**Del Médico, Ana Paula**
**Vitelleschi, María Susana**
*Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística*

# ANÁLISIS PROCRUSTES GENERALIZADO. UNA APLICACIÓN EN EL ÁREA AGRÍCOLA

**RESUMEN:** En diversas situaciones experimentales las observaciones de varias variables sobre un conjunto de individuos u objetos se obtienen a través de distintas condiciones experimentales, temporales o ambientales; pudiéndolas resumir en las denominadas matrices o tablas multivías, en las cuales cada dato es originado por tres modos o vías: individuos x variables x condiciones. En los últimos años se han desarrollado métodos multivariados que permiten analizar tablas de tres o más modos permitiendo recoger la verdadera estructura presente en los datos y generar conclusiones más completas que las obtenidas a través de la aplicación de una técnica de análisis multivariado tradicional que trabaja con tablas de dos modos (individuos x variables). Uno de los métodos que permite abordar la problemática de los datos a varios modos o vías es el Análisis Procrustes Generalizado (APG). En este trabajo se aplica dicha técnica a un conjunto de datos proporcionados por la Estación Experimental Agropecuaria del INTA de Marcos Juárez; que provienen de ensayos comparativos de variedades de trigo pan de ciclo largo, realizados en Corral de Bustos y Cavanagh, campaña 2011/2012. Se evaluaron variables cuantitativas referidas a la calidad y al rendimiento. Lo que constituyó una tabla múltiple de tres modos: individuos x variables x ambientes. El APG permitió realizar un análisis simultáneo obteniendo una estructura consenso capaz de sintetizar toda la información disponible; como así también, posibilitó observar qué variedades estaban más afectadas por el ambiente.

**Palabras claves:** Datos de tres modos; Análisis Procrustes Generalizado; caracterización de trigo pan.

**Abstract:** In numerous experimental situations, there are data sets obtained from the observation of several variables in a group of individuals through different experimental conditions. These data are originated by three ways: units, variables and conditions. There are several multivariate methods for analyzing three way data. Generalized Procrustes Analysis (GPA) is a multivariate method that can work with three way data. This paper aims at characterizing varieties of bread wheat through trials that come from the INTA Experimental

Station in Marcos Juárez. In order to reach a conclusion, quantitative variables in two different environmental situations (Corral de Bustos and Cavanagh) are analyzed. The structure of data is explored using GPA. This method provides useful analytic and graphic tools to characterize varieties of bread wheat.

**Keywords:** Three-way data; Generalized Procrustes Analysis; Characterization of bread wheat.

## I. INTRODUCCIÓN

Las técnicas estadísticas multivariadas posibilitan el estudio simultáneo de un grupo de variables intercorrelacionadas medidas sobre un conjunto de individuos u objetos, permitiendo obtener representaciones simplificadas de bases de datos voluminosas. Dichas técnicas son utilizadas como herramientas para sintetizar la información.

Los datos multivariados son arreglados en una tabla o matriz en la que cada fila corresponde a una unidad de observación y cada columna a una variable en estudio; es decir son "datos de dos modos o vías". Se denomina "modo o vía" al conjunto de índices de la tabla; siendo un modo el conjunto de variables y otro el de las observaciones.

En muchas investigaciones las observaciones de un conjunto de variables sobre un grupo de individuos u objetos pueden presentar diferentes estructuras de comportamiento, asociados principalmente a variables de caracterización como distintas condiciones experimentales, momentos en el tiempo o puntos geográficos, entre otras. Para cada estructura de comportamiento se tiene una tabla de datos de dos modos; por lo tanto, la información puede presentarse en varias tablas de individuos por variables. Estas diferentes estructuras pueden quedar ocultas en los análisis de la información en su conjunto, si son analizadas como datos de dos modos. Por tal motivo, esta información puede ser estudiada desde la óptica de las tablas múltiples; es decir, teniendo en cuenta la existencia de diversos grupos, lo que requiere realizar, por un lado, análisis parciales de cada uno de ellos y, por otro, un análisis global en el que la influencia individual de cada uno de los grupos esté equilibrada.

