$ mediante la siguiente función:

$$f_{\boldsymbol{x}}(t) = \frac{x_1}{\sqrt{2}} + x_2 sen(t) + x_3 cos(t) + x_4 sen(2t) + x_5 cos(2t) + \cdots$$

Con los gráficos obtenidos para valores $-\pi < t < \pi$, Andrews propone hacer una comparación de semejanza, indicando que observaciones cercanas deberán generar representaciones gráficas semejantes.

Goodchild & Vijakan (1974), Rubio (1983) hacen notar que cuando el número de variables es par las variancias se reducen a una constante, ya que no depende de "t". No obstante, en el caso general la dependencia en "t" no es grande, y dado que $f_{\mathbf{x}}(t)$ es univariada, puede

utilizarse la prueba T de Student con los grados de libertad correspondiente a la matriz de covariancias W obtenida de las variables X.

Semmar et al. (2008) hacen ver la utilidad de las representaciones gráficas de Andrews en la identificación de datos atípicos multivariados, ya que gráficos semejantes y compactos indican una fuerte estructura de grupo, y para aquellos casos en los que se tiene una observación multivariada atípica se genera una representación alejada que puede ser identificada visualmente con facilidad.

Embrechts & Herzberg (1991) proponen algunas variaciones a la propuesta por Andrews, y sugiere evaluar las alternativas: hacer un proceso de cambio de escala y ordenamiento; utilizar variables estandarizadas; hacer uso de polinomios de Chebyshev, o polinomios de Legendre. Asimismo, indica que los gráficos de Andrews son una herramienta muy útil en el análisis exploratorio de datos, en particular en lo que se refiere a la representación gráfica de observaciones multivariadas.

Maravelakis & Bersimis (2009) hacen mención a diversas propuestas para obtener representaciones gráficas que constituyen variantes a la propuesta de Andrews $f_x(t) = \boldsymbol{a}(t)'\boldsymbol{x}$, considerando:

$$\boldsymbol{a}(t) = (sen(2t),\ cos(2t),\ sen(4t),\cdots)$$
$$\boldsymbol{a}(t) = (1, sen(t)+cos(t),\ sen(t)-cos(t),\ sen(2t)+cos(2t),\ sen(2t)-cos(2t),\cdots)$$

No obstante, la ventaja de la propuesta de Andrews radica en las propiedades estadísticas asociadas a la función.

2.3 Series Finitas de Fourier

El método sugerido por Andrews (1972) propone resumir en gráficos bidimensionales la información correspondiente a "*k*" variables, captada para un conjunto de "*n*" elementos a los cuales se desea clasificar. Es decir, a partir de la matriz de información

$$\mathbf{X} = (\mathbf{x}_i) = (x_{ij}), \qquad i = 1, 2, \cdots, n; \quad j = 1, 2, \cdots, k$$

y considerando el siguiente vector móvil:

$$\boldsymbol{a}(t)=\left[\frac{1}{\sqrt{2}}, sen(t),\ cos(t),\ sen(2t),\ cos(2t),\ \cdots\right]', \qquad -\pi \le t \le \pi \tag{1}$$

El vector que proyecta al vector $\mathbf{x}_i$ en la dirección del vector $\boldsymbol{a}(t)$ es:

$$\begin{aligned}\text{Proy}_{\boldsymbol{a}(t)}[\mathbf{x}_i] &= \frac{\mathbf{x}_i\boldsymbol{a}(t)}{\|\boldsymbol{a}(t)\|^2}\boldsymbol{a}(t)\\ &= \frac{\mathbf{x}_i\boldsymbol{a}(t)}{\|\boldsymbol{a}(t)\|}\frac{\boldsymbol{a}(t)}{\|\boldsymbol{a}(t)\|}\\ &= f_{\mathbf{x}_i}(t)\frac{\boldsymbol{a}(t)}{\|\boldsymbol{a}(t)\|^2}\end{aligned}$$

La longitud y sentido de la proyección del vector $\mathbf{x}_i$ sobre el vector $\boldsymbol{a}(t)$ es:

$$\text{Comp}_{\boldsymbol{a}(t)}[\mathbf{x}_i] = \frac{\mathbf{x}_i\boldsymbol{a}(t)}{\|\boldsymbol{a}(t)\|} = \frac{f_{\mathbf{x}_i}(t)}{\|\boldsymbol{a}(t)\|} \tag{2}$$

Puesto que todas las proyecciones de los vectores $\mathbf{x}_i$ están afectadas por la misma constante $\|\boldsymbol{a}(t)\|$; luego, dichas proyecciones pueden ser estudiadas mediante la comparación de los productos escalares:

$$f_{\mathbf{x}_i}(t) = \mathbf{x}_i\boldsymbol{a}(t) = \frac{x_1}{\sqrt{2}} + x_2 sen(t) + x_3 cos(t) + x_4 sen(2t) + x_5 cos(2t) + \cdots \tag{3}$$

Esta función propuesta por Andrews (1972) permite generar una representación bidimensional $(t, f_{\mathbf{x}_i}(t))$ de cada observación multivariada: Si las observaciones están cerca una de otra se obtendrán representaciones semejantes y cercanas, y si las observaciones están lejos una de otra se obtendrán representaciones diferentes y/o lejanas. Luego, mediante la comparación de dichas representaciones es posible detectar semejanzas entre las observaciones multivariadas, presencia de datos atípicos, y otras peculiaridades que de otro

modo sería difícil de detectar. Por otro lado, Andrews (1979) y Rubio (1983) señalan que las proyecciones tienen las siguientes propiedades:

Propiedad 1

Las representaciones de las funciones preservan las distancias de las observaciones multivariadas.

El cuadrado de la distancia euclidiana entre las representaciones de dos observaciones multivariadas se define como:

$$\begin{aligned}\left\| f_{\mathbf{x}_i}(t) - f_{\mathbf{x}_k}(t) \right\|_E^2 &= \int_{-\pi}^{\pi} \left[f_{\mathbf{x}_i}(t) - f_{\mathbf{x}_k}(t) \right]^2 dt \\ &= \int_{-\pi}^{\pi} \left[\mathbf{x}_i \boldsymbol{a}(t) - \mathbf{x}_k \boldsymbol{a}(t) \right]^2 dt \\ &= \left(\mathbf{x}_i - \mathbf{x}_k \right)' \left[\int_{-\pi}^{\pi} \boldsymbol{a}(t)' \boldsymbol{a}(t) dt \right] \left(\mathbf{x}_i - \mathbf{x}_k \right)\end{aligned}$$

Es decir,

$$\left\| f_{\mathbf{x}_i}(t) - f_{\mathbf{x}_k}(t) \right\|_E^2 = \left(\mathbf{x}_i - \mathbf{x}_k \right)' \left[\int_{-\pi}^{\pi} \boldsymbol{a}(t)' \boldsymbol{a}(t) dt \right] \left(\mathbf{x}_i - \mathbf{x}_k \right) \tag{4}$$

Ahora, del vector definido en (1) se tiene:

$$\boldsymbol{a}(t)'\boldsymbol{a}(t) = \left[\frac{1}{\sqrt{2}}, sen(t),\ cos(t),\ sen(2t),\ cos(2t),\ \cdots \right]' \begin{bmatrix} 1/\sqrt{2} \\ sen(t) \\ cos(t) \\ \vdots \end{bmatrix}$$

$$= \begin{bmatrix} 1/2 & \frac{1}{\sqrt{2}} sen(t) & \frac{1}{\sqrt{2}} \cos(t) & \frac{1}{\sqrt{2}} sen(2t) & \cdots \\ \frac{1}{\sqrt{2}} sen(t) & sen^2(t) & sen(t)\cos(t) & sen(t)sen(2t) & \cdots \\ \frac{1}{\sqrt{2}} \cos(t) & sen(t)\cos(t) & \cos^2(t) & sen(t)\cos(t) & \cdots \\ \frac{1}{\sqrt{2}} sen(2t) & sen(t)sen(2t) & sen(t)\cos(t) & sen^2(2t) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Con lo cual se obtiene:

$$\left\| f_{\mathbf{x}_i}(t) - f_{\mathbf{x}_k}(t) \right\|_E^2 = \left(\mathbf{x}_i - \mathbf{x}_k\right)' \left[\int_{-\pi}^{\pi} \boldsymbol{a}(t)'\boldsymbol{a}(t)dt \right] \left(\mathbf{x}_i - \mathbf{x}_k\right)$$

$$\int_{-\pi}^{\pi} \boldsymbol{a}(t)'\boldsymbol{a}(t)dt = \begin{bmatrix} 1/2\int_{-\pi}^{\pi} dt & \frac{1}{\sqrt{2}}\int_{-\pi}^{\pi} sen(t)dt & \frac{1}{\sqrt{2}}\int_{-\pi}^{\pi} cos(t)dt & \frac{1}{\sqrt{2}}\int_{-\pi}^{\pi} sen(2t)dt & \cdots \\ \frac{1}{\sqrt{2}}\int_{-\pi}^{\pi} sen(t)dt & \int_{-\pi}^{\pi} sen^2(t)dt & \int_{-\pi}^{\pi} sen(t)cos(t)dt & \int_{-\pi}^{\pi} sen(t)sen(2t)dt & \cdots \\ \frac{1}{\sqrt{2}}\int_{-\pi}^{\pi} cos(t)dt & \int_{-\pi}^{\pi} sen(t)cos(t)dt & \int_{-\pi}^{\pi} cos^2(t)dt & \int_{-\pi}^{\pi} cos(t)sen(2t)dt & \cdots \\ \frac{1}{\sqrt{2}}\int_{-\pi}^{\pi} sen(2t)dt & \int_{-\pi}^{\pi} sen(t)sen(2t)dt & \int_{-\pi}^{\pi} cos(t)sen(2t)dt & \int_{-\pi}^{\pi} sen^2(2t)dt & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

$$\int_{-\pi}^{\pi} \boldsymbol{a}(t)'\boldsymbol{a}(t)dt = \begin{bmatrix} \pi & 0 & 0 & 0 & \cdots \\ 0 & \pi & 0 & 0 & \cdots \\ 0 & 0 & \pi & 0 & \cdots \\ 0 & 0 & 0 & \pi & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} = \pi\, \mathbf{I}$$

De donde se deduce que:

$$\begin{aligned} \left\| f_{\mathbf{x}_i}(t) - f_{\mathbf{x}_k}(t) \right\|_E^2 &= \left(\mathbf{x}_i - \mathbf{x}_k\right)' \left[\int_{-\pi}^{\pi} \boldsymbol{a}(t)'\boldsymbol{a}(t)dt \right] \left(\mathbf{x}_i - \mathbf{x}_k\right) \\ &= \left(\mathbf{x}_i - \mathbf{x}_k\right)' [\pi\, \mathbf{I}] \left(\mathbf{x}_i - \mathbf{x}_k\right) \\ &= \pi \left(\mathbf{x}_i - \mathbf{x}_k\right)' \left(\mathbf{x}_i - \mathbf{x}_k\right) \\ &= \pi \left\| \mathbf{x}_i - \mathbf{x}_k \right\|_E^2 \end{aligned}$$

Por consiguiente:

$$\left\| f_{\mathbf{x}_i}(t) - f_{\mathbf{x}_k}(t) \right\|_E = \sqrt{\pi} \left\| \mathbf{x}_i - \mathbf{x}_k \right\|_E \qquad (5)$$

Basados en esta igualdad se deduce que la comparación de las similitudes entre dos observaciones multivariadas se puede realizar, sin pérdida de información, a través de una comparación de sus representaciones mediante las series finitas de Fourier de la forma (3)

Propiedad 2

Las representaciones de las funciones corresponden a una serie de proyecciones unidimensionales.

Como se mostró en la ecuación (2), la longitud y sentido de la proyección del vector $\mathbf{x}_i$ sobre el vector $\boldsymbol{a}(t)$ es

$$\text{Comp}_{\boldsymbol{a}(t)}[\mathbf{x}_i] = \frac{\mathbf{x}_i \boldsymbol{a}(t)}{\|\boldsymbol{a}(t)\|} = \frac{f_{\mathbf{x}_i}(t)}{\|\boldsymbol{a}(t)\|}$$

Puesto que todas las proyecciones de los vectores $\mathbf{x}_i$ están afectadas por la misma constante $\|\boldsymbol{a}(t)\|$; luego, las magnitudes de las proyecciones pueden ser comparadas por medio de la comparación de las representaciones mediante las series finitas de Fourier de la forma (3).

