Componentes principales y coordenadas principales: estudio comparativo basado en una aplicación a la taxonomía numérica

Arce, Osvaldo E. A.[1]; Nora E. De Marco[1]; María R. Santillán[2]

[1] Facultad de Agronomía y Zootecnia.
[2] Facultad de Ciencias Económicas.
Universidad Nacional de Tucumán. E-mail: ova.arce@gmail.com

➤ **Resumen** — Arce, Osvaldo E. A.; Nora E. De Marco; María R. Santillán. 2009. "Componentes principales y coordenadas principales: estudio comparativo basado en una aplicación a la taxonomía numérica". *Lilloa* 46 (1-2). El objetivo del trabajo es realizar un estudio comparativo de las ordenaciones obtenidas mediante la aplicación de componentes principales y coordenadas principales a una matriz de datos mixtos correspondiente a los taxones argentinos del género *Echinochloa* (*Poaceae*), bajo diferentes condiciones de aplicación. Se utilizaron los datos sin estandarizar y estandarizados por desvío estándar o rango. En coordenadas principales se usaron: distancia Euclidiana, disimilaridades Manhattan, Bray Curtis, Canberra y el coeficiente de similaridad de Gower. Para la comparación de resultados obtenidos se emplearon varias técnicas. Los análisis se corrieron en el paquete NTSys. En los casos que fueron necesarios se aplicaron correcciones por autovalores negativos por los métodos de Lingoes y Cailliez. El uso de los diagramas de Shepard y correlaciones entre matrices resultó muy útil para juzgar las ordenaciones. La estandarización resultó el elemento más importante para la obtención de ordenaciones apropiadas. El coeficiente de Gower manejó apropiadamente la naturaleza mixta de las variables. La presencia de autovalores negativos no introdujo distorsiones importantes en espacios de dimensión reducida.

Palabras claves: ordenación, *Echinochloa*, estadística multivariada, autovalores negativos, diagramas de Shepard, NTSys, coeficiente de Gower.

➤ **Abstract** — Arce, Osvaldo E. A.; Nora E. De Marco; María R. Santillán. 2009. "Principal components and principal coordinates: a comparative study based on an implementation to numerical taxonomy". *Lilloa* 46 (1-2). The objective of the present paper is to compare ordinations obtained from principal components and principal coordinates using a mixed data matrix corresponding to the Argentinean taxa of *Echinochloa* (*Poaceae*) under different application conditions. The following coefficients were used in principal coordinates: Euclidean distance, Manhattan, Bray Curtis and Canberra dissimilarities, and Gower similarity coefficient. Unstandardized and range or standard deviation standardized data were used. Ordination comparisons were accomplished using several techniques. All the analyses were run on the package NTSys. Corrections for negative eigenvalues were applied when necessary by means of Lingoes and Cailliez methods. Using Shepard diagrams and matrix to matrix correlations was very useful in order to judge ordinations. Standardization was the most important element to obtain appropriate ordinations. Gower coefficient handled appropriately the variables mixed nature. No important distortions in reduced dimensionality spaces were obtained when negative eigenvalues were present.

Keywords: Ordination, *Echinochloa,* multivariate statistics, negative eigenvalues, Shepard diagrams, NTSys, Gower coefficient.

INTRODUCCIÓN

La diversidad de los organismos biológicos despertó la curiosidad del hombre desde sus inicios. En un principio se comenzaron a observar y diferenciar los animales y plantas. Luego se les pusieron nombres. Cuando el adelanto tecnológico permitió la movilidad de un lugar a otro del planeta el número de seres vivos conocidos se fue incrementando y surgió la necesidad de reunirlos en grupos definidos.

Así es como surge la Taxonomía (palabra de origen griego que significa "ley o norma de ordenación"), que es la ciencia de la clasificación (De La Sota, 1982).