El Análisis Procrustes Generalizado (APG) es un método que permite analizar tablas múltiples (individuos x variables x condiciones).

En este trabajo se aplica el APG sobre datos que fueron proporcionados por la Estación Experimental Agropecuaria del INTA de Marcos Juárez, sobre diferentes variedades de trigo

pan de ciclo largo.

## II. MATERIALES

Los datos analizados en este trabajo corresponden a un conjunto de 20 variedades de trigo pan de ciclo largo proporcionados por la Estación Experimental Agropecuaria de la ciudad de Marcos Juárez. Los ensayos fueron realizados en campo de productores de las localidades de Corral de Bustos y Cavanagh, durante el ciclo agrícola 2011/2012. Se evaluaron las variables: Rendimiento (REND, Kg/ha), Peso hectolítrico (PESOh, Kg/hl), Proteína en grano (PROTg, %), Rendimiento de harina (RENDh, %), Gluten húmedo (GLUTh, %), Alveograma W (W, 10-4 Julios), Alveograma P/L (PL, mm. de agua) y Volumen de panificación (VOL, $cm^3$). Cada localidad representa un *ambiente*, por lo tanto los modos de la matriz de datos resultante son: variedades, características y ambientes.

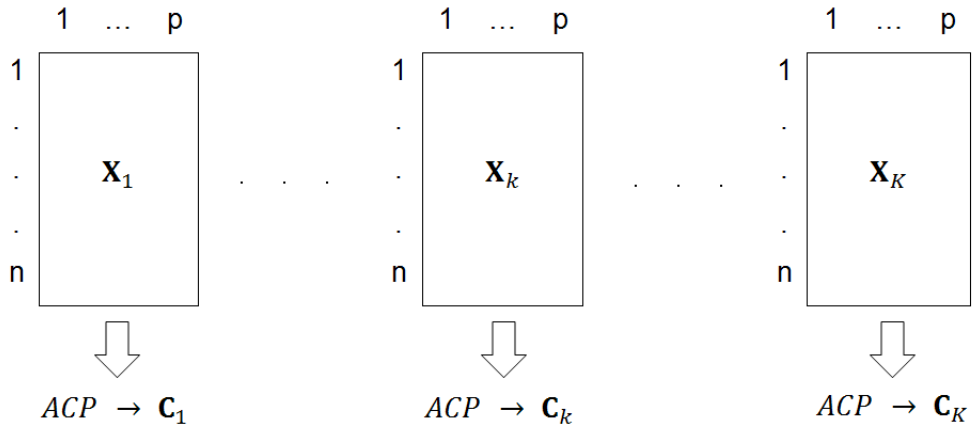
## III. MÉTODOS

El APG, propuesto por Gower en 1975, armoniza las configuraciones individuales (las representaciones geométricas en el plano) a través de pasos algebraicos que transforman a cada configuración individual. Estos pasos incluyen traslación, rotación y escalamiento de las coordenadas de sus puntos mediante dos premisas: mantener la distancia relativa entre los elementos de las configuraciones individuales y minimizar la suma de cuadrados entre puntos homólogos, es decir, puntos que corresponden a un mismo elemento bajo diferentes configuraciones.

La configuración consenso es obtenida como el promedio de todas las configuraciones individuales transformadas.

La información inicial para la aplicación de esta técnica está constituida por *K* tablas de datos conformadas por *n* filas (los individuos) y *p* columnas (las variables). La *k*-ésima tabla se denota con $\mathbf{X}_k$, (*k*=1,2,..., *K*). Se aplica a cada tabla un Análisis de Componentes Principales (ACP) originando nuevas configuraciones que se simbolizan con $\mathbf{C}_k$ (Figura 1). Siendo $\mathbf{C}_k$ una matriz de dimensión *nxp,* donde la *i*-ésima fila corresponde a las coordenadas del *i*-ésimo individuo valorizado en las *p* componentes principales; es decir, proporciona las coordenadas de un punto en los *p* ejes; dicho punto se representa con $P_i^{(k)}$.