Propiedad 3

Las representaciones de las funciones preservan promedios.

$$\mathrm{E}\left[f_{\mathbf{x}_i}(t) \right] = \mathrm{E}\left[\mathbf{x}_i \boldsymbol{a}(t) \right] = \mathrm{E}\left[\mathbf{x}_i \right] \boldsymbol{a}(t) = \boldsymbol{\mu}\, \boldsymbol{a}(t) = f_{\boldsymbol{\mu}}(t) \qquad (6)$$

$$\hat{\mathrm{E}}\left[f_{\mathbf{x}_i}(t) \right] = \hat{\boldsymbol{\mu}}\, \boldsymbol{a}(t) = f_{\hat{\boldsymbol{\mu}}}(t) = f_{\bar{X}}(t) = \underline{\bar{\boldsymbol{X}}}\, \boldsymbol{a}(t) \qquad (7)$$

Considerando que

$$\bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_{i,k}$$

Se deduce que:

$$\begin{aligned}
f_{\bar{X}}(t) &= \bar{\mathbf{x}}\, \boldsymbol{a}(t) = \frac{\bar{x}_1}{\sqrt{2}} + \bar{x}_2 sen(t) + \bar{x}_3 cos(t) + \bar{x}_3 sen(2t) + \cdots \\
&= \frac{1}{\sqrt{2}} \left(\frac{1}{n} \sum_{i=1}^{n} x_{i,1} \right) + \left(\frac{1}{n} \sum_{i=1}^{n} x_{i,2} \right) sen(t) + \left(\frac{1}{n} \sum_{i=1}^{n} x_{i,3} \right) cos(t) + \left(\frac{1}{n} \sum_{i=1}^{n} x_{i,4} \right) sen(2t) + \cdots \\
&= \frac{1}{n} \left[\left(\sum_{i=1}^{n} x_{i,1} \Big/ \sqrt{2} \right) + \left(\sum_{i=1}^{n} x_{i,2} \right) sen(t) + \left(\sum_{i=1}^{n} x_{i,3} \right) cos(t) + \left(\sum_{i=1}^{n} x_{i,4} \right) sen(2t) + \cdots \right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{\sqrt{2}} + x_{i,2} sen(t) + x_{i,3} cos(t) + {}_{i,4}\, sen(2t) + \cdots \right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[f_{\mathbf{x}_i}(t) \right] \\
&= \frac{1}{n} \sum_{i=1}^{n} f_{\mathbf{x}_i}(t)
\end{aligned}$$

De esta propiedad se deduce que la representación gráfica del centro de gravedad de un grupo de elementos se puede obtener con el promedio de las observaciones de los elementos que conforman el grupo.

Propiedad 4

Las representaciones de las funciones preservan variancias.

$$\mathrm{Var}\left[f_{\mathbf{x}_i}(t)\right] = \mathrm{Var}\left[\mathbf{x}_i\boldsymbol{a}(t)\right] = \boldsymbol{a}(t)'\mathrm{Var}\left[\mathbf{x}_i\right]\boldsymbol{a}(t) = \boldsymbol{a}(t)'\mathbf{V}\boldsymbol{a}(t)$$

Un caso particular se presenta cuando las variables x son incorrelacionadas. Esto que se lograría si se trabaja con los componentes principales en lugar de las variables originales. Si aún más se asume que las variancias de todas las variables son iguales se tiene que:

$$\begin{aligned}\mathrm{Var}\left[f_{\mathbf{x}_i}(t)\right] &= \mathrm{Var}\left[\mathbf{x}_i\boldsymbol{a}(t)\right] = \boldsymbol{a}(t)'\mathrm{V}\boldsymbol{a}(t) = \boldsymbol{a}(t)'\sigma^2\mathrm{I}\,\boldsymbol{a}(t)\\ &= \sigma^2\boldsymbol{a}(t)'\boldsymbol{a}(t)\\ &= \sigma^2\left[\frac{1}{2} + sen^2(t) + cos^2(t) + sen^2(2t) + \cdots\right]\end{aligned}$$

Si "v" denota el número de variables x que están estudiando, se deduce que:

$$\begin{aligned}\mathrm{Var}\left[f_{\mathbf{x}_i}(t)\right] &= \sigma^2\left[\frac{1}{2} + \frac{v-1}{2}\right] \quad , \quad si\ \ v\ \ es\ \ impar\\ &= \sigma^2\left[\frac{v-1}{2} + sen^2\left(vt/2\right)\right], \quad si\ \ v\ \ es\ \ par\end{aligned} \tag{8}$$

Como $0 \le sen^2\left(vt/2\right) \le 1$, se obtiene entonces que:

$$\frac{\sigma^2(v-1)}{2} \le \mathrm{Var}\left[f_{\mathbf{x}_i}(t)\right] \le \frac{\sigma^2(v+1)}{2} \tag{9}$$

Rubio (1983) indica que estas cuatro propiedades permiten realizar comparaciones de un conjunto de observaciones multivariadas sin pérdida de información utilizando sus representaciones generadas con las series finitas de Fourier propuestas por Andrews.

2.4 Detección de datos atípicos

En este punto una pregunta es: ¿cuán lejos del comportamiento usual debe estar una observación para que sea considerado como un dato atípico? Grubbs (1969) señala que las pruebas estadísticas pueden ser utilizadas para evaluar si existen razones suficientes para decir que una observación es muy diferente debido a razones no aleatorias y propias de la distribución. Grubbs (1969) y Dixon (1950) proponen dos pruebas alternativas para decidir si un dato, recopilado para una variable, puede ser considerado como atípico. En ambos casos se asume que los datos recopilados corresponden a una muestra aleatoria de una población Normal. Las pruebas sugeridas por éstos autores son utilizadas en diversos programas estadísticos como Minitab, R, SPSS, StatGraphics, etc. Para el caso de datos multivariados también se han propuesto pruebas estadísticas de significación, como es el caso de Canori & Prescott (1992), Pan & Wang (1994), y Filzmoser (2004). En estos casos también se asume que los datos provienen de una distribución normal multivariada. Tanto para datos univariados como para datos multivariados siempre existe la duda si el supuesto de normalidad puede aceptarse como válido, ya que de no ser asi, el procedimiento de prueba de hipótesis ya no sería aplicable. En estos casos es cuando se puede hacer uso de los métodos no paramétricos como el que se propone más adelante.

Pan & Wang (1994) proponen una prueba estadística para detectar observaciones atípicas multivariadas basada en el criterio de razón de máxima verosimilitud, e indican lo siguiente: Si $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p$ es una muestra aleatoria de una distribución p-dimensional $F_p(\mu, \Sigma)$, donde μ y $\Sigma > 0$ son el vector media y la matriz covariancia, respectivamente, y ambos son desconocidos. La detección de la presencia de datos atípicos puede ser llevada a cabo mediante la prueba siguiente: Si se elige la hipótesis nula de que no hay datos atípicos en la muestra,

$$H: \ \boldsymbol{x}_i \sim F_p(\mu, \Sigma) \quad i = 1, 2, \cdots, n$$

luego un posible modelo alternativo que puede ocurrir cuando se tiene múltiples observaciones atípicas sería el modelo con una media desplazada

$$\begin{aligned} K: \ & \boldsymbol{x}_i \sim F_p(\mu, \Sigma) \qquad (i \notin I) \\ & \boldsymbol{x}_i \sim F_p(\mu + a_i, \Sigma) \quad (i \in I) \end{aligned}$$

Donde $\boldsymbol{I} = \{i_1, i_2, \cdots, i_k\}$ es un subconjunto de índices de $\{1, 2, \cdots, n\}$ con un entero positivo k y $a_1, a_2, \cdots, a_k$ son parámetros de desplazamiento promedio desconocidos. Pan & Wang (1994) indican que Siotani (1959) y Wilks (1963) discutieron esta propuesta bajo el supuesto de una distribución normal multivariada.

Bajo la propuesta anterior, si se tiene observaciones atípicas dentro de una muestra, al menos uno de los coeficientes $a_1, a_2, \cdots, a_k$ debe ser diferente de cero; es decir, la presencia de un dato atípico afecta al comportamiento de la tendencia y por consiguiente el valor del promedio se ve afectado por dicha observación.

Acuña & Rodriguez (2004) y Muñoz García & Amón Uribe (2013) mencionan los siguientes métodos para la detección de datos atípicos multivariados.

- La distancia de Mahalanobis
- Métodos de componentes principales
- Búsqueda de proyecciones
- Análisis de grupos (clustering).

Ben-Gal (2005) y Barnett & Lewis (1994) hacen mención a la diferencia entre los métodos estadísticos paramétricos, basados usualmente en el supuesto de normalidad, y los métodos no paramétricos de minería de datos en los cuales no se asume un modelo para el comportamiento de los datos, lo que resulta una gran ventaja en el análisis de grandes conjuntos de datos.

2.5 Inferencia Booststrap

Las representaciones gráficas mencionadas en párrafos anteriores describen métodos que permiten representar observaciones multivariadas en un espacio bidimensional, y que pueden ser analizadas más fácilmente mediante la apreciación visual de las mismas. Un análisis objetivo de las diferencias entre las representaciones puede ser llevado a cabo mediante métodos de inferencia estadística.

Benjamini & Braun (2002) hacen mención sobre algunos aportes de John Tukey en los Problemas de Comparaciones Múltiples (PMC) y hacen notar la preferencia de Tukey sobre el uso de intervalos de confianza antes que las pruebas de significación, y que resulta importante reconocer los diferentes objetivos principales de la actividad estadística, que define como: acción, indicación y santificación. Según Benjamini & Braun (2002), Tukey señala que los intervalos de confianza son identificados con indicaciones mientras que las pruebas de significación son identificadas con santificaciones y acciones, y que al favorecer el uso de intervalos de confianza (o el análisis exploratorio de datos, métodos gráficos, estadísticas Bayesiana, etc.), los estadísticos implícitamente rechazan el uso de las pruebas de significación.

En los procesos de inferencia estadística se han discutido muchos métodos paramétricos y no paramétricos. Indudablemente que el caso de la estadística paramétrica resulta deseable y aplicable cuando se cumplen los supuestos exigidos por el marco teórico de cada método. En aquellos casos donde no se conoce el comportamiento aleatorio del estadístico, ni la distribución teórica de la variable en estudio, es posible aplicar el método Bootstrap. López & Elosua (2004) señalan: "La aplicación del enfoque Bootstrap permite obtener estimaciones de medidas de precisión, así como la realización de contrastes de hipótesis en aquellas situaciones en las que no se dispone de información acerca de la distribución muestral de un estadístico o en casos en los que la distribución muestral es dependiente de parámetros desconocidos".

Para Losilla (1994), la idea sugerida por Efron (1979) consiste en: "Si una muestra aleatoria contiene la máxima información disponible sobre la población, ¿por qué no proceder como si la muestra fuese la población y, entonces, estimar la distribución muestral de un estadístico generando nuevas muestras mediante un muestreo aleatorio con reposición, a partir de los datos de la muestra original".

Hesterberg, et al. (2003) señalan: "el método Bootstrap permite cuantificar incertidumbres mediante el cálculo de errores estándar e intervalos de confianza, y permiten realizar pruebas de significación. Además, presentan las siguientes ventajas:"

- Menos supuestos. No requiere que la distribución de los datos sea Normal (o de otro tipo), ni que el tamaño de muestra sea grande.

- Mayor precisión. Las pruebas de permutaciones y algunos métodos Bootstrap son más precisos en la práctica, que los métodos clásicos.
- Generalidad. Los métodos de remuestreo son notablemente similares para una amplia gama de estadísticos y no requiere de nuevas fórmulas para cada estadístico. No se requiere de fórmulas especiales para cada procedimiento.
- Promueve la comprensión. Los procedimientos Bootstrap desarrollan nuestra intuición proporcionando analogías concretas respecto a conceptos teóricos

Efron (1979), Losilla (1994), Sánchez (2012), y otros autores hacen mención de los siguientes métodos de estimación Bootstrap:

- **Bootstrap no paramétrico**

  La estimación Bootstrap se basa en el supuesto de que la función de distribución empírica $\hat{F}_n$, obtenida a partir de una muestra correspondientes a extracciones aleatorias $\{x_1, x_2, \cdots, x_n\}$ de n variables aleatorias X_i con idéntica función de distribución F:

  $$X_1, X_2, \cdots, X_n \overset{iid}{\approx} F$$

  es la estimación máximo verosímil no paramétrica de F, fundamentada en la asignación de probabilidad 1/n a todos y cada uno de los datos muestrales.

- **Bootstrap paramétrico**

  En el caso de que la función de distribución F_θ a partir de la cual se han extraído los datos sea conocida excepto por su parámetro θ, puede recurrirse a la estimación de este parámetro considerando que $\hat{\theta}$ es una buena estimación de θ, obtenida a partir de los datos muéstrales y, por tanto, que $F_\theta = F_{\hat{\theta}}$.

  En esta caso, en lugar de extraer muestras aleatorias a partir de la muestra original X_o, pueden extraerse directamente por muestreo Monte Carlo con base en el modelo de distribución $F_{\hat{\theta}}$.

- **Bootstrap suavizado**

  El Bootstrap suavizado consiste en sustituir la distribución empírica $F_{\hat{\theta}}$ por otra distribución $F_{\hat{\theta}}^{s}$ previamente suavizada. El proceso de suavización puede realizarse mediante una "perturbación", término utilizado por Holmes (1990), de cada observación de la muestra original, añadiéndole una variable aleatoria con distribución suave, como por ejemplo, la normal o la uniforme, como lo señalan Silverman y Young (1987).

Varios autores tales como: Alonso (2002), Boos (2003), Solanas & Sierra (1992), Sánchez (2012) señalan que el método Bootstrap propuesto por Efron (1994) se basa en la utilización del principio de analogía o sustitución (plug-in) que constituye uno de los métodos más simples utilizados para obtener un estimador de un parámetro poblacional $\theta = T(F)$, donde *T* es un funcional definido en una clase convexa de funciones de distribución, y *F* es la distribución subyacente de los datos. Un estimador de sustitución o plug-in es $\hat{\theta} = T(\hat{F})$ donde $\hat{F}$ es un estimador de *F*.

Por otro lado, Solanas & Sierra (1992) y Reyes & Ramírez (2002) señalan que la justificación teórica está basada en dos consideraciones:

1) La Función de Distribución Empírica, $\hat{F}(x)$, estima a la función de distribución verdadera, $F(x)$; y el teorema Glivenko-Cantelli muestra que $\hat{F}(x)$ converge en probabilidad a $F(x)$, decir,

$$\lim_{n\to\infty} sup\left|\hat{F}_n(x) - F(x)\right| = 0$$

   Como lo indican Bickel & Freedman (1981), intuitivamente, en la aplicación de la ley de los grandes números, cuando se incrementa el tamaño de la muestra, ésta contiene mayor información acerca de la población, y para n=N, $\hat{F}(x) = F(x)$.