Recibido: 21/08/08 – Aceptado: 27/07/09

Una doctrina dentro de esta ciencia es el feneticismo, el cual se basa en el estudio de las relaciones taxonómicas fenéticas, entendiéndose como tales a aquellos arreglos por similitud total basados en todos los caracteres disponibles para los objetos u organismos bajo estudio sin una ponderación de los mismos (Crisci y López Armengol, 1983; De la Sota, 1982; Sneath y Sokal, 1973).

Debido a que los caracteres empleados en estudios de este tipo deben ser cuantificados con precisión es que al enfoque feneticista de la taxonomía se le ha llamado taxonomía numérica. Esta taxonomía emplea entonces técnicas numéricas, entendiéndose como tales, aquéllas que mediante operaciones matemáticas calculan la afinidad entre unidades taxonómicas a base del estado de sus caracteres.

Al trabajar con caracteres cuantitativos o cualitativos codificados es que la taxonomía debió valerse de las técnicas proporcionadas por la estadística. Por otra parte, como cada entidad está caracterizada por múltiples atributos, es la estadística multivariada la principal proveedora de herramientas de análisis para estudios taxonómicos de tipo numérico.

Las técnicas clasificatorias basadas en información estrictamente numérica comenzaron a desarrollarse a mediados del siglo XX. Sin embargo el gran auge de las mismas se da con la difusión masiva de las computadoras a fines de los ‘80. Numerosos paquetes estadísticos han sido desarrollados desde entonces, lo cual ha puesto estas herramientas de análisis a disposición de toda la comunidad científica y técnica.

Entre todas las técnicas usadas en taxonomía numérica, los métodos basados en autovalores y autovectores (ordenación) tales como componentes principales (Peña, 2002; Hair *et al.*, 1999; Legendre y Legendre, 1998; Gnanadesikan, 1997; Jobson, 1992; Johnson y Wichern, 1992; Everitt y Dunn, 1991; Jolliffe, 1986; Anderson, 1984; Dillon y Goldstein, 1984; Seber, 1984; Karson, 1982; Mardia *et al.*, 1979; Morrison, 1967), y coordenadas principales (también llamado escalado multidimensional métrico) (Peña, 2002; Legendre y Legendre, 1998; Legendre y Anderson, 1998; Jobson, 1992; Jolliffe, 1986; Seber, 1984; Gower y Digby, 1981; Gower, 1966) han sido ampliamente usados. Otras técnicas han sido utilizadas también, aunque en menor grado, como ser análisis de factores, escalas multidimensionales no métricas y análisis canónico.

En taxonomía numérica estas técnicas se usan para obtener grupos a partir de representaciones gráficas bi y/o tridimensionales, es decir, se usan como una alternativa al análisis de conglomerados (“cluster” análisis) y con fines clasificatorios.

La calidad de las representaciones gráficas obtenidas a partir de los métodos de ordenación ha sido motivo de numerosos trabajos. Al ser dichas representaciones en espacios de dimensión reducida el elemento que el investigador tomará en cuenta para extraer conclusiones válidas sobre su trabajo, es que se debe asegurar que éstas sean lo más fieles posibles a las matrices de proximidades en las que éstos se basan.

Moss (1968) fue uno de los primeros autores en plantear que la aplicación de diferentes técnicas puede conducir a resultados diferentes y, en consecuencia, a conclusiones taxonómicas distintas. Realiza estudios comparativos aplicando diferentes técnicas (componentes principales, coordenadas principales y escalas multidimensionales no métricas) a los mismos datos.

Las primeras presentaciones de estos métodos y discusión de sus propiedades se presentan en el clásico libro de Sneath y Sokal (1973).

Rohlf (1972) trabaja en la comparación de distintos métodos de ordenación y usa algunos datos simulados muy simples. Su interés radica principalmente en el efecto de los datos faltantes. Propone algunas medidas que pueden ser usadas para evaluar la calidad de las representaciones gráficas.

Thorpe (1980) trabaja sobre razas de la serpiente *Natrix natrix* y compara varios métodos de ordenación: componentes principales, coordenadas principales, escalas multidimensionales no métricas. Los datos con los que trabaja representan un modelo taxo-

nómico conocido. Llega a la conclusión que la estandarización es recomendable.