***Figura 1:*** *Obtención de las configuraciones iniciales para APG*




Los tres pasos de transformación Procrustes (escalamiento, rotación y traslación), en términos matriciales, pueden ser expresados del siguiente modo:

$$\mathbf{Y_k} = \rho_k \mathbf{C}_k \mathbf{H}_k + \mathbf{T}_k$$

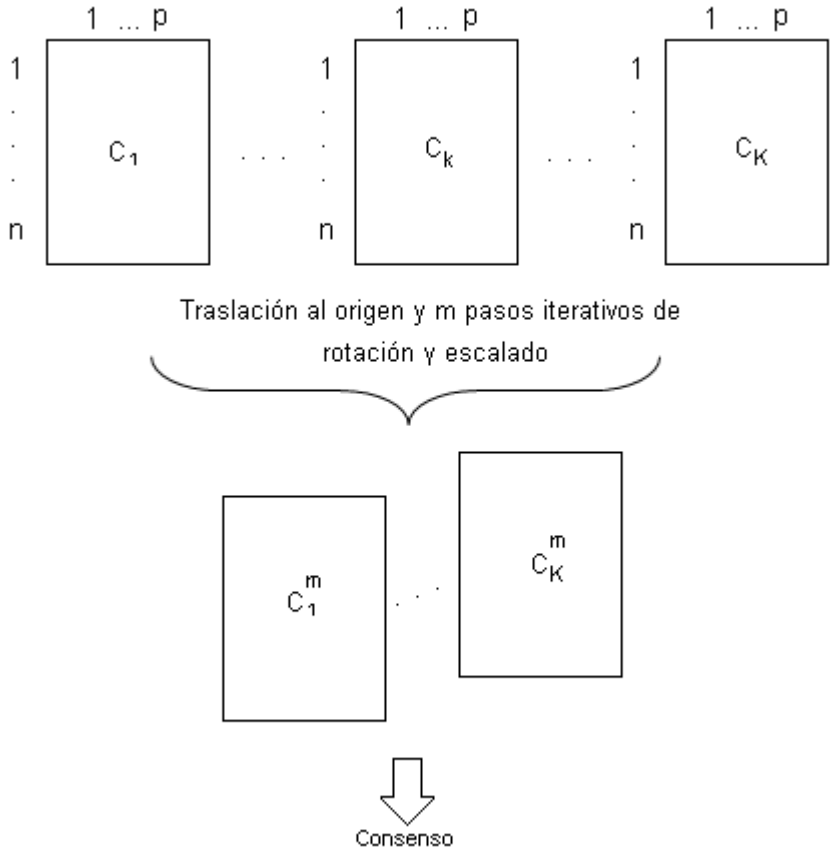
donde $\mathbf{Y_k}$ representa la transformación Procrustes, $\rho_k$ el factor de escala, $\mathbf{H}_k$ la matriz ortogonal de rotación de dimensión *pxp* y $\mathbf{T}_k$ la matriz traslación de dimensión *nxp*. Estos tres últimos elementos son encontrados minimizando la Suma de Cuadrados Residuales (SCR):

$$SCR = \sum_{i=1}^{n} \sum_{k=1}^{K} \Delta^2 \left( P_i^{(k)}, G_i \right)$$

donde $\Delta^2 \left( P_i^{(k)}, G_i \right)$ es la distancia Euclídea entre el punto $P_i^{(k)}$ y el centroide de los *K* puntos análogos $P_i^{(k)}$, denominado $G_i$.