2) La propiedad de consistencia permite a la distribución muestral Bootstrap $\hat{F}^{*}(\hat{\theta}^{*})$ aproximar a $F(\hat{\theta})$ de una muestra dada, cuando el número de remuestreos B es grande y permite aproximar $\hat{F}(x)$ a $F(x)$.

Flores (2005), Sánchez (2012), y otros autores indican que una prueba de hipótesis se basa en un estadístico $T = T(x_1, x_2, \cdots, x_n)$ que mide la discrepancia entre los datos $(x_1, x_2, \cdots, x_n)$ y la hipótesis nula. Si el valor del estadístico es "t", el nivel de evidencia contra la hipótesis nula se mide por la probabilidad:

$$p = P_{H_0}\left[T > |t|\right] = P\left[T > |t| \ / H_0\right]$$

La cual es también llamada "p-valor". Cuanto menor sea el p-valor mayor será la evidencia en contra de H_0. Alonso (2002) señala que la aproximación Bootstrap a esta probabilidad, para el caso de pruebas unilaterales a la derecha, es:

$$p^*(x) = \frac{1}{B}\sum_{i=1}^{B} I\left(b_i^* > |x|\right)$$

donde

$$I\left(b_i^* > |x|\right) = 1, \quad si \quad b_i^* > |x|$$
$$= 0, \quad de \ otro \ modo$$

Esta aproximación será apropiada cuando el número de muestras bootstrap B sea lo suficientemente grande. Efron (1994), Sánchez (2012), y otros autores señalan que el valor de B puede ser determinado por el comportamiento del coeficiente de variación del error estándar bootstrap condicional sobre una muestra dada.

$$CV\left(\hat{\sigma}^*(\hat{\theta}) / x\right) = \left[\frac{\hat{\Delta} + 2}{4B}\right]^{1/2}$$

Donde $\hat{\Delta}$ es la curtosis de la distribución bootstrap de $\hat{\theta}$. También señalan para valores de CV mayores de 0.10 se observa pocas mejoras con respecto a valores de B mayores a 100. No obstante, indican que para el caso de estimaciones por intervalos y para el cálculo de cuantiles es recomendable un mayor número de remuestras.

Boos (2003) señala que un correcto conocimiento del remuestreo Bootstrap para obtener un error estándar o un intervalo de confianza no necesariamente proporciona una compresión adecuada de cómo debe ser el muestreo en una prueba de hipótesis. El punto clave para obtener el p-valor es que el muestreo debe realizarse bajo una apropiada hipótesis nula,

mientras que, para la estimación del error estándar e intervalos de confianza, el muestreo es irrestricto.

En la prueba de hipótesis $H_0: \theta = \theta_0$, $H_1: \theta \neq \theta_0$, Becher (1993) hace mención a las siguientes recomendaciones de Hall y Wilson (1991): la primera recomendación señala que el muestreo debe realizarse de $\hat{\theta}^* - \hat{\theta}$, en vez de muestrear $\hat{\theta}^* - \theta_0$, y como segunda recomendación señala para la prueba ϕ y el valor α deberían estar basados en la distribución Bootstrap de $(\hat{\theta}^* - \hat{\theta}) / \hat{\sigma}^*$, y no en la distribución Bootstrap de $(\hat{\theta}^* - \hat{\theta}) / \hat{\sigma}$. Hall y Wilson (1991) proponen realizar la prueba de la siguiente manera:

$$\phi(x) = \begin{cases} 0, & si \ |\hat{\theta}^* - \hat{\theta}| / \hat{\sigma} \leq \hat{t}, \\ 1, & de \ otra \ manera \end{cases}$$

Donde $\hat{t}$ es calculado como el número que cumple con la condición:

$$P_r^* \left[\left| \hat{\theta}^* - \hat{\theta} \right| / \hat{\sigma}^* > \hat{t} \right] = \alpha$$

y donde P_r^* representa la probabilidad medida bajo la distribución Bootstrap.

III. MATERIALES Y MÉTODOS

3.1 Materiales

- Computadora Toshiba. Satellite. Core i7, con 6GB de memoria RAM. Disco duro de 500 GB. Sistema operativo Windows 10.
- Programa R. Versión R x64 3.3.2
- Programa Minitab 17
- Programa Excel 2016
- Impresora Láser: RICOH Aficio PostScript 3'. Modelo MP C305 PCL 6 Color
- Papel bond

3.2 Metodología de la investigación

3.2.1 Tipo de investigación

La presente investigación ha sido definida como un trabajo exploratorio no experimental por que el objetivo principal es proponer un procedimiento metodológico alternativo para detectar datos atípicos multivariados sin el requisito de normalidad exigido en los métodos paramétricos usuales.

3.2.2 Hipótesis de la investigación

El procedimiento propuesto si permite detectar datos atípicos multivariados con ciertos niveles de confianza generados por la aplicación de la inferencia Bootstrap.

3.2.3 Semejanza entre observaciones multivariadas

Tal como lo indican Embrechts y Herzberg (1991), y otros autores, las representaciones pueden ser afectadas por las magnitudes de los datos asociados a las variables, y al orden en que se incluyen dentro de la serie de Fourier. Para hacer que dicha influencia no afecte a las representaciones se recomienda trabajar con los datos estandarizados.

Asumiendo que no se conoce el modelo probabilístico de las distancias entre las representaciones mediante series finitas de Fourier, se propone desarrollar un procedimiento alternativo mediante el método Bootstrap, para construir la distribución empírica que permita comparar las distancias entre las observaciones multivariadas.

Basados en las representaciones mediante series finitas de Fourier, los datos base para estudiar la semejanza entre las observaciones multivariadas son las distancias entre sus representaciones; es decir, para un conjunto de valores "t", se obtienen las distancias:

$$D_{k,m}(t) = \left| f_{\mathbf{x}_k}(t) - f_{\mathbf{x}_m}(t) \right| \qquad (10)$$

Esta distancia representa el alejamiento que existe entre las representaciones de dos observaciones multivariadas en el instante "t".

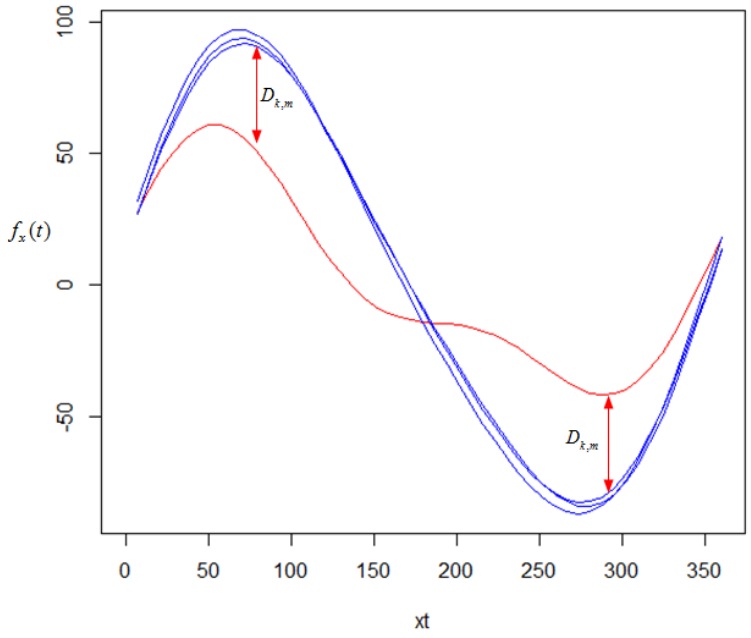


Figura 1. Distancias entre representaciones de Fourier

Considerando que $-\pi \le t \le \pi$, los valores de "t" se eligen mediante una muestra sistemática con arranque aleatorio, con el propósito de que los valores elegidos tengan una buena cobertura en dicho intervalo. Por otro lado, como los valores de la función $f_{\mathbf{x}_k}$ no producen cambios repentinos para valores cercanos de "t", se deduce que un muestreo sistemático permite seleccionar un conjunto de distancias que será representativo de todas las diferencias que existen entre dos series finitas de Fourier.

3.2.4 Prueba estadística

Una característica propia de las observaciones atípicas es que su presencia afecta de manera significativa a las medidas estadísticas tales como la media aritmética y a la variancia. Es decir, la presencia de un dato atípico hace que el valor de la distancia promedio y de la variancia sean mayores. Considerando que, para analizar las semejanzas entre las observaciones multivariadas, los datos base son las distancias, se tendrá que el valor del promedio será siempre mayor o igual que cero. Un valor alto será un indicativo de que las representaciones son significativamente diferentes y también las observaciones multivariadas que las generan.

En el presente trabajo se propuso utilizar como un indicador de la semejanza entre las observaciones multivariadas a los promedios de las distancias existentes entre las representaciones gráficas de Andrews. El indicador básico de similitud será entonces la distancia promedio entre las representaciones. Una distancia promedio igual a cero será un indicativo que las observaciones son idénticas entre si.

Por tanto, los datos base para hacer comparaciones entre las observaciones multivariadas serán las distancias $\{d_1, d_2, \cdots, d_m\}$ obtenidas para una muestra sistemática con arranque aleatorio, correspondiente a un conjunto de "b" valores comprendidos en el intervalo: $-\pi < t < \pi$. Puesto que los valores de una serie finita de Fourier cambian gradualmente para valores "t" dentro de este intervalo, una muestra sistemática resulta apropiada para generar una muestra representativa.

Puesto que no se conoce la distribución probabilística de las distancias, para el proceso de inferencia se hace uso del método no paramétrico Bootstrap para construir la población empírica asociada a la muestra sistemática de distancias entre las representaciones mediante series finitas de Fourier.

Como lo señalan Alonso (2002), Boos (2003), Solanas & Sierra (1992), Sánchez (2012), y otros autores, el método Bootstrap propuesto por Efron (1994) se basa en la utilización del principio de analogía o sustitución (plug-in). Sobre la base de una muestra aleatoria inicial $D=(d_1,d_2,\cdots,d_m)$ con la cual se genera la distribución empírica $\hat{F}$ como aproximación de las distribución desconocida *F*. Las aproximaciones Bootstrap del promedio, la variancia y la distribución acumulada del estadístico $\hat{\theta}=b(D)$ son dados por:

$$\begin{aligned} b^* &= E^*\left[b^*(D^*)\right] \\ S^{*2} &= Var^*\left[b^*(D^*)\right] \\ H^*(x) &= Pr^*\left[b^*(D^*)\le x\right] \end{aligned} \qquad (11)$$

donde D^* es una muestra obtenida de $\hat{F}$, y E^*, Var^*, *y* Pr^* denotan la esperanza, la variancia y la probabilidad Bootstrap generadas con la muestra D^*.

El procedimiento de Monte Carlo para estimar las expresiones definidas en (11) es:

1. Generar B muestras independientes $D^{*(b)}$ a partir de $\hat{F}$, con b=1,2,…, B
2. Calcular el valor de $b_i^*=b^*(D_i^*)$, para i=1,2, … , B
3. Obtener los estimados Bootstrap:

$$\begin{aligned} b^* &= \frac{1}{B}\sum_{i=1}^{B} b_i^* \\ S^{*2} &= \frac{1}{B-1}\sum_{i=1}^{B}(b_i^*-b^*)^2 \\ H^*(x) &= \frac{1}{B}\sum_{i=1}^{B} I(b_i^*\le x) \end{aligned}$$

donde

$$I(b_i^* \le x) = 1, \quad si \quad b_i^* \le x$$
$$= 0, \quad de \ otro \ modo$$

3.2.5 Algoritmo propuesto

Para verificar si una observación puede o no ser considerada como atípica se propuso realizar la comparación de los promedios de dos distribuciones Bootstrap: el primero generado con todas las observaciones de la muestra original de distancias, y el segundo generado con las distancias existentes sin considerar la observación en evaluación. El objetivo es dar respuesta a la siguiente hipótesis:

$$H_0 : \mu_j \ge \mu_c$$
$$H_1 : \mu_j < \mu_c$$

Donde μ_c representa el promedio con todas las observaciones, y μ_j representa el promedio sin considerar la observación $\boldsymbol{x}_j$. Si realmente existe una observación atípica, su representación mediante una serie de Fourier estará alejada de las otras representaciones y esto ocasionará distancias muy elevadas, y por consiguiente, el valor del promedio con toda la muestra debe estar influenciado por estas distancias elevadas. Por lo tanto, si se retira la observación $\boldsymbol{x}_j$ se tendrá un nuevo promedio μ_j que deberá ser menor al promedio completo μ_c .

Considerando que una serie Finita de Fourier no presenta cambios bruscos para diferentes valores de "t", con el propósito de expandir la muestra a lo largo del intervalo $-\pi < t < \pi$, y evitar posibles sesgos de una muestra simple aleatoria al concentrar la muestra en un pequeño intervalo de valores de "t". Además, como lo señala Lhor (2000), la muestra sistemática es más precisa que una muestra simple aleatoria cuando la variancia de la muestra sistemática es mayor que la variancia general, y que para este el caso de la comparación de series Finitas de Fourier se logra considerando valores de "t" diferentes dentro del intervalo $-\pi < t < \pi$.