Pimentel (1981) realiza un trabajo similar al de Thorpe para especies de *Abronia*. Trabaja con componentes principales, coordenadas principales, escalas multidimensionales no métricas y mapeo linear. Aplica el coeficiente de Gower.

Hartmann (1988) compara métodos de ordenación empleando datos de dientes de *Homínidos*. Analiza el efecto de la estandarización de datos y propone comparar el ajuste entre diferentes métodos usando coeficientes de correlación entre matrices de distancias y matrices derivadas a partir de los puntos en los espacios de dimensión reducida.

En todos los casos se concluye que las diferentes combinaciones de tipos de datos, estandarización, coeficientes de disimilaridad / similaridad y técnica empleada pueden producir distintos resultados.

En los trabajos mencionados previamente los autores trabajaron con variables de tipo mixto, es decir, datos con variables cuantitativas y cualitativas codificadas. Según Thorpe (1980) la codificación otorga a los datos no numéricos un carácter de numéricos y propone estandarizarlos como si se trataran de variables de este tipo. Pimentel (1981) realiza su estudio sobre *Abronia* aplicando un concepto similar.

Crisci y López Armengol (1983) aplican componentes principales a una matriz de especies del género *Bulnesia* constituida por 23 variables cuantitativas y 20 categóricas codificadas. También trabajan con las variables codificadas como si se trataran de variables numéricas. El uso de matrices de datos que contienen variables de tipo mixto es muy común en estudios de taxonomía numérica, no habiéndose dado la importancia que el tema tiene en la bibliografía existente sobre análisis estadístico multivariado.

Componentes principales y la mayoría de las medidas de di/similaridad existentes, empleadas en coordenadas principales, no han sido diseñadas para manejar matrices de datos mixtos. Gower (1971) presenta una alternativa, la única encontrada por los autores, para manejar datos de este tipo.

En trabajos posteriores (Peña, 2002; Legendre y Legendre, 1998; Gower y Legendre, 1986; Gower, 1985) se comenzó a dar importancia a las propiedades matemáticas de las matrices obtenidas a partir de distintos coeficientes de disimilaridad o similaridad, ya que la metricidad y euclinidad de los mismos son esenciales para la obtención de representaciones apropiadas de los datos en espacios de dimensión reducida. Se discute asimismo como la estandarización por rango puede llevar a la euclinidad a ciertos coeficientes de disimilaridad.

Legendre y Legendre (1998) proponen dos métodos, Cailliez (1983) y Lingoes (1971), para corregir la presencia de autovalores negativos y asegurar la euclinidad de disimilaridades y similaridades no métricas. Legendre y Anderson (1998), desarrollan un paquete de software (DistPCoA) para aplicar estas correcciones.

Bramardi (2000) y Rohlf (1990) presentan la técnica del árbol de distancia mínima ("minimum-length spanning tree") como una manera adicional de evaluar la calidad de las representaciones gráficas al superponerlo sobre la representación obtenida a partir de cualquier ordenación.

Rohlf (2009) desarrolla la versión 2.2 del paquete *NTSys-pc* para su aplicación en problemas de taxonomía numérica incorporando todos los procedimientos que aparecen en la bibliografía sobre el tema desde los mencionados por Sneath y Sokal (1973) hasta los más modernos, como análisis de "procrustes".

En este trabajo se realizaron todos los análisis con dicho paquete estadístico.

El objetivo general del presente trabajo es realizar un estudio comparativo de las ordenaciones obtenidas mediante la aplicación de las técnicas de componentes principales y coordenadas principales a una matriz de datos correspondiente a los taxones argentinos del género *Echinochloa* (*Poaceae*: *Panicoideae*: *Paniceae*), bajo diferentes condiciones de aplicación.

Los objetivos parciales son:

– Analizar el efecto de la estandarización de datos por desvío estándar o rango sobre

todas las variables y sólo sobre variables cuantitativas en una matriz de datos mixtos.