Una iteración es completada una vez que todas las configuraciones se han transformado. Luego, una configuración consenso es calculada como la media de todas las configuraciones individuales transformadas y se inicia una nueva iteración. El proceso se repite hasta que el cambio entre dos pasos consecutivos en las sumas de cuadrados residuales sea menor que un valor prefijado (Figura 2). Una tolerancia de convergencia de 0.0001 se considera satisfactoria (Gower, 1975).

***Figura 2:*** *Obtención de la configuración consenso en APG*




Gower propone al finalizar el APG realizar un Análisis de la Variancia (ANOVA) que permita identificar sobre SCR la participación relativa de los individuos y de las condiciones. La variabilidad total de un APG puede ser particionada en forma de tabla de análisis de la variancia una vez que el proceso iterativo finalice.

## IV. RESULTADOS

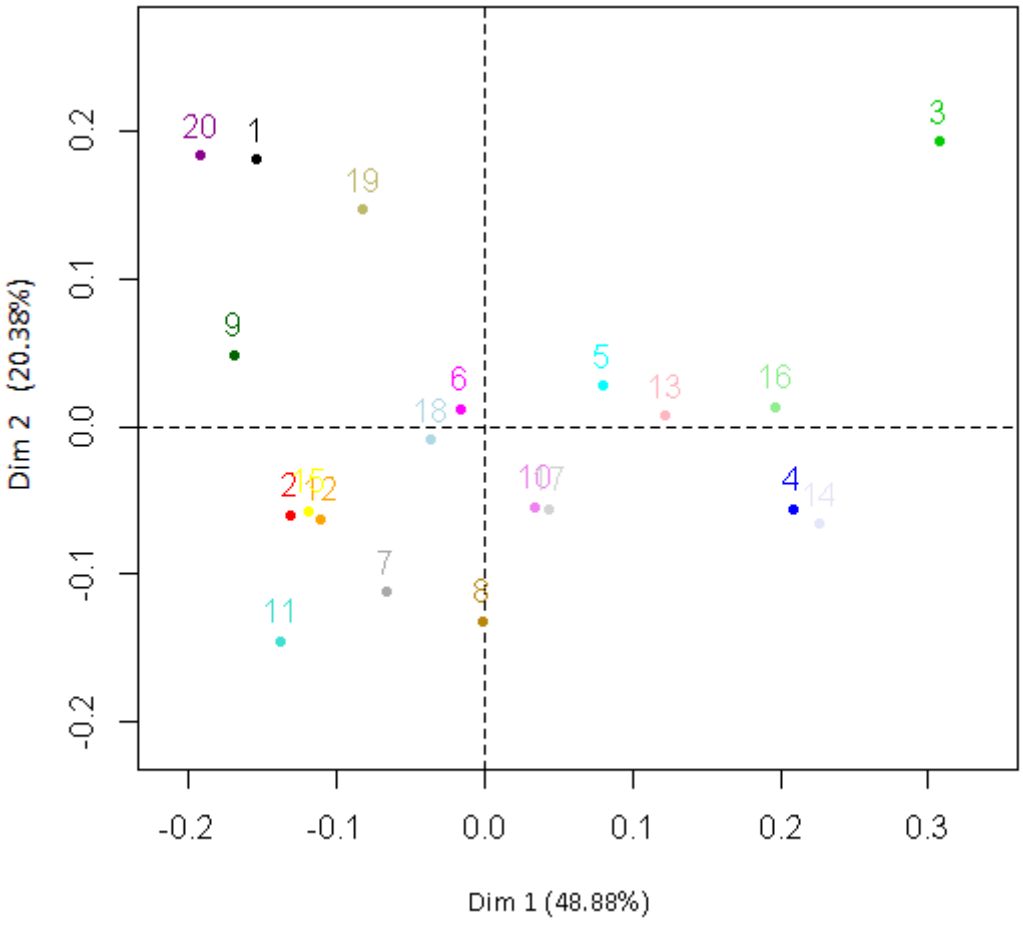
Todos los resultados son obtenidos con el software R (versión 3.0.2).