El algoritmo Bootstrap que se propuso para evaluar si una muestra de observaciones multivariadas tiene observaciones atípicas es el siguiente:

Paso 1 Leer la matriz de datos

$$\mathbf{Y} = (\mathbf{y}_i) = (y_{ij}), \qquad i = 1,2,\cdots,n; \quad j = 1,2,\cdots,k$$

Paso 2 Obtener la matriz de los datos estandarizados (paso opcional)

$$\mathbf{X} = (\mathbf{x}_i) = (x_{ij}), \qquad i = 1,2,\cdots,n; \quad j = 1,2,\cdots,k$$

$$x_{ij} = \frac{y_{ij} - \overline{y}_j}{S_j}$$

Paso 3 Generar una muestra sistemática con arranque aleatorio $\{t_1, t_2, \cdots, t_m\}$ de "m" valores de "t" comprendidos en el intervalo: $-\pi < t < \pi$.

Paso 4 Obtener las series finitas de Fourier, con la información de cada elemento a agrupar $\mathbf{x}_i$.

$$f_{\mathbf{x}_i}(t) = \mathbf{x}_i \boldsymbol{a}(t) = \frac{x_{i,1}}{\sqrt{2}} + x_{i,2} sen(t) + x_{i,3} cos(t) + x_{i,4} sen(2t) + x_{i,5} cos(2t) + \cdots$$

Paso 5 Obtener las distancias entre las series finitas de Fourier.

Para la muestra de valores t, $\{t_1, t_2, \cdots, t_m\}$ obtener:

$$D_{j,k}(t_i) = \left| f_{\mathbf{x}_j}(t_i) - f_{\mathbf{x}_k}(t_i) \right|, \quad \text{i} = 1,2,\cdots,\text{m}; \quad \text{j} = 1,2,\cdots,\text{n}; \quad \text{k} = \text{i}+1,\cdots,\text{n}$$

Paso 6 Obtener la distribución empírica de todas las distancias mediante B muestras Bootstrap.

Repetir el siguiente proceso B veces:

Del conjunto de distancias generadas $D_{j,k}(t_i)$ elegir una muestra con reemplazo de tamaño m, $\{d_1, d_2, \cdots, d_m\}$, y con esta muestra obtener la distancia promedio $\overline{d}_i^*$.

Obtener el estimado Bootstrap:

$$b_c^* = \frac{1}{B}\sum_{i=1}^{B} \bar{d}_i^*$$

Paso 7 Realizar el proceso de inicialización siguiente:

Para j=1,2, . . . , n, hacer: pvalor(j) = 0

Paso 8 Para j=1, 2, . . ., n, repetir el siguiente proceso:

Mediante B muestras Bootstrap, obtener la distribución empírica de todas las distancias sin considerar la observación "j".

Repetir el siguiente proceso B veces:

Del conjunto de distancias generadas $D_{j,k}(t_i)$, eliminando la fila "j" y la columna "j", elegir una muestra con reemplazo de tamaño m, $\{d_1, d_2, \cdots, d_m\}$, y con esta muestra obtener la distancia promedio $\bar{w}_j^*$.

Obtener el valor

$$H_j^*(\bar{w}_j^*) = \frac{1}{B}\sum_{i=1}^{B} I(\bar{d}_i^* < \bar{w}_j^*)$$

$$Pvalor^*(j) = H_j^*(\bar{w}_c^*)$$

donde

$$I(\bar{d}_i^* < \bar{w}_j^*) = 1, \quad si \quad \bar{d}_i^* < \bar{w}_j^*$$
$$= 0, \quad de\ otro\ modo$$

Paso 9 Generar el reporte de resultados, sobre las pruebas:

Para j=1,2, . . , n, se hace la prueba:

$$H_0 : \mu_j \geq \mu_c$$
$$H_1 : \mu_j < \mu_c$$

Si $Pvalor^*(j) < \alpha$ Se rechaza H_0

Si $Pvalor^*(j) \geq \alpha$ No se rechaza H_0

La observación “j” puede ser considerada como atípica cuando se rechaza la hipótesis nula, que indica que el promedio de las distancias no se encuentra afectada por la presencia de la observación “j”.

Paso 10 Generar gráficos de las representaciones de Fourier, y el listado de las probabilidades Bootstrap estimadas.

Los programas desarrollados con este algoritmo fueron elaborados en el lenguaje R, y se presentan en los anexos.

3.2.6 Metodología de análisis

Para probar el algoritmo propuesto se hace uso del método de análisis hipotético deductivo basado en tres conjuntos de datos preparados de manera específica con el propósito de simular la presencia o ausencia de datos atípicos dentro de una muestra multivariada. Se trata de investigar si el algoritmo es capaz de identificar la presencia de las observaciones inusuales incluidas de manera ex profesa dentro del conjunto de datos.

Si bien el procedimiento propuesto se puede aplicar a cualquier número de observaciones multivariadas, existe una limitación en la capacidad de cómputo disponible debido a que las matrices que se generan en el procedimiento crecen exponencialmente con la cantidad de datos a analizar.

Por otro lado, los reportes incluyen la medición de la posibilidad de que cada observación sea considerada como atípica. Esto ocasiona que los reportes sean más extensos cuando se tenga una gran cantidad de datos.

Por las razones expuestas en los dos párrafos anteriores se limita la cantidad de datos en los tres conjuntos de datos preparados para probar el algoritmo.

Adicionalmente se hace uso de un cuarto conjunto de datos que consiste de las Horas-hombre perdidas por huelgas entre los años 1996 y 2012. Esta información fue tomada del Ministerio de Trabajo y Promoción de Empleo.

Los datos utilizados como prueba del algoritmo se sintetizan en el siguiente cuadro:

Conjuntos de datos utilizados como prueba del algoritmo

Conjunto de datos	Número de variables	Número de observaciones	Fuente
1	2	16	Datos simulados
2	4	10	Datos simulados
3	4	31	Datos simulados
4	4	17	Horas-hombre perdidas por huelgas 1996-2012. Ministerio de Trabajo y Promoción de Empleo

Los datos simulados fueron generados aleatoriamente sin un modelo definido, y a esos datos se adicionaron algunas observaciones ligeramente diferentes, y otras observaciones muy alejadas de la tendencia. El objetivo fue tener observaciones diferentes para ver si el algoritmo propuesto era o no capaz de identificarlos.

Los datos utilizados en los cuatro conjuntos de datos se muestran en el Anexo 2.

IV. RESULTADOS Y DISCUSIÓN

4.1 Análisis del primer conjunto de datos

El conjunto de datos N°1 (ver Anexo 2) se preparó con el propósito de mostrar como un análisis univariado puede fallar en la detección de datos atípicos multivariados.

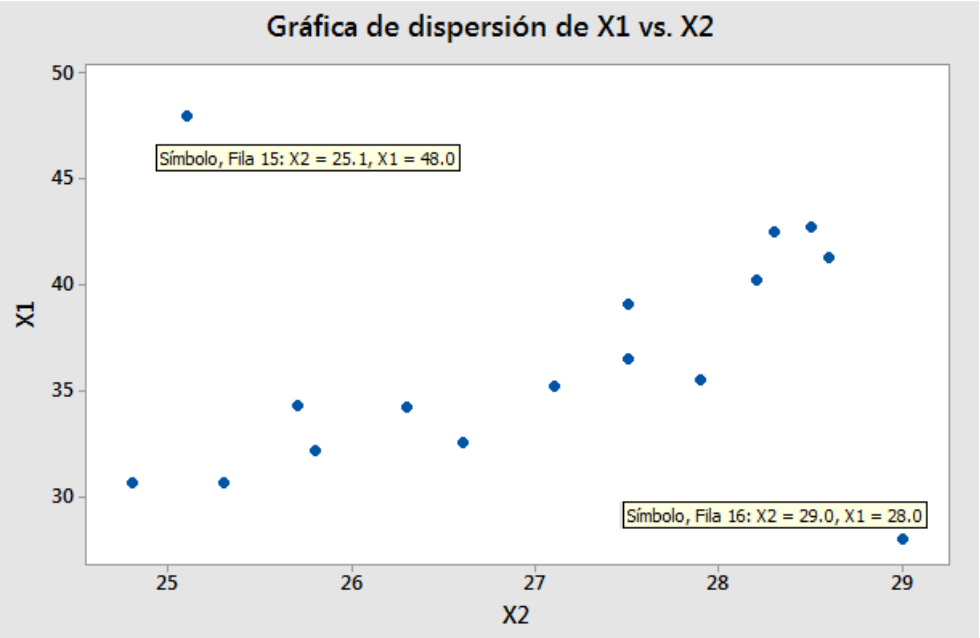


Figura 2. Gráfico de dispersión del primer conjunto de datos

Se puede observar que los datos de las filas 15 y 16 están alejados de la tendencia general, y que por tanto podrían ser considerados como datos atípicos. No obstante, el análisis individual de las variables en dicho conjunto de datos conduce a los siguientes resultados:

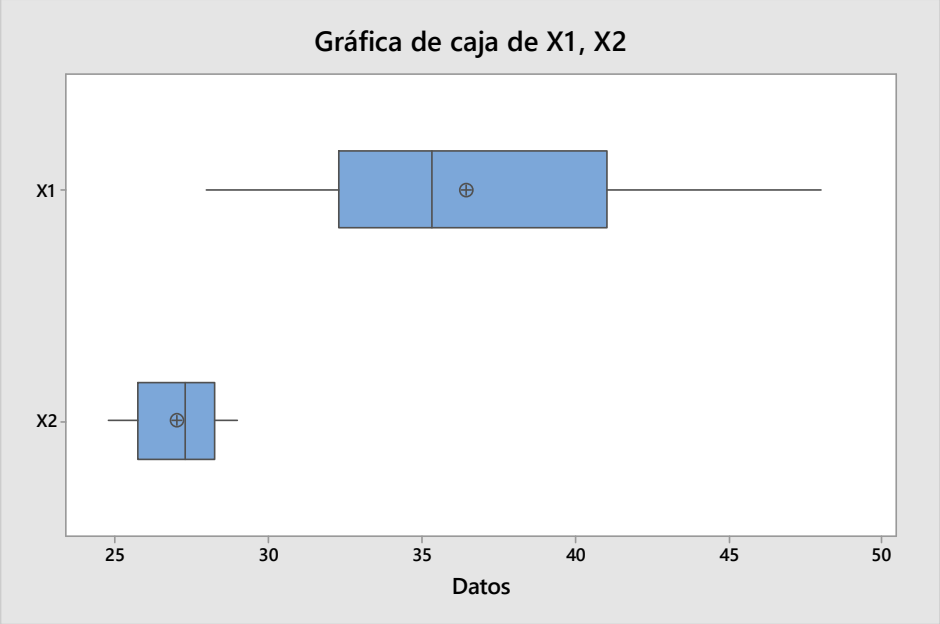


Figura 3. Gráfico de cajas para las variables en el primer conjunto de datos

Prueba de valores atípicos: X1, X2

Prueba de Grubbs

Variable	N	Media	Desv.Est.	Mín.	Máx.	G	P
X1	16	36.48	5.39	28.00	48.00	2.14	0.340
X2	16	27.013	1.375	24.800	29.000	1.61	1.000

* NOTA * No hay valor atípico en el nivel de significancia de 5%

Figura 4. Prueba de datos atípicos de Grubbs

Tanto el gráfico de cajas como la prueba de Grubbs no detectan la presencia de los datos atípicos presentes en las filas 15 y 16 del primer conjunto de datos. Ahora, al usar el algoritmo propuesto en este trabajo se obtiene el siguiente resultado:

```
Promedio total estimado :   4.467219

Valores Alpha para cada observación

Ho : Promedio sin este dato >= Promedio total
H1 : Promedio sin este dato <  Promedio total

   Promedios Alpha Sig       valt
1   4.544071 0.915      1.0317290
2   4.650661 1.000      2.3391571
3   4.625339 1.000      1.9720059
4   4.623650 1.000      1.9200657
```

```
5    4.462027 0.460      -0.0653202
6    4.525059 0.840       0.6947494
7    4.646682 1.000       2.2757936
8    4.434191 0.280      -0.4387839
9    4.437832 0.295      -0.3893597
10   4.384527 0.055      -1.0370978
11   4.370668 0.045      -1.2117479
12   4.568517 0.960       1.2266355
13   4.562731 0.955       1.1844861
14   4.639299 1.000       2.1912108
15   3.864350 0.000    * -8.5691719
16   4.181172 0.000    * -4.0240418

Coeficiente de variabilidad:  5.239

Número de muestras:      200
```

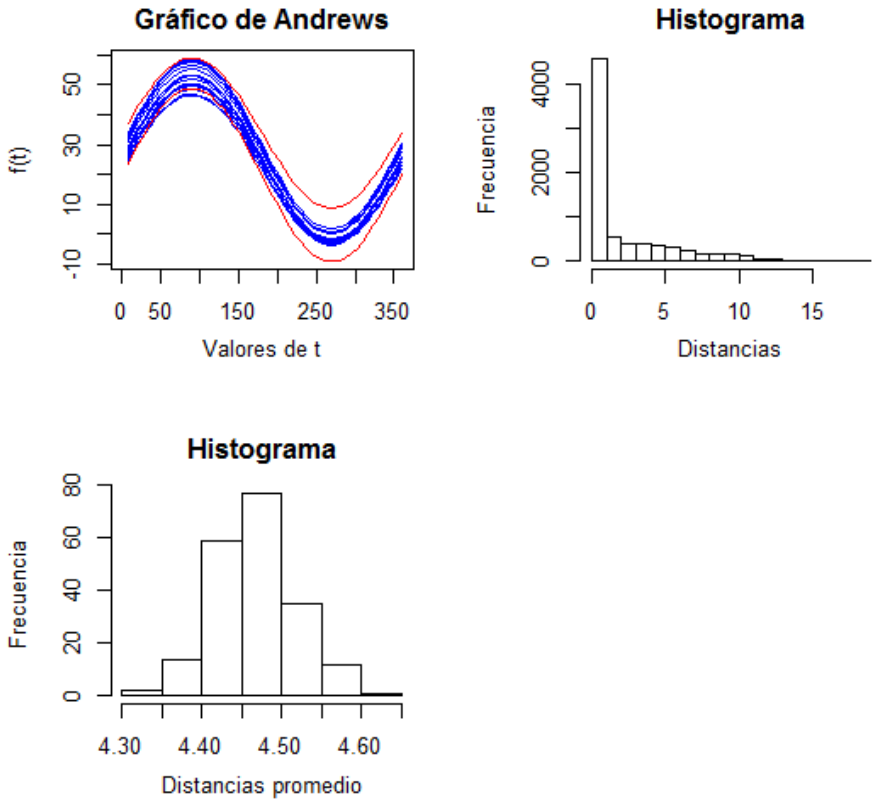


Figura 5. Representaciones de Fourier y distribución de distancias para el primer conjunto de datos.