– Estudiar distintos coeficientes de disimilaridad y similaridad y su efecto en las representaciones gráficas obtenidas.

– Evaluar los resultados obtenidos al aplicar las técnicas de componentes principales y coordenadas principales.

– Comparar los resultados obtenidos a partir de distintos métodos de corrección de autovalores negativos.

– Evaluar las soluciones obtenidas en espacios de dimensión reducida, a partir de la aplicación, a la matriz de datos, de diferentes combinaciones de centrado, estandarización, coeficientes de disimilaridad / similaridad y técnica de análisis empleada.

METODOLOGÍA

Las evaluaciones metodológicas generalmente involucran simulaciones hechas con computadoras o el estudio a partir de datos reales que tienen una estructura taxonómica conocida. La importancia de evaluaciones del último tipo radica en el hecho de que modelos matemáticos generados no producen información que pueda ser justificada biológicamente (Pimentel, 1981).

Thorpe (1980) dice que si se parte de datos generados por computadora, las técnicas bajo estudio serán la única base para la construcción del modelo taxonómico creando de esta manera una lógica circular.

Por este motivo se trabajó con datos reales con una estructura taxonómica conocida. En este estudio el modelo taxonómico conocido corresponde a De Marco (2002), es decir, la estructura de agrupamientos en la matriz de datos se conocía con anterioridad a su análisis estadístico.

Modelo taxonómico.— El género *Echinochloa* en la Argentina (De Marco, 2006; Zuloaga *et al*., 1994), se encuentra representado por 7 especies y 2 variedades: *E. colona*, *E. crusgalli* var. *crusgalli*, *E. crusgalli* var. *mitis*, *E. cruspavonis*, *E. chacöensis*, *E. helodes*, *E. oryzoides*, *E. polystachya* var. *polystachya*, *E. polystachya* var. *spectabilis*.

Los caracteres exomorfológicos analizados y considerados como relevantes para delimitar taxones, determinaron la existencia de dos grandes grupos, el primero de los cuales se encuentra constituido por las siguientes entidades: *E. colona* (C), *E. crusgalli* var. *crusgalli* (VC), var. *mitis* (VM), *E. cruspavonis* (CR), *E. chacoensis* (CH); y el segundo conformado por *E. oryzoides* (O), *E. helodes* (H), *E. polystachya* var. *polystachya* (VP) y var. *spectabilis* (VS). Los caracteres de separación para estos grupos son: la propagación vegetativa que es cespitosa o rizomatosa, el ciclo de la planta que es anual o perenne y la longitud de la espiguilla que varía de 2-5 mm o de (4,5) 5-7 mm de longitud.

El análisis de otros caracteres como la forma de la espiguilla, su longitud y el ápice de la lemma inferior permiten delimitar los siguientes subgrupos dentro del primer gran grupo: El subgrupo (C), evidencia uniformidad y está definido claramente el carácter predominante que es la presencia de lemma mútica o mucronada. El subgrupo (CR) se manifiesta en forma homogénea y separado del anterior, los caracteres que contribuyeron son la forma de la espiguilla, que es lanceolada y la lemma inferior aristada. El otro subgrupo está formado por las entidades VC y VM que no se separan claramente por los caracteres exomorfológicos, lo cual sugiere que pudo haber procesos de hibridación entre ellos.

Dentro del segundo gran grupo se encuentra los siguientes subgrupos: el subgrupo O en el que los caracteres que contribuyeron a su separación son la longitud de la espiguilla y lemma inferior con arista hasta de 3 cm de longitud. El subgrupo H presenta uniformidad, con sus espiguillas lanceoladas, lemma inferior aristada y su inflorescencia linear y nutante. Los subgrupos formados por las entidades VP y VS, en donde caracteres como el de nudos y vainas glabras, nudos setosos-hirsutos y vainas hirsutas los delimitan. Sin embargo, a pesar de que poseen caracteres diferenciales, son muy similares entre sí.