Se aplica APG a la tabla de tres modos (individuos x variables x ambientes) generada por el conjunto de 20 variedades de trigo pan (individuos) que fueron cultivadas en Corral de Bustos y Cavanagh (ambientes) y a las cuales se les evaluó el rendimiento, el peso hectolí-trico, la proteína en grano, el rendimiento de harina, el gluten húmedo, el alveograma, el

alveograma P/L y el volumen de panificación (variables).

La Figura 3 es la representación bidimensional de la configuración consenso encontrada con la aplicación del APG sobre las dos configuraciones obtenidas. El plano principal representa el 62,26% de la variabilidad total de los datos. Se observa que la variedad de trigo pan 3 se diferencia del resto, ya que se encuentra alejada de las demás variedades. En cambio, algunas variedades de trigo pan tienen sus centroides muy próximos entre sí; es decir, son aquellas variedades con estructura similar, tales como las 2, 12 y 15; 10 y 17; 4 y 14. Por otro lado, el primer eje diferencia las variedades de trigo pan 3, 4, 14 y 16 con las 1, 9 y 20, entre otras. En relación al segundo eje, se pone en evidencia la discrepancia entre las variedades de trigo pan 1, 3, 19 y 20 con las 7, 8 y 11.

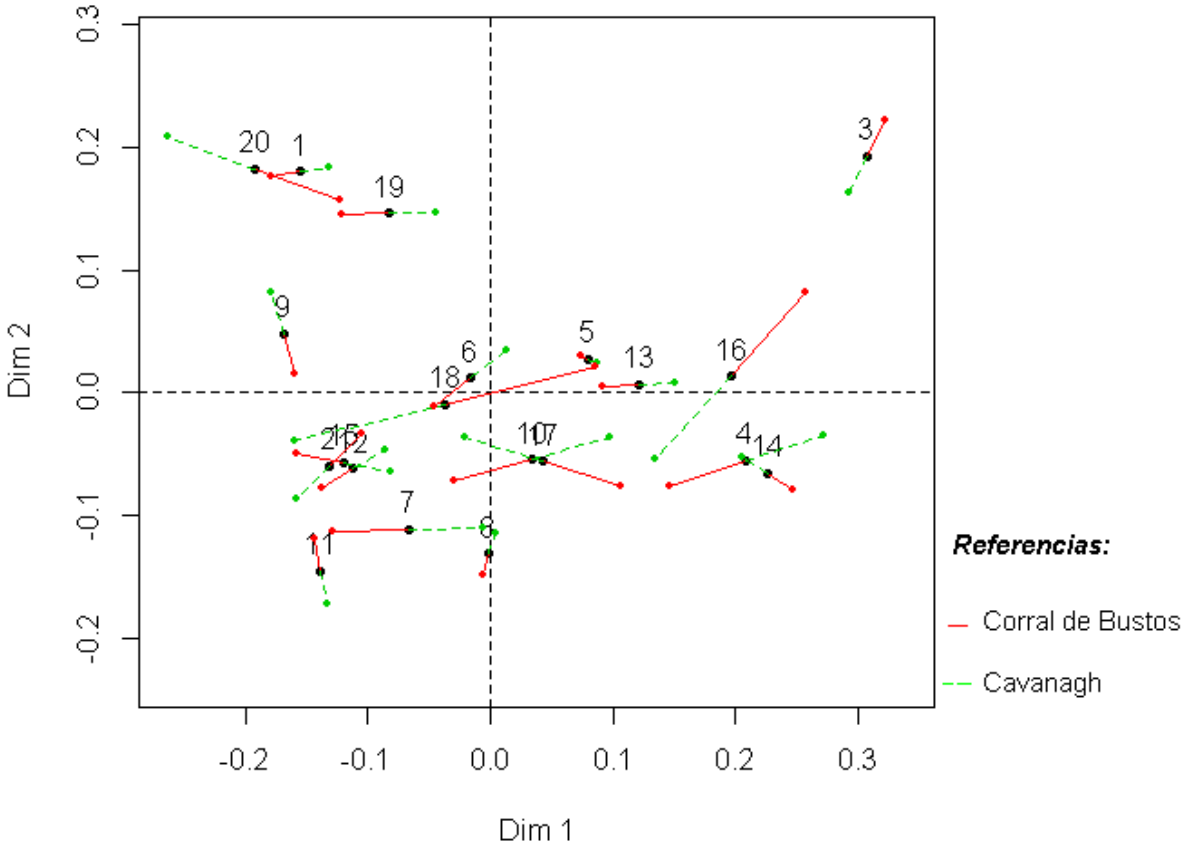
***Figura 3:*** *Representación de los individuos medios sobre los dos primeros ejes principales*