Como se puede observar el método detectó los dos casos atípicos que se presentan en las observaciones 15 y 16, y para los que se tiene un "alpha" estimado de cero, por lo que se concluye que dichas observaciones inusuales. También se puede apreciar que la distribución de las distancias es asimétrica a la derecha, como era de esperar, y que la distribución de

promedios lograda con 200 muestras Bootstrap, tiene un comportamiento aproximadamente Normal. Los valores T, calculados como referencia, indican que las observaciones 15 y 16 muestran una fuerte diferencia con la tendencia general. También se puede observar que eliminando una de dichas observaciones hace que la distancia promedio disminuya significativamente de 4.467 a 3.864 cuando se elimina la observación 15, y disminuye a 4.181 cuando se elimina la observación 16.

El reporte que se genera sin la observación 16 es el siguiente:

```
Promedio total estimado :   4.175261

Valores Alpha para cada observación

Ho : Promedio sin este dato >= Promedio total
H1 : Promedio sin este dato <  Promedio total

   Promedios Alpha Sig         valt
1   4.206205 0.685      0.359031124
2   4.344294 0.995      1.939594869
3   4.314093 0.990      1.597006143
4   4.316169 0.990      1.644396860
5   4.174438 0.480     -0.009460144
6   4.241246 0.865      0.786205674
7   4.340115 0.995      2.025497571
8   4.077302 0.045     -1.215024510
9   4.076322 0.045     -1.174983779
10  4.099428 0.105     -0.908435873
11  4.075640 0.045     -1.202112597
12  4.290832 0.985      1.334666398
13  4.237364 0.855      0.722365862
14  4.348429 0.995      2.046938266
15  3.541535 0.000   * -8.083081428

Coeficiente de variabilidad:  5.338

Número de muestras:      200
```

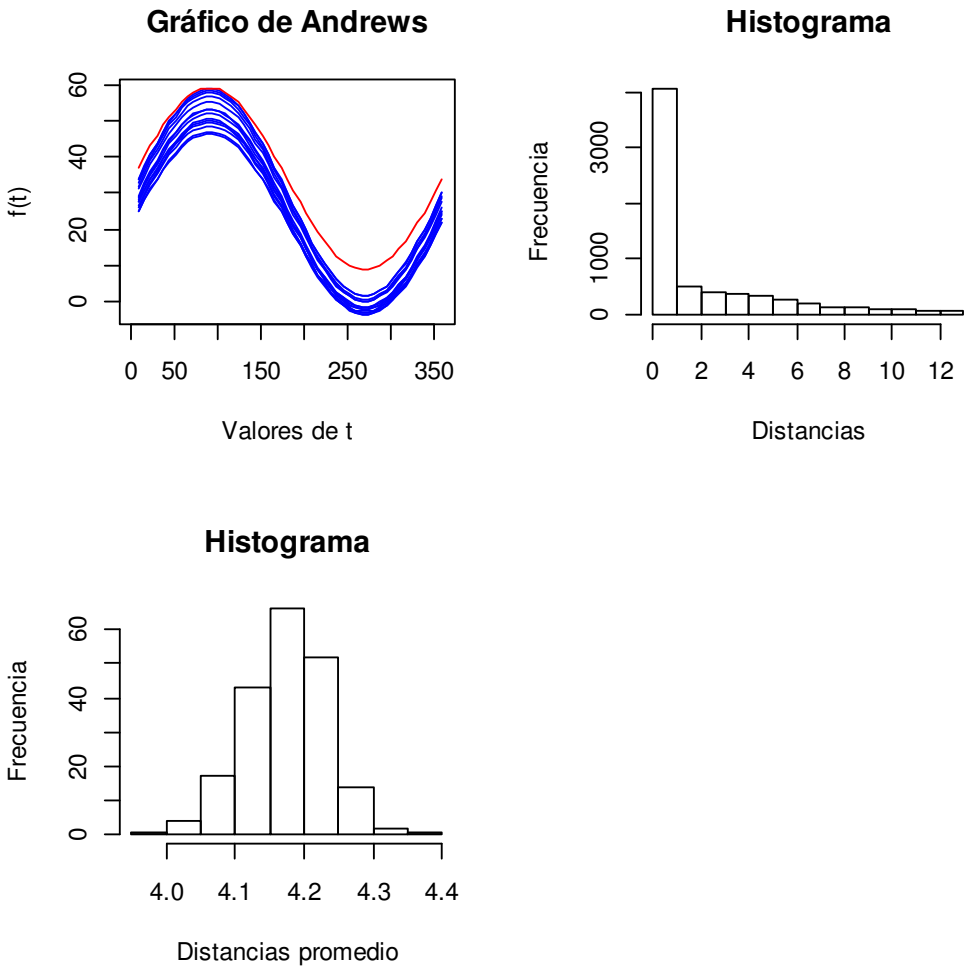


Figura 6. Representaciones de Fourier y distribución de distancias para el primer conjunto de datos, sin la observación 16.

En este reporte se observa que el método sigue detectando a la observación 15 como atípica, y que no aparecen otras observaciones inusuales con la eliminación de la observación 16. También se observa que en la distribución de las distancias promedio el rango ahora va entre 4.0 y 4.4, diferente al caso con todas las observaciones donde el rango estaba entre 4.30 y 4.70.

4.2 Análisis del segundo conjunto de datos

El segundo conjunto de datos (Ver Anexo 2), se ha elaborado con el propósito de evaluar el algoritmo con más de dos variables. Para éstos datos se obtiene el siguiente gráfico de cajas:

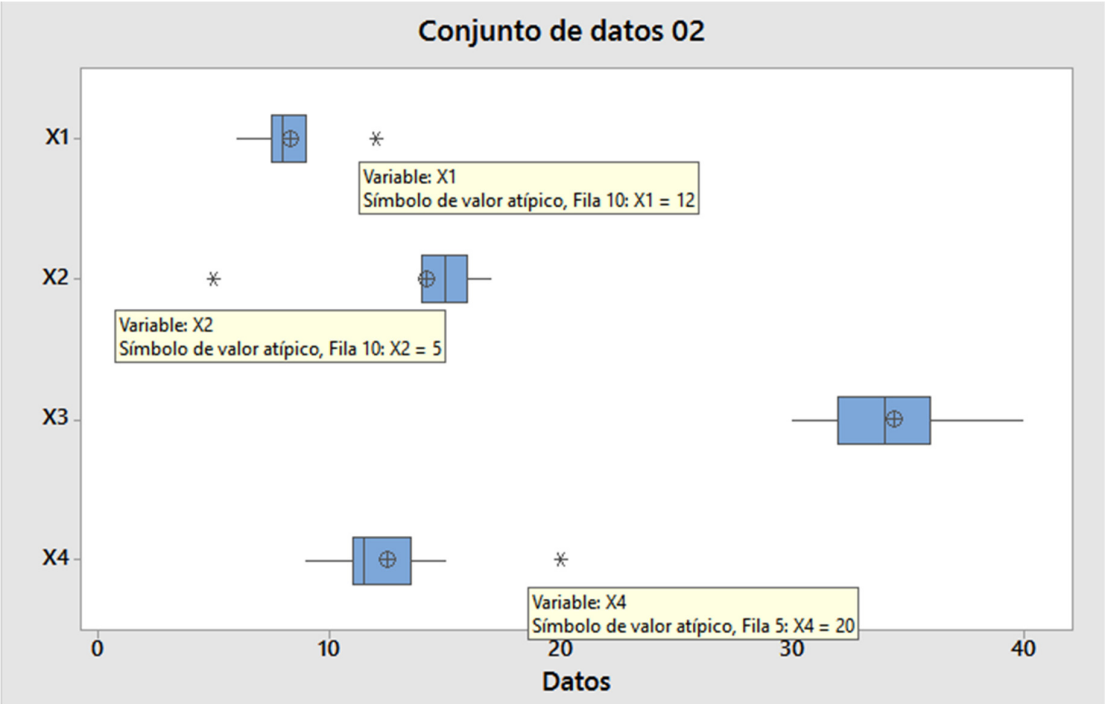


Figura 7. Gráfico de cajas para las variables en el segundo conjunto de datos

En el gráfico de cajas se precia que la observación 10 es considerada atípica para la variable X1 y X2, y que la observación 5 es considerada atípica para la variable X4.

Con el algoritmo propuesto en este trabajo se obtiene el siguiente resultado:

```
Promedio total estimado :   7.273847

Valores Alpha para cada observación

Ho : Promedio sin este dato >= Promedio total
H1 : Promedio sin este dato <  Promedio total

   Promedios Alpha Sig       valt
1   7.884102 0.995      1.7211097
2   7.856268 0.995      1.5537346
3   7.677164 0.950      1.1395908
4   7.650063 0.925      1.0813699
5   5.308353 0.000    * -6.4917067
```

```
6    7.868311 0.995       1.7501086
7    7.864264 0.995       1.6467697
8    7.445005 0.780       0.4860692
9    7.756294 0.985       1.4057955
10   5.277351 0.000    * -6.4317114

Coeficiente de variabilidad:  7.861

Número de muestras:      200
```

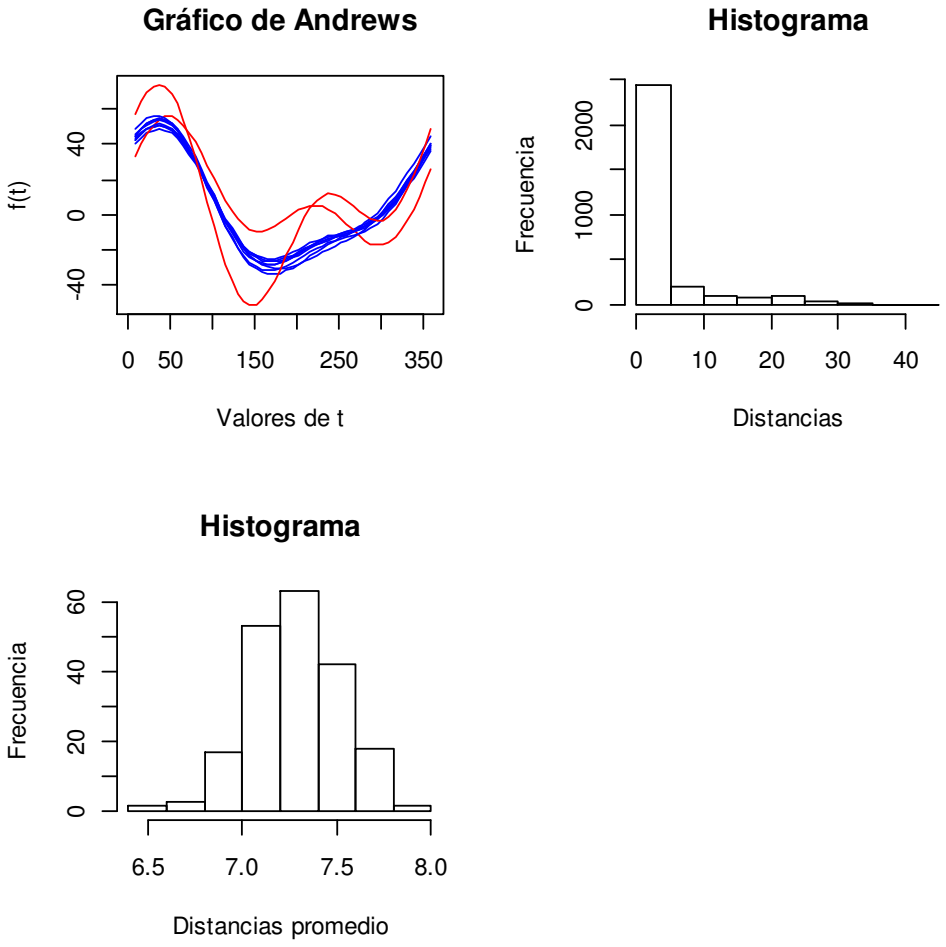


Figura 8. Representaciones de Fourier y distribución de distancias para el segundo conjunto de datos.