Puesto que VC y VM no se diferencian claramente, a los fines de este trabajo se los considerará como un grupo único. Por lo

tanto el número de grupos en estudio es de 8.

Se evaluaron 10 individuos de cada uno de los 9 taxones. Se midieron 9 variables cuantitativas, 11 binarias y 10 cualitativas multiestado. La descripción de las variables se encuentra en De Marco (2002). Las variables cualitativas fueron codificadas mediante códigos numéricos.

En los gráficos de ordenación las referencias de los taxones correspondientes son las siguientes: ● VS, ⋆ VP, ○ C, + O, × H, □ CR, Δ VC, ◊ VM, ∇ CH.

A lo largo de este trabajo se adoptará la convención de asignar los individuos a las *n* filas de la matriz y las variables a las *p* columnas.

El análisis de componentes principales se aplicó a datos centrados, en todos los casos, tal como lo sugieren Legendre y Legendre (1998). Se trabajó con los datos sin estandarizar y estandarizados por desvío estándar o rango. El coeficiente de correlación es la covarianza de las variables centradas y estandarizadas por desvío estándar (Johnson y Wichern, 1992), razón por la cual calcular la covarianza de variables así estandarizadas o la correlación de variables no estandarizadas proporcionará la misma matriz de asociación y, por lo tanto, idénticos autovalores.

La estandarización fue aplicada a todas las variables primero y luego sólo a las variables cuantitativas. Los siguientes coeficientes fueron aplicados: varianza-covarianza y correlación, en componentes principales; y coeficiente de similaridad general de Gower (Bramardi, 2000; Gower, 1971) y distancias Euclidiana, disimilaridades Manhattan, Bray Curtis y Canberra (Gower y Legendre, 1986; Gower, 1985), en coordenadas principales.

Autovalores negativos.— Autovalores negativos pueden generarse al usar medidas de distancia semimétricas o no métricas. También pueden encontrarse durante el análisis de algunas distancias métricas que no garantizan una completa representación Euclidiana (Gower y Legendre, 1986). El problema que surge aquí es que los correspondientes ejes de ordenación serán imaginarios.

Para estos casos existen dos métodos de corrección disponibles que permiten obtener una representación euclídea en todos los casos.

Método de Lingoes:

$$d_{ij}^{'} = \sqrt{d_{ij}^2 + 2c}$$

donde *c* es el valor absoluto del autovalor negativo más grande de la corrida del análisis de coordenadas principales (Lingoes, 1971).

Método de Cailliez:

$$d_{ij}^{'} = d_{ij} + c$$

donde *c* es el mayor autovalor de una matriz no simétrica especial (Cailliez, 1983).

Detalles sobres estos métodos se encuentran en Legendre y Legendre (1998) y se incluyen en el paquete estadístico DistPCoA (Legendre y Anderson, 1998).

Para la evaluación de ordenaciones se utilizaron las siguientes técnicas:

Autovalores y porcentaje de varianza explicada.— Esta es la técnica más usada y ampliamente recomendada en la bibliografía (Hair *et al*., 1999; Legendre y Legendre, 1998; Gnanadesikan, 1997; Jobson, 1992; Johnson y Wichern, 1992; Everitt y Dunn, 1991; Jolliffe, 1986; Anderson, 1984; Dillon y Goldstein, 1984; Seber, 1984; Karson, 1982; Mardia *et al*., 1979; Morrison, 1967) para evaluar representaciones gráficas obtenidas a partir de un análisis de componentes principales.

Es equivalente al coeficiente de determinación r^2 empleado en análisis de regresión.

Si se consideran un espacio de dimensión *m*, con $m << p$, su expresión será:

$$\frac{\sum_{k=1}^{m} \lambda_k}{\sum_{k=1}^{p} \lambda_k}$$

e indica el porcentaje de la variabilidad total explicada por los primeros *m* autovalores.