En la Figura 4 se puede observar que las trayectorias son irregulares en magnitud y, en algunos de los casos, muestran considerable dispersión. Las variedades de trigo pan 16, 17, 18 y 20 son las que poseen mayores discrepancias entre las caracterizaciones en ambos

ambientes (mayor efecto de ambiente). Por el contrario, las variedades de trigo pan 1, 5, 12 y 13, mostraron mayor concordancia entre ambas caracterizaciones (menor efecto de ambiente).

***Figura 4:*** *Trayectorias de las Variedades de Trigo Pan*




La diferencia entre las configuraciones originales también puede ser analizada con el aporte de cada variedad de trigo pan a la suma de cuadrados residual en el ANOVA asociado al APG (PANOVA), estos resultados se presentan en la siguiente tabla:

***Tabla 1:*** *ANOVA asociado al APG (PANOVA)*

| ***Variedad*** | ***Consenso*** | ***Residual*** | ***Total*** |
|---|---|---|---|
| 1 | 7.6466198 | **0.2215327** | 7.868152 |
| 2 | 3.6911874 | 0.3917942 | 4.082982 |
| 3 | **16.1718328** | 0.8782296 | 17.050062 |
| 4 | 5.4742564 | 0.5577144 | 6.031971 |
| 5 | 3.0124859 | **0.1286902** | 3.141176 |
| 6 | 2.8500418 | 0.7270257 | 3.577068 |
| 7 | 1.9095987 | 0.5503189 | 2.459918 |
| 8 | 2.6177527 | 0.2839430 | 2.901696 |
| 9 | 3.6463202 | 0.3049202 | 3.951240 |
| 10 | 2.1228657 | 0.7005751 | 2.823441 |
| 11 | 4.7333221 | 0.3875339 | 5.120856 |
| 12 | 2.4360633 | **0.1439917** | 2.580055 |
| 13 | 1.7910686 | **0.1995060** | 1.990575 |
| 14 | 7.9021584 | 0.2701829 | 8.172341 |
| 15 | 2.1648013 | 0.2951807 | 2.459982 |
| 16 | 6.3138007 | **1.2933451** | 7.607146 |
| 17 | 1.7485952 | **1.2743322** | 3.022927 |
| 18 | 0.9041574 | **1.8099449** | 2.714102 |
| 19 | 3.3774556 | 0.3730388 | 3.750494 |
| 20 | 7.9515640 | **0.7422519** | 8.693816 |

Se puede observar en la Tabla 1 que la variedad de trigo pan 3 tiene la mayor suma de cuadrados de consenso, marcando su carácter diferencial con el resto. Por otro lado, a partir de la suma de cuadrados residual, se observa que las variedades de trigo pan 16, 17, 18 y 20 son las que poseen mayores diferencias entre los dos ambientes (mayor efecto ambiente). Por el contrario, las variedades de trigo pan 1, 5, 12 y 13 no presentan diferencias importantes entre ambas caracterizaciones; es decir, serían las poblaciones con un comportamiento más homogéneo, siendo en consecuencia las más estables (menor efecto de ambiente). Por lo tanto, el PANOVA permite corroborar las interpretaciones anteriores.