Como se puede observar el método detectó los dos casos atípicos que se presentan en las observaciones 5 y 10, para los que se tiene un "alpha" estimado de cero, por lo que se concluye que dichas observaciones son atípicas. También se puede apreciar que la distribución de las distancias es asimétrica a la derecha, y que la distribución de las distancias promedio para 200 muestras Bootstrap, tiene un comportamiento aproximadamente Normal

con un rango entre 6.4 y 8.0. Además, como se observa, los promedios sin los datos 5 y 10 son 5.31 y 5.28, respectivamente. Por otro lado, los valores T, calculados como referencia, indican que las observaciones y 5 y 10 muestran una fuerte diferencia con la tendencia general. Estos resultados son similares a los observados en el gráfico de cajas, en donde se aprecia similares resultados.

4.3 Análisis del tercer conjunto de datos

Con el tercer conjunto de datos (Ver Anexo 2) se obtiene el siguiente gráfico de cajas:

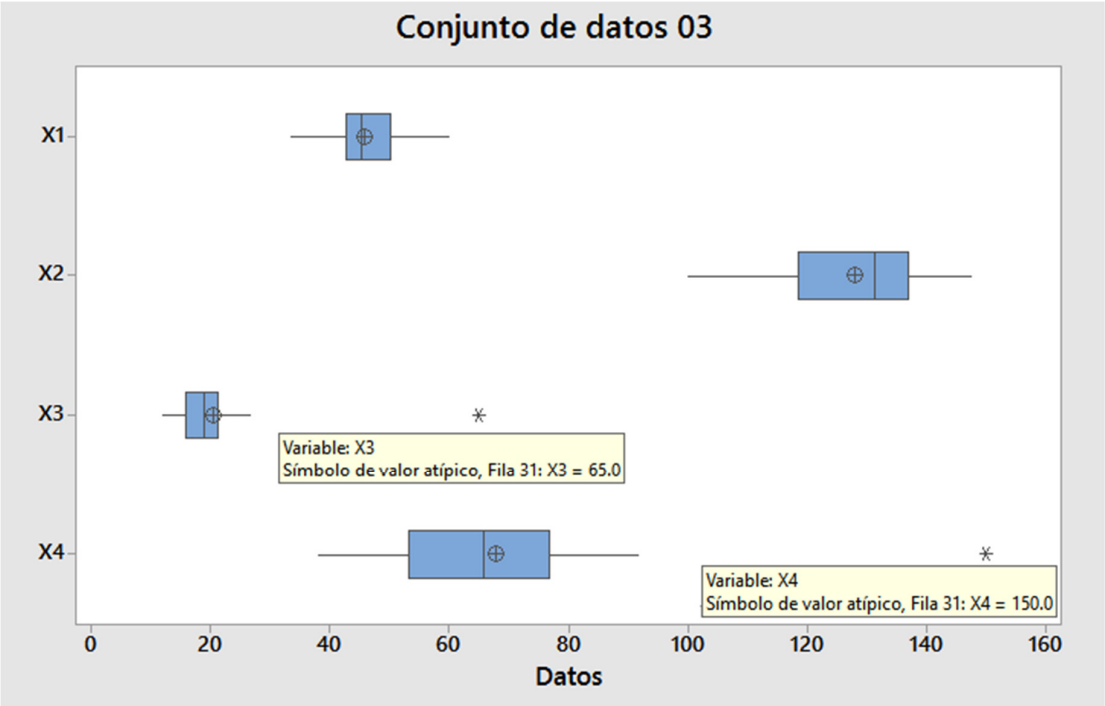


Figura 9. Gráfico de cajas para las variables en el tercer conjunto de datos

En este gráfico de cajas se precia que la observación 31 es considerada atípica para la variable X3 y X4: Para las variables X1 y X2 no se detectan valores atípicos.

Con el tercer conjunto de datos (Ver Anexo 2) se obtiene el siguiente reporte con el algoritmo propuesto:

```
Promedio total estimado :   18.0761048

Valores Alpha para cada observación

Ho : Promedio sin este dato >= Promedio total
H1 : Promedio sin este dato <  Promedio total

   Promedios Alpha Sig         valt
1   18.33148 0.955       1.08331987
2   18.33778 0.960       1.14468161
3   18.27823 0.890       0.88437929
4   18.22773 0.810       0.67805772
5   17.88040 0.105      -0.85927644
6   18.31569 0.950       1.08340217
7   18.10307 0.600       0.12055643
8   18.39663 0.975       1.37652056
9   18.12091 0.615       0.19842816
10  18.17102 0.705       0.40209404
11  17.64724 0.000   *  -1.87390987
12  18.19198 0.765       0.51229949
13  18.35133 0.960       1.22987767
14  18.19144 0.760       0.48603232
15  18.40854 0.980       1.54780044
16  18.23171 0.820       0.64911877
17  18.10394 0.600       0.13076828
18  18.06304 0.490      -0.05622248
19  18.40886 0.980       1.47738252
20  17.95725 0.230      -0.50501237
21  18.32909 0.955       1.07976290
22  18.47413 0.995       1.76915337
23  17.94942 0.195      -0.56694371
24  18.21408 0.795       0.60727951
25  18.36679 0.965       1.18538950
26  18.28682 0.920       0.95312092
27  18.13895 0.640       0.26774712
28  18.08570 0.560       0.04323640
29  18.38948 0.975       1.36974890
30  18.26034 0.870       0.79024419
31  15.08087 0.000   * -16.38963613

Coeficiente de variabilidad:  7.457

Número de muestras:      200
```

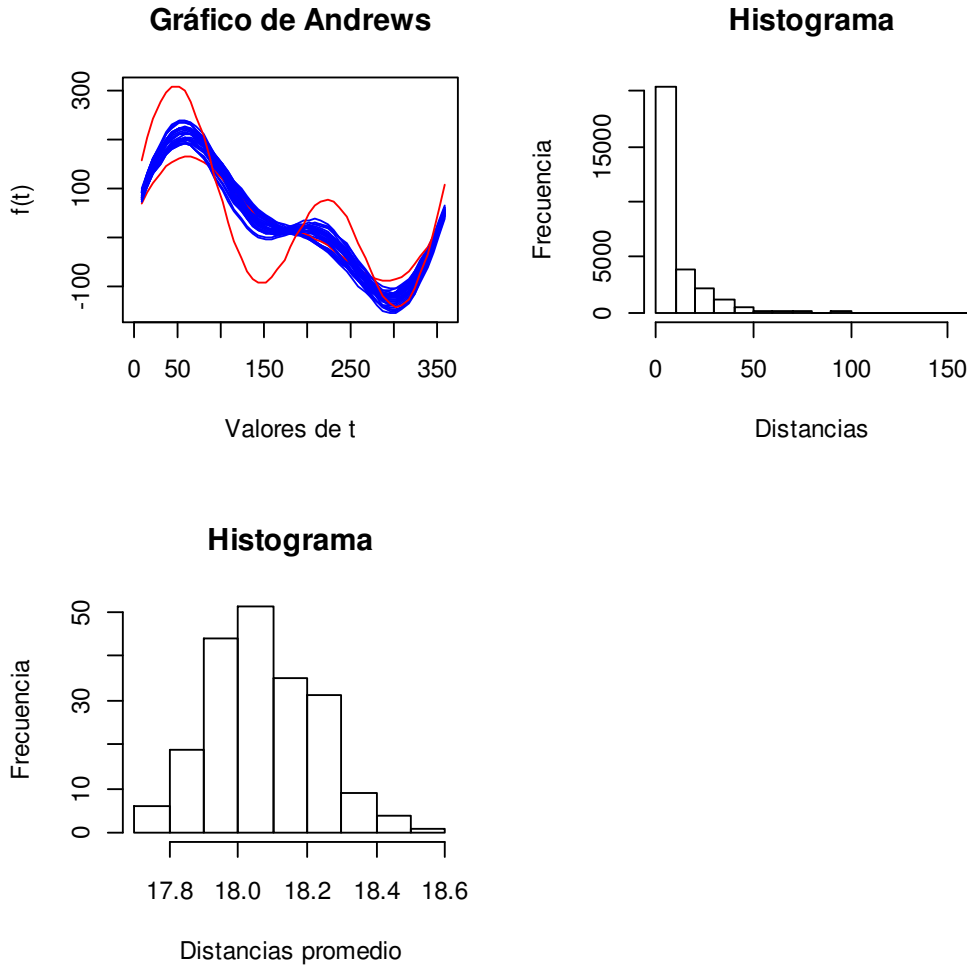


Figura 10. Representaciones de Fourier y distribución de distancias para el tercer conjunto de datos

Es este reporte se puede observar el método detectó los dos casos atípicos que se presentan en las observaciones 11 y 31, para los que se tiene un "alpha" estimado de cero, por lo que se concluye que dichas observaciones son atípicas. También se puede apreciar que la distribución de las distancias es asimétrica a la derecha, como era de esperar, y que la distribución de las distancias promedio para 200 muestras Bootstrap, tiene un comportamiento aproximadamente Normal, con un rango aproximado entre 17.8 y 18.6, Como se observa, los promedios sin los datos 11 y 31 son 17.647 y 15.081, respectivamente, están fuera del rango de la distribución de promedios. En los valores T, calculados como referencia, se aprecia que para la observación 31 se tiene un valor de -16.3896 que indica una fuerte diferencia con la tendencia general, y que por tanto este es una observación claramente atípica; mientras que para la observación 11 se tiene un valor -1.8739, el cual

muestra que esta observación difiere de las otras observaciones, no obstante, la diferencia no es tan grande, y si se observa el gráfico se puede apreciar que no hay tantas diferencias. Una inspección más minuciosa de los datos permitirá decidir si se considera a dicha observación realmente atípica.

Con relación a los resultados del gráfico de cajas, se observa que este gráfico no detecta la observación 11 como atípica, en cambio que el algoritmo propuesto si lo hace

4.4 Análisis del cuarto conjunto de datos

El cuarto conjunto de datos (ver Anexo 2) contiene el número de huelgas, las variaciones porcentuales del número de huelgas, las variaciones porcentuales del número de trabajadores comprometidos, y las variaciones porcentuales del número de Horas-hombre perdidas, para el periodo 1996 a 2012. Esta información fue tomada de la página web del Ministerio de Trabajo y Promoción del Empleo. Con estos datos se obtiene el siguiente reporte:

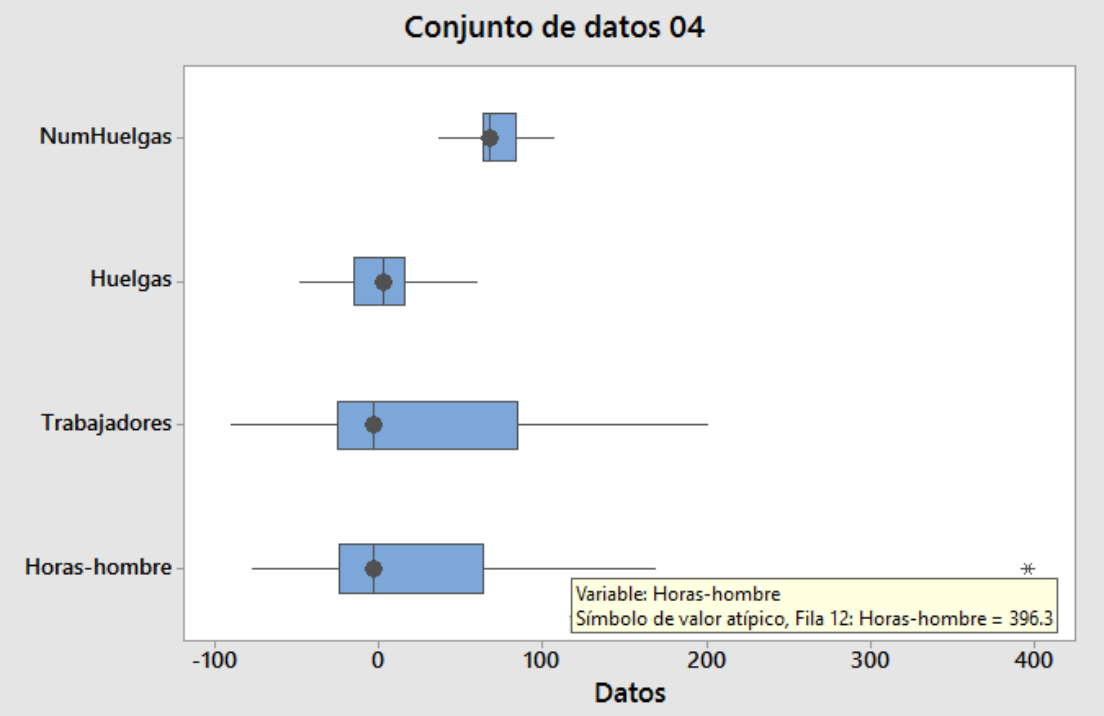


Figura 11. Gráfico de cajas para las variables en el cuarto conjunto de datos

En este gráfico de cajas se precia que la observación 12, correspondiente a la información del año 2007, es considerada atípica solamente para la variable Horas-hombre, y que para las otras variables no se detectan valores atípicos. En general se observa que las distribuciones son asimétricas a la derecha, lo que significa que en algunos años se tuvo variaciones porcentuales más altas de lo usual en lo que corresponde a las variaciones del número de huelgas, las variaciones número de trabajadores comprometidos y las variaciones de horas hombres perdidas.