Normalmente se considera que una representación es apropiada si los dos o tres primeros autovalores explican un alto porcentaje de la variabilidad total.

Legendre y Legendre (1998) indican que esta es también una medida válida en el caso de coordenadas principales ya que, algunas veces, los autovalores obtenidos a partir de coordenadas principales son los mismos (excepto por un factor de escala) que los obtenidos a partir de componentes principales.

Peña (2002) establece que esta medida indica el grado de bondad de ajuste de la representación gráfica en el caso de coordenadas principales, ya que en esta técnica los autovalores no corresponden a varianzas.

El módulo EIGEN de NTSys brinda esta información.

Congruencia entre el modelo taxonómico estudiado previamente y el obtenido mediante ordenaciones.— Esta metodología fue utilizada por Hartman (1988), Pimentel (1981) y Thorpe (1980).

El esquema 1 resume el procedimiento usado.

La congruencia se evaluó determinando si el número de grupos obtenidos mediante las ordenaciones concordaba con el número presente en el modelo taxonómico conocido.

Gráficos en 2 dimensiones con árboles de recorrido mínimo superpuestos.— El árbol de recorrido mínimo se calcula a partir de una matriz de disimilaridades o similaridades. Es útil para su superposición sobre ordenaciones para ayudar a detectar distorsiones locales, es decir, pares de puntos que se ven próximos en un gráfico pero que en realidad están separados si otras dimensiones son tomadas en cuenta (Rohlf, 2009). También se utilizan para una más fácil identificación de agrupamientos.

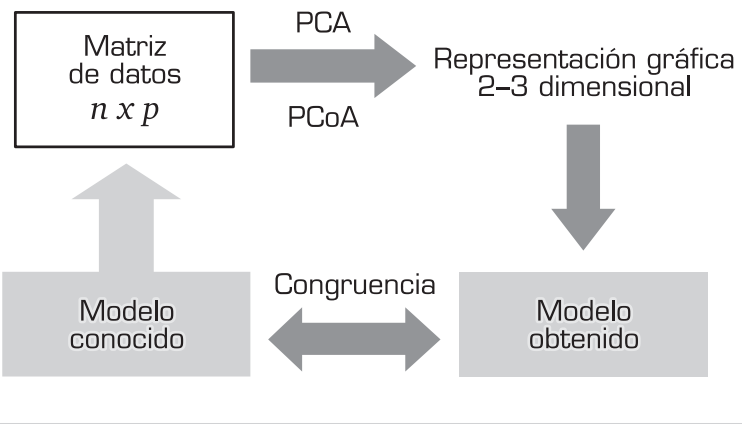
Gower y Ross (1969) muestran que este árbol equivale a obtener un agrupamiento ("clustering") no jerárquico de los objetos por el método de agrupamiento simple.

NTSys incluye estos árboles en el procedimiento MST.

Diagramas de Shepard.— Estos diagramas fueron originalmente propuestos por Kruskal (1964). Consisten en graficar una matriz contra la otra, elemento a elemento, ignorando las diagonales.

McCune y Grace (2002), Gnanadesikan (1997) y Everitt y Dunn (1991) indican que en estos diagramas puede observarse si existe una relación monótona entre las matrices, lo cual indica una apropiada configuración de puntos en el espacio de dimensión reducida. La monotonicidad asegura que las distancias interpuntos recuperadas concuerden con las disimilaridades originales, o sea,

Esquema 1



cuanto mayor sea la disimilaridad entre dos objetos, mayor será la distancia interpunto en la representación euclidiana de estos objetos. Dicho en otras palabras, las relaciones de orden entre las distancias interpuntos en la representación euclidiana está en exacta concordancia con las relaciones de orden entre las disimilaridades originales. Si lo que se compara es una medida de similaridad con la distancia, la relación que se busca es la inversa a la anterior, es decir, a mayor distancia, menor similaridad.

También se puede observar qué tan próxima o similar es la matriz de disimilaridades o similaridades derivada con respecto a la original. Legendre y Legendre (1998) muestran que cuanto más próxima está la nube de puntos a la diagonal del diagrama, más parecidas serán ambas configuraciones.