## V. CONSIDERACIONES FINALES

El tratamiento de tablas múltiples permite un enfoque mucho más completo que el de tablas a doble entrada. En el que cada una tiene identidad propia, esto es, tiene un papel activo en los resultados globales; proporcionando indicadores apropiados para medir las semejanzas y las diferencias entre las estructuras internas de cada uno de los grupos considerados.

El APG se ha convertido en una metodología con una gran versatilidad para el tratamiento de datos de tres modos o vías.

Los resultados obtenidos a través del APG poseen información mucho más rica en relación a la interpretación del efecto ambiente, que las que se hubieran obtenido al analizar las tablas de datos a dos modos con las técnicas tradicionales. Se logró identificar a las variedades de trigo pan que fueron menos afectadas por el ambiente, siendo las mismas 1, 5, 12 y 13, entre otras; como así también a las 16, 17, 18 y 20, entre otras, que resultaron ser más afectadas por el ambiente. También, se pudo determinar que las variedades de trigo pan 2, 12 y 15; 10 y 17; 4 y 14 eran similares entre sí. Y se detectó que la variedad de trigo pan 3 se diferenciaba del resto de las variedades.

En síntesis, el APG posibilitó caracterizar a las variedades de trigo pan, obteniendo una representación superpuesta de las variedades vistas a través de cada ambiente, permitiendo observar qué variedades estaban más afectadas por el mismo. Además, logró determinar qué variedades eran similares entre sí y qué variedades presentaban diferencias.

## REFERENCIAS BIBLIOGRÁFICAS

Bruno, C.; Balzarini, M. (2010). Ordenaciones de material genético a partir de información multidimensional. Revista de la Facultad de Ciencias Agrarias. Universidad Nacional de Cuyo. 42(2): 183-200.

Cuadras, C. (2012). "Nuevos Métodos de Análisis Multivariente". CMC Editions, Barcelona.

Dijksterhuis, G. ; Gower, J. 1991. The interpretation of Generalizad Procrustes Análisis and allied Methods. Food Quality and Preference. p. 67-87.

Gower, J. (1975). Generalized Procrustes Analysis. Psychometrika, 40:33-51.

Gower, J. (2004). The geometry of biplot scaling. Biometrika, 91 705-714.

Gower, J. and Dijksterhuis, G. (2004). “Procrustes Problems”. Oxford: Oxford University Press

Gower, J. (2010). Procrustes methods. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), pp. 503–508.

James, E. G. (2007). Matrix Algebra: Theory, Computations, and Applications in Statistics. Springer, New York.

Johnson, R. and Wichern, D. (2007). Applied Multivariate Statistical Analysis. 6a ed. Pearson Prentice Hall, New Jersey.

Kroonenberg, P. M. (2008). "Applied Multiway Data Analysis". John Wiley & Sons, Inc. Hoboken, New Jersey.

Lê, S.; Josse, J. And Housson, F. (2008). FactorMineR: an R package for multivariate analysis. Journal of Statistical Software, 25 (1):1-18.

Morand E. and Pagès J. (2006). Procrustes multiple factor analysis to analyse the overall perception of food products. Food quality and preference17:36-42.

Mures Quintana, M. J.; Vallejo Pascual, M. E.; García Gallego, A. (2006). "Comparación empírica de técnicas estadísticas para tablas de tres entradas: La construcción en Castilla y León en el período 2002-2004". Pecunia 3: 95-140.

Torcida, S.; Perez, I. (2012). Análisis de Procrustes y el estudio de la variación morfológica. Revista argentina de Antropología Biológica 14,(1): 131-141.

**AGRADECIMIENTOS**

Las autoras agradecen a la Estadística Beatriz Masiero y a la Ingeniera Leticia Mir por brindar la base de datos utilizada en este trabajo.