Con el cuarto conjunto de datos se obtiene el siguiente reporte con el algoritmo propuesto:

```
Promedio total estimado :   94.566236

Valores Alpha para cada observación

Ho : Promedio sin este dato >= Promedio total
H1 : Promedio sin este dato <  Promedio total

   Promedios Alpha Sig        valt
1   97.63213 0.960       1.2283562
2   94.96827 0.585       0.1646705
3   98.50489 0.985       1.6675026
4   88.69403 0.000   *  -2.3658281
5   92.81068 0.145      -0.7330507
6   90.50394 0.000   *  -1.6758166
7   93.30929 0.210      -0.5164829
8   96.90157 0.935       0.9933159
9   96.24164 0.860       0.7098720
10  96.89485 0.935       0.9286014
11  98.44398 0.985       1.5726021
12  73.39383 0.000   * -10.4628215
13  97.34501 0.950       1.1309664
14  96.91019 0.935       1.0008182
15  98.12278 0.965       1.4700708
16  97.55507 0.955       1.2839369
17  98.64007 0.985       1.7411368

Coeficiente de variabilidad:  7.857

Número de muestras:      200
```

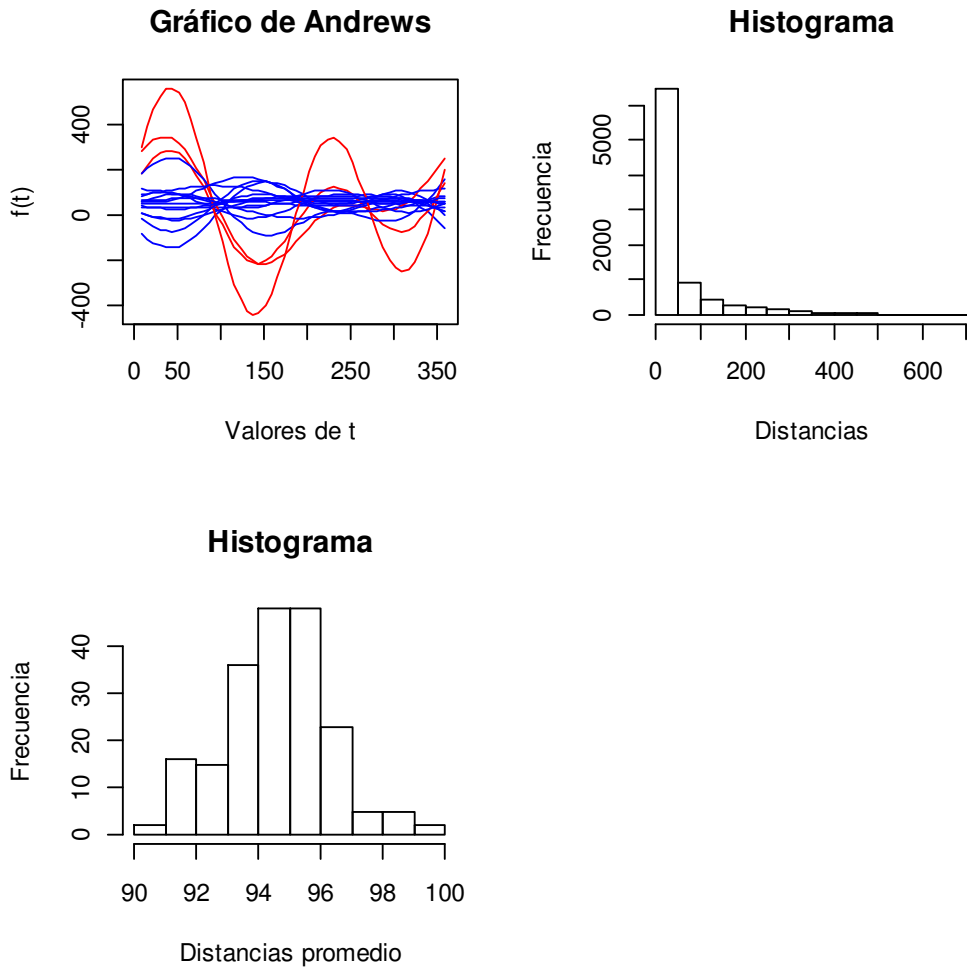


Figura 12. Representaciones de Fourier y distribución de distancias para el cuarto conjunto de datos. Huelgas, Trabajadores comprendidos, y Horas-hombre perdidas. 1996-2012

Es este reporte se puede observar el método detectó tres casos atípicos que se presentan en las observaciones 4, 6 y 12, para los que se tiene un "alpha" estimado de cero, por lo que se concluye que dichas observaciones son atípicas. También se puede apreciar que la distribución de las distancias es asimétrica a la derecha, como era de esperar, y que la distribución de las distancias promedio para 200 muestras Bootstrap, tiene un comportamiento aproximadamente Normal, de manera similar a lo observado en las aplicaciones anteriores. En los valores T, calculados como referencia, se aprecia que para la observación 04 se tiene un valor de -2.3658, para la observación 06 se tiene un valor -1.6758, y para la observación 12 se tiene un valor -10.4628, los cuales muestran que estas

observaciones difieren de la tendencia general. En cuanto a la información contenida en el cuarto conjunto de datos, las observaciones 04, 06, y 12 corresponden a los años 1999, 2001, y 2007, años en los cuales las variaciones porcentuales del número de huelgas, número de trabajadores comprometidos, y del número de Horas-hombre pérdidas fueron significativamente superiores a lo observado en otros años, lo que indica momentos de incertidumbre laboral que difieren lo observado en otros años. La evidencia es muy clara para el año 2007, que muestra un comportamiento muy diferente. Para los años, 1999, y 2001 también se observa diferencias importantes, y para decidir si se trata o no de observaciones atípicas es necesario hacer una inspección más minuciosa de lo ocurrido en dichos años. Para el año 2001, el método no proporciona una evidencia clara, pero si brinda información acerca de la posibilidad de que dicha observación presente realmente un comportamiento atípico que pueden alterar un estudio de la tendencia.

Con relación a lo observado en el gráfico de cajas, se tiene una diferencia con respecto a las observaciones 04 y 06 que no son considerada como atípicas con estos gráficos.

V. CONCLUSIONES

Sobre la base de las pruebas realizadas y los resultados obtenidos se puede concluir lo siguiente:

1. El método sugerido si permite detectar datos atípicos multivariados.
2. Las representaciones gráficas de Andrews (1972), obtenidas mediante series finitas de Fourier constituyen una excelente representación bidimensional que permite una apreciación visual de la similitud entre observaciones multivariadas.
3. El método no paramétrico propuesto es simple en su concepto, pero requiere de un gran procesamiento de cálculo numérico.
4. El método puede ser no eficiente cuando las observaciones multivariadas son semejantes, sin la presencia real de datos atípicos. Cuando las distancias entre las representaciones gráficas muestran una gran homogeneidad puede ocurrir que pequeñas diferencias podrían hacer que el método propuesto reporte a las observaciones como atípicas.
5. El método propuesto también podría ser utilizado para generar una metodología de agrupación de observaciones multivariadas.
6. El método sugerido puede ser aplicado a datos medidos en una escala ordinal, por intervalos o de razón. La razón de esto es que los valores de las variables son utilizados como coeficientes en la serie finita de Fourier propuesta por Andrews (1972). Una aplicación del método puede ser en la detección de datos atípicos cuando se hacen mediciones en escala Likert.
7. Una dificultad para usar el método sugerido es la capacidad de cómputo disponible debido a que las matrices que se generan en el procedimiento crecen exponencialmente con la cantidad de datos a analizar.

VI. RECOMENDACIONES

Sobre la base de las pruebas realizadas y los resultados obtenidos se puede recomendar lo siguiente:

1. Las representaciones gráficas de Andrews (1972), obtenidas mediante series finitas de Fourier pueden ser susceptibles al orden en que se pongan las variables incluidas en el análisis. Por ello se recomienda que las primeras columnas del conjunto de datos estén asociadas a las variables más importantes en la investigación.
2. Complementar el método con la apreciación de las representaciones gráficas para decidir si hay o no una real diferencia.
3. Evaluar la presencia de datos atípico haciendo uso de varios tamaños de muestra para evaluar la consistencia de los resultados.

VII REFERENCIAS BIBLIOGRÁFICAS

Acuña, E., & Rodriguez, C. (2004). A Meta analysis study of outlier detection methods in classification. *En línea*, 23. Obtenido de https://academic.uprm.edu/eacuna/paperout.pdf

Alonso, A. (2002). Un ejemplo de bootstrap suavizado. *Lecturas Matemáticas, 23*, 11-24.

Andrews D.F. (1972). Plots of High-dimensional Data. *Biometrics, 28*(1), 125-136.

Barnett, V., & Lewis, T. (1994). *Ouliers in statisctical data.* Toronto: John Wiley & Sons.

Becher, H., Hall, P., & Wilson, S. (1993). Bootstrap Hypothesis Testing Procedures. *Biometrics*, 1268-1272.

Beckman, R., & Cook, R. (1983). Outliers. *Technometrics, 25*(2), 161-163.

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

Benjamini, Y., & Braun, H. (2002). Tukey's Contributions to Multiple Comparisons. (I. o. Statistics., Ed.) *The Annals of Statistics, 30*(6), 1576-1594.

Bickel, P., & Freedman, D. (1984). Asymptotic Normality and the Bootstrap in Stratified Sampling. *The Annals of Statistics, 12*(2), 470-482. Obtenido de http://www.jstor.org/stable/2241388

Boos, D. D. (2003). Introduction to the Bootstrap World. *Statistical Science, 18*(2), 169-174.

Buja A., C. D. (1996). Interactive High-Dimensional Data Visualization". Journal of Computational and Graphical Statistics, Vol. 5, No. 1 (Mar., 1996), pp. 78-99.The American Statistical Association. *Journal of Computational and Graphical Statistics, 5*(1), 78-99.

Canori, C., & Prescott, P. (1992). Sequential Aplication of Wilk's Multivariate Outlier Test. *Journal of de Royal Statistical Society. Series C (Applied Statistics), Vol. 41, N°2*, 355-364. Obtenido de http://www.jstor.org/stable/2347567

Dixon, W. (1950). Analysis of Extreme Values. *The Aol 21, N°4*, 488-506. Obtenido de http://www.jstor.org/stable/2236602

Efron B., T. R. (1994). *An Introduction to the Bootstrap* . New York.: Chapman & Hall/CRC.

Efron, B. (1979). Bootstrap methods: Another look at the . *The Annals of Statistics, 7*(1), 1-26.

Embrechts , P., & Herzberg , A. (1991). Variations of Andrews' Plots. *International Statistical Review, 59*(2), 175-194.

Fienberg, S. (1979). Graphical Methods in Statistics. *The American Statistician, 33*(4), 65-177.

Filzmoser, P. (2004). A Multivariate Outlier Detection Method. (V. U. Technology, Ed.) *Department of Statistics and Probability Theory*, 1-5. Obtenido de http://www.statistik.tuwien.ac.at/public/filz/papers/minsk04.pdf

Flores, J. G. (2005). Aplicación del método Bootstrap al contraste de hipótesis en la investigación educativa. (U. d. Sevilla, Ed.) *Revista de Educación*, 251-265.

Goodchild, N. A., & Vijakan, K. (1974). Significance test in plots of multi-dimensional data in two dimensions. *Biometrics 30*, 209-210.

Grubbs, F. (1969). Procedures for Detecting Outlyng Observations in samples. (L. Taylor & Francis, Ed.) *Technometrics, Vol 11, N°1*, 1-21. Obtenido de http://www.jstor.org/stable/1266761

Hall, P., & Wilson, s. (1991). Two Guidelines for Bootstrap Hypothesis Testing. *Biometrics, 47*(2), 757-762.

Hesterberg, T., Monaghan, S., Moore, D., & Clipson, A. (2003). *The Practice of Business Statistics.* New York: W. H. Freeman and Company.

Holmes, S., Morris, C., Tibshirani, R., & Efron, B. (2003). Bradley Efron: A Conversation with Good Friends. *Statistical Science, Vol. 18, No. 2, Silver Anniversary of the Bootstrap, 18*(2), 268-281. Obtenido de http://www.jstor.org/stable/3182856

Inselberg, A. (1985). The Plane with Parallel Coordinates. *Visual Computer, 1*, 69-91.

Lohr, S. L. (2000). Muestreo: Diseño y Análisis. México: Thomson.

López, A., & Elosua , P. (2004). Estimaciones bootstrap para el coeficiente de determinación: un estudio de simulación". Facultad de Psicología. Universidad del País Vasco. *Revista Electrónica de Metodología Aplicada, 9*(2), 1-14.

Losilla Vidal, J. M. (1994). Herramientas para un laboratorio de estadística fundamentado en Técnicas de Monte Carlo. Barcelona: Universitat Autónoma de Barcelona.

Maravelakis, P., & Bersimis, S. (2009). The use of Andrews curves for detecting the out-of-control variables when a multivariate control chart signals. *Stat Papers (2009) 50:51–65, Springer-Verlag*, 51-56.

Moustafa, R. (2009). QGPCP: Quantized Generalized Parallel Coordinate Plots for Large Multivariates Data Visualization. *Journal of Computational and Graphical Satatistics. Vol. 18, N° 1*, 32-51. Obtenido de http://www.jstor.org/stable/25703552

Muñoz García, J., & Amón Uribe, I. (2013). Técnicas para detección de outliers multivariantes. *Revista de Telecomunicaciones e Informática, 3*(5), 11-25.