Además si la nube de puntos está próxima a la diagonal y sigue una tendencia lineal, las posiciones relativas de los puntos habrán sido recuperadas con precisión. A veces la relación entre ambos espacios no es lineal, lo cual indica que las posiciones relativas recuperadas han sufrido una distorsión.

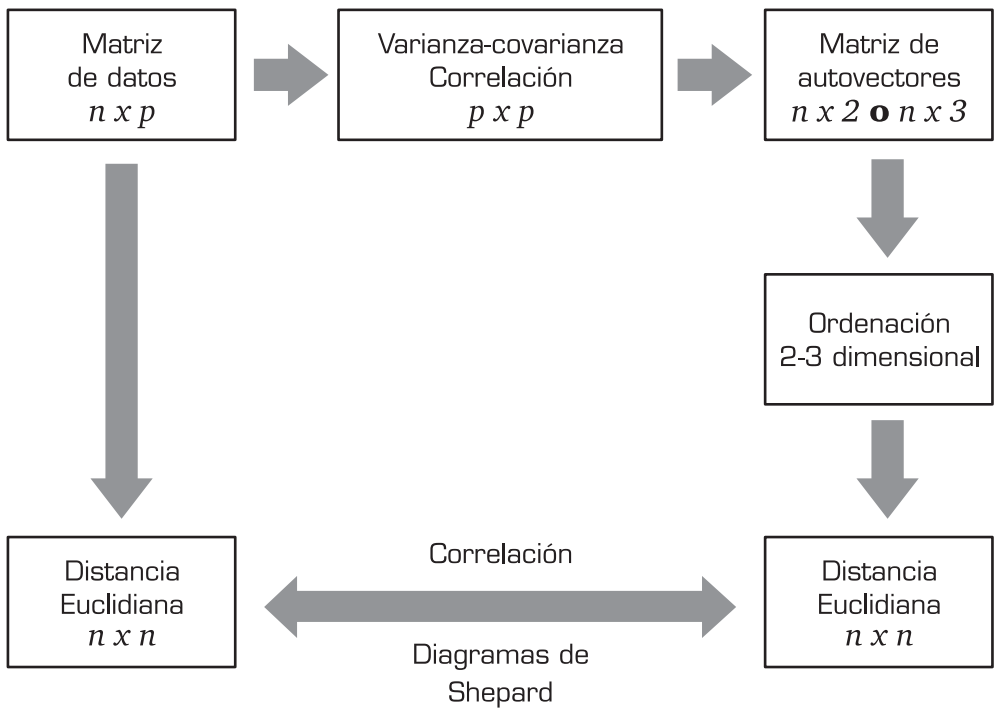
Estos diagramas están incluidos en el procedimiento MXCOMP de NTSys.

Correlación entre matrices.— Es una medida de la correspondencia entre los elementos de dos matrices. Esta técnica es recomendada por Rohlf (1972) para evaluar cuán próxima es la configuración de puntos en espacios de dimensión reducida en relación a la configuración de los mismos en el espacio original *p*-dimensional.