Pan, J.-X., & Wang, X.-R. (1994). Unbiasedness of a Multivariate Outlier Test for Elliptically Contoured Distributions. *Lectures Notes-Monograph Series. Vol. 243, Multivariate Analysis and Its Applications., 24*, 457-460. Obtenido de http://www.jstor.org/stable/4355825

Reyes, R., & Ramírez, G. (2002). Prueba Bootstrap para la hipótesis de no preferencia en estudios toxicológicos con variables dicotómicas. *Agrociencia, 36*(3), 329-335,. Obtenido de http://www.redalyc.org/articulo.oa?id=30236306

Rubio, J. (1983). Análisis de Grupos: Una Aplicación para la Evaluación del Sector Agrícola del Perú. *Tesis de grado.*, 6-45.

Sánchez, A. (2012). *Introducción al Bootstrap.* http://www.bubok.es/libros/209771/Introduccion-al-Bootstrap.

Semmar, N., Urien, S., Bruguerolle, B., & Simon, N. (2008). Independent-model diagnostics for a priori identification and interpretation of outliers from a fullpharmacokinetic database: correspondence analysis, Mahalanobis distance and Andrews curves. *PharmacokinetPharmacodyn*, 159-183.

Silverman, B., & Young, G. (1987). To Smooth or Not to Smooth? *Biometrika, 74*(3), 469-479.

Siotani, M. (1959). The extreme value of the generalized distances of the individual points in the multivariate normal sample. *Annals of the Institute of Statistical Mathematics*, 183-204.

Solanas, A., & Sierra, V. (1992). Bootstrap: fundamentos e introducción a sus aplicaciones. *Anuario de Psicología*, 143-154.

Wegman, E. (1990). HGyperdimensional Data Analysis Using Parallel Coordinates. *Journal of the American Statisticas Association, 85*, 664-675.

Wilks, S. S. (1963). Multivariate Statistical Outliers. *Sankhyā: The Indian Journal of Statistics, 25*(4), 407-426.

VIII ANEXOS

Anexo 1 Programas en lenguaje R

```
# datos = Matriz de datos n*m,
# n= Número de observaciones multivariadas de orden "m"
# m= Número de variables o factores en cada dato multivariado.
# datosz = Matriz de datos estandarizados

prepara=function(datos,n,m)
{
datosz=array(0,dim=c(n,m))
for (j in 1:m)
{
datosz[,j]=(datos[,j]-mean(datos[,j]))/sd(datos[,j])
}
return(datosz)
}

# Función de Andrews. Representación de serie finita de Fourier
#
# datos = Vector de una observación multivariada
# p= número de puntos a evaluar entre 0 y 360°

fourier=function(vector,m,p)
{
datos=vector
ft=rep(0,p)
rn=runif(1,0,1)*0.1745
alpha=0
parejas=floor((m-1)/2)

# genera las magnitudes de las proyecciones
for (i in 1:p)
  {
  val = datos[1]/sqrt(2)

  alpha=(i*2*pi+rn)/p
  if (parejas>0)
    {
    for (j in 1:parejas)
    {  val=val+datos[2*j]*sin(j*alpha)+datos[2*j+1]*cos(j*alpha) }
    }
  if ((parejas*2+1)<m)
     { val=val+datos[m]*sin((parejas+1)*alpha) }
```

```
  ft[i]=val
  }
return(ft)
}


# Genera matriz básica de distancias  Dist(i,j,k)
#
# Dist[i,j,k] matriz salida:  i=1,2,…,n;  j=1,2,…n, k=1,2,…, 30
# datos = Matriz de datos n*m,
# n= Número de observaciones multivariadas de orden "m"
# m= Número de variables o factores en cada dato multivariado.
# p= número de puntos a evaluar entre 0 y 360°

basica=function(datos,n,m,p)
{

dista=array(0,dim=c(n,n,p))
proy=array(0,dim=c(n,p))

for (i in 1:n)
{
  datoi=datos[i,]
  bb=fourier(datoi,m,p)
  for (j in 1:p) { q=bb[[j]];proy[i,j]=q }
}
for (i in 1:(n-1))
{
  for (j in (i+1):n)
  {
  if (i != j )  #    !=  es diferente de
    {
    for (k in 1:p)  { dista[i,j,k]= abs(proy[i,k]-proy[j,k]) }
    }
  }
}
return(dista)
}


#  Genera distribución de distancias promedio con "b" muestras
     Bootstrap
#
# dist(i,j,k)=  Matriz de distancias
#               i=1,2,…,n;  j=1,2,…n, k=1,2,…, 30
# n= Número de observaciones multivariadas de orden "m"
# b= número de muestras bootstrap

dispro=function(dista,n,b)
{
muestra={}
for (j in 1:30)
  {
  ww=dista[,,j]
  muestra=c(muestra,ww[upper.tri(ww)])
  }
```

```
m=length(muestra)
estimado=array(0,dim=c(b,2))
data=1:m
for (i in 1:b)
{
  #  selecciona los subíndices de una muestra bootstrap
  mcr=sample(data,m,T)
  mcr=sort(mcr)
  estimado[i,1]=mean(muestra[mcr])
  estimado[i,2]=var(muestra[mcr])
}
return(estimado)
}

# Genera promedio de distancias de "b" muestras Bootstrap,
# excluyendo a la observación "j"
#
# dist(i,j,k)=  Matriz de distancias
#               i=1,2,…,n;  j=1,2,…n, k=1,2,…, 30
# n= Número de observaciones multivariadas de orden "m"
# j= Observación que se excluye del cálculo de la variancia
# b= número de muestras bootstrap

pro1=function(dista,n,j,b)
{
muestra={}
for (k in 1:30)
  {
  ww=dista[-j,-j,k]
  muestra=c(muestra,ww[upper.tri(ww)])
  }
m=length(muestra)

estima=rep(0,b)
resu=rep(0,2)
data=1:m

for (i in 1:b)
{
#  selecciona los subíndices de una muestra bootstrap
  mcr=sample(data,m,T)
  mcr=sort(mcr)
  estima[i]=mean(muestra[mcr])
}
resu[1]=mean(estima)
resu[2]=var(estima)
return(resu)
}



# Genera gráfico de las representaciones de Fourier

# datos(i,j)=  Matriz de datos,   i=1,2,…,n;  j=1,2,…m
# n= Número de observaciones multivariadas de orden "m"
# p= número de puntos de cada representación
```

```
grafica=function(datos,n,m,p,alpha)
{
datoi=array(0,dim=c(m))
proy=array(0,dim=c(n,p))
xt=array(0,dim=c(p))
for (i in 1:p)  { xt[i]=i*360/p }

for (i in 1:n)
{
  datoi=datos[i,]
  b=fourier(datoi,m,p)
  for ( j in 1:p) { proy[i,j]=b[[j]]}
}

ftmin=min(proy)
ftmax=max(proy)
color="blue"
if (alpha[1] < 0.01) { color="red" }
plot(xt,proy[1,],col=color,type="l",
      xlim=c(0,360),ylim=c(ftmin,ftmax),
      xlab="Valores de t",ylab="f(t)", main="Gráfico de Andrews")

for (i in 2:n)
  {
  datoi=proy[i,]
  color="blue"
  if (alpha[i]<0.01) { color="red" }
  lines(xt,datoi,col=color)
  }
}

# Genera reporte de resultados

# distpro= Vector de promedios obtenidos con las muestras Bootstrap
# proest = Vector de promedios sin considerar la observación en
#          evaluación (para probar si es una observación atípica)
# alpha  = El p-valor Bootstrap calculado para la prueba de valores
#          atípicos
# n      = Número de observaciones multivariadas de orden "m"
# valt   = Valor T calculado para la prueba de hipótesis

reporte=function(distpro,proest,alpha,n,valt)
{
Sig=rep(" ",n)
Num=rep(0,n)
bla=" "
prome=mean(distpro)

print(c(" Promedio total estimado :  ",prome),quote=F)
print(bla,quote=F)
print("Valores Alpha para cada observación ",quote=F)
print("                              ",quote=F)
print("Ho : Promedio sin este dato >= Promedio total ",quote=F)
print("H1 : Promedio sin este dato <  Promedio total ",quote=F)
print("                              ",quote=F)
```

```
for (i in 1:n)
{
  Sig[i]=" "
  if (alpha[i] < 0.01) { Sig[i]=" *" }
}
Promedios=proest
Alpha=alpha
bb=data.frame(Promedios,Alpha,Sig,valt)
print(bb)
}


# Programa principal
#
#  b= número de muestras Bootstrap

detecta=function(b)
{
# Lee archivo de datos Excel en formato  CSV, delimitado por comas

datos=read.table(file.choose(), header=TRUE, sep=";")

#  Inicializa  valores.
m=length(datos[1,])
n=length(datos[,1])
p=30

# Obtiene datos estandarizados (opcional)
# datosz=prepara(datos,n,m)
# datos=datosz

par(mfrow=c(2,2))

dista=array(0,dim=c(n,n,30))
proy=array(0,dim=c(n,30))
distpro=array(0,dim=c(b))
proj =array(0,dim=c(n))
valt =array(0,dim=c(n))
estad=rep(0,2)
alpha=array(0,dim=c(n))
prome=array(0,dim=c(b,2))

# Genera las distancias entre las representaciones de Fourier
dista=basica(datos,n,m,p)

# genera una distribución de promedios base a "b" muestras
prome=dispro(dista,n,b)
distpro=prome[,1]
distpro=sort(distpro)

pro=mean(distpro)
vpro=var(distpro)

for (j in 1:n)
{
   estad=pro1(dista,n,j,b)
```

```
   proj[j]=estad[1]
   varj=estad[2]
   sdp=sqrt(varj+vpro)
   k=0
   valt[j]=(proj[j]-pro)/sdp
   for (i in 1:b)
    {
    if ( proj[j] > distpro[i] ) { k=k+1 }
    }
   alpha[j]= k/b
}

reporte(distpro,proj,alpha,n,valt)
varpro= mean(prome[,2])/b
cv=sqrt(varpro)*100/mean(distpro)
cv=round(cv,digits=3)
print("                            ",quote=F)
print(c("Coeficiente de variabilidad: ",cv),quote=F)
print("                            ",quote=F)
print(c("Número de muestras:     ",b),quote=F)
print("                            ",quote=F)

# Genera gráfica de las representaciones de Fourier y de las
# distribuciones: de distancias y de promedios

grafica(datos,n,m,50,alpha)
hist(dista, xlab="Distancias",ylab="Frecuencia", main="Histograma")
hist(distpro,     xlab="Distancias     promedio",ylab="Frecuencia",
     main="Histograma")
}
```

Anexo 2 Datos de prueba

Conjunto de datos N°1.

X1	X2
32.2	25.8
35.2	27.1
34.3	25.7
34.2	26.3
41.3	28.6
40.2	28.2
35.5	27.9
30.7	24.8
30.7	25.3
42.5	28.3
42.7	28.5
39.1	27.5
32.6	26.6
36.5	27.5
48.0	25.1
28.0	29.0

Conjunto de datos N°2.

X1	X2	X3	X4
8.0	15.0	32.0	11.0
9.0	14.0	34.0	12.0
8.0	14.0	30.0	11.0
6.0	16.0	35.0	9.0
15.0	21.0	15.0	20.0
9.0	15.0	32.0	11.0
9.0	16.0	34.0	12.0
8.0	14.0	39.0	11.0
6.0	16.0	35.0	13.0
12.0	5.0	40.0	32.0

Conjunto de datos N°3.

X1	X2	X3	X4
44.5	116.8	12.1	61.9
34.5	124.4	18.3	67.4
46.6	136.0	26.2	55.7
42.6	137.7	17.6	50.0
39.8	102.0	20.7	76.1
46.6	126.6	20.8	78.9
51.1	139.4	26.6	49.4
50.9	132.4	24.9	58.2
48.8	143.9	18.3	81.2
49.8	125.7	14.3	48.2
42.6	108.5	16.4	38.2
42.6	136.3	19.1	47.6
49.3	118.6	14.7	65.0
45.3	146.1	14.2	66.8
45.1	119.4	17.2	65.7
35.5	114.1	16.9	65.1
48.8	137.0	18.8	45.3
40.2	147.5	14.8	53.1
53.4	132.1	18.9	71.0
37.3	106.4	14.0	77.3
42.6	134.0	19.2	51.0
47.0	128.2	21.1	65.9
50.3	131.4	15.9	91.7
58.2	130.2	20.2	78.9
43.3	130.2	23.6	76.0
41.3	137.8	15.9	76.9
33.4	141.5	14.6	73.9
43.6	106.8	24.1	63.2
53.0	135.8	18.9	70.5
50.1	135.0	22.3	80.2
60.0	100.0	65.0	150.0

Conjunto de datos N°4. Huelgas, Trabajadores comprendidos, y Horas-hombre perdidas. 1996-2012.

Año	Huelgas	Variación porcentual		
		Huelgas	Trabajadores comprendidos	Horas-hombre perdidas
1996	77	-24.5	28.6	33.5
1997	66	-14.3	-47.0	-77.2
1998	58	-12.1	-9.7	1.2
1999	71	22.4	200.5	124.1
2000	37	-47.9	-89.9	-74.9
2001	40	8.1	109.3	169.1
2002	64	60.0	107.5	86.7
2003	68	6.3	62.8	-3.4
2004	107	57.4	-21.6	-33.9
2005	65	-39.3	-35.0	-17.8
2006	67	3.1	2.9	-6.7
2007	73	9.0	145.8	396.3
2008	63	-13.7	-29.3	-31.4
2009	99	57.1	6.2	-4.5
2010	83	-16.2	-15.3	-11.9
2011	84	1.2	-12.5	40.6
2012	89	6.0	-3.5	4.4

Fuente: Ministerio de Trabajo y Promoción del Empleo
http://www.trabajo.gob.pe/archivos/file/estadisticas/huelgas/2012/huelgas2012.pdf