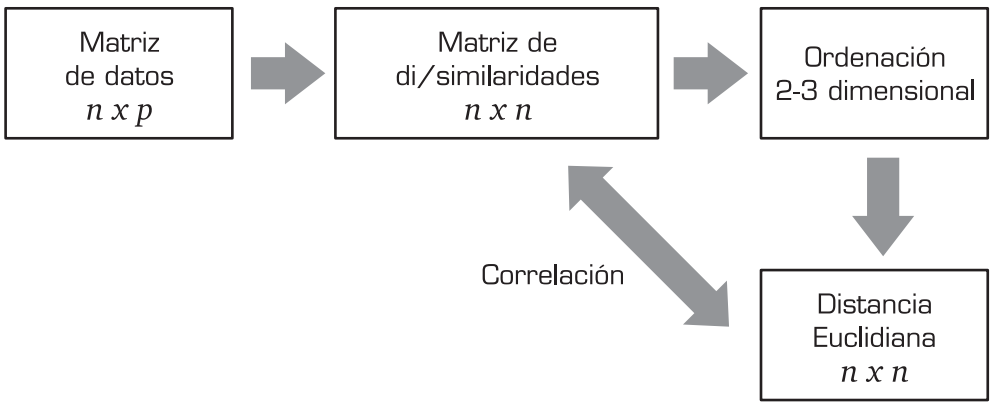
Legendre y Legendre (1998) sugieren utilizar la distancia Euclídea en el espacio original y en el reducido en el caso de componentes principales. En coordenadas principales y escalas multidimensionales no métricas aconsejan calcular las distancias Euclídeas entre los objetos en el espacio reducido y compararlas con las disimilaridades o similaridades en las cuales se basó la ordenación.

Este procedimiento también fue utilizado por Hartman (1988) con el fin de evaluar

Esquema 2



Esquema 3



las correspondencias entre las configuraciones obtenidas a partir de distintos análisis de los mismos datos y con el fin de comparar la similitud entre los resultados obtenidos.

NTSys permite calcular estas correlaciones en el procedimiento MXCOMP.

Concordancia entre el espacio original y el derivado para una ordenación en particular.— Los esquemas 2 y 3 resumen los procedimientos utilizados.

En el caso de componentes principales se trabajó de la manera indicada en el esquema 2.

Coeficiente Codificación	Estandarización	Autovalores	% Var. explicada	Correlación orig.–deriv. en 2 y 3 dim.	Grupos identificados
varcov *varne*	centrado	234.2879 36.9095 22.3580	72.7269 84.1843 91.1246	0.97882 0.98947	0
varcov* *varde*	centrado desvío estándar todas las variables	10.9144 5.2534 3.3983	36.3815 53.8928 65.2203	0.90527 0.93862	8
varcov *vardec*	centrado desvío estándar sólo variables cuantitativas	8.7915 3.6024 2.5782	40.3805 56.9266 68.7684	0.92172 0.95762	8
varcov *varrg*	centrado rango todas la variables	1.8088 0.7888 0.4958	41.6909 59.8711 71.2983	0.91065 0.95244	8
varcov *varrgc*	centrado rango sólo variables cuantitativas	6.4281 2.2339 1.7291	47.4834 63.9853 76.7580	0.90657 0.97270	10

Tabla 1. Autovalores, porcentaje de varianza explicada, correlación entre espacio euclidiano original y derivado en 2 dimensiones, ídem en 3 dimensiones, y número de grupos identificados en la representación bidimensional bajo distintas condiciones de estandarización. Palabras en cursiva corresponden a la codificación empleada en otras tablas y figuras. varcov = varianza–covarianza; corr = correlación; (*) esta opción es equivalente a la correlación entre variables.

	varne	*varde*	*vardec*	*varrg*	*varrgc*
varne	1				
varde	0.20884	1			
vardec	0.28319	0.92527	1		
varrg	0.12544	0.97792	0.89460	1	
varrgc	0.16950	0.79844	0.92389	0.89032	1

Tabla 2. Correlaciones entre distancias Euclidianas en espacios derivados en dos dimensiones. Referencias en Tabla 1.

En coordenadas principales se trabajó tal como lo muestra el esquema 3.

Concordancia entre espacios derivados mediante distintos análisis aplicados a la misma matriz de datos.— Se correlacionaron las matrices de distancia Euclidiana obtenidas a partir de espacios bidimensionales, para las distintas combinaciones estudiadas, a fin de juzgar la similitud entre las soluciones obtenidas (módulo MXCOMP en NTSys). A las matrices de correlaciones obtenidas se les aplicó análisis de conglomerados por el método UPGMA ("unweighted pair-group arithmetic average clustering"), mediante el módulo SAHN de NTSys. El procedimiento se indica en el esquema 4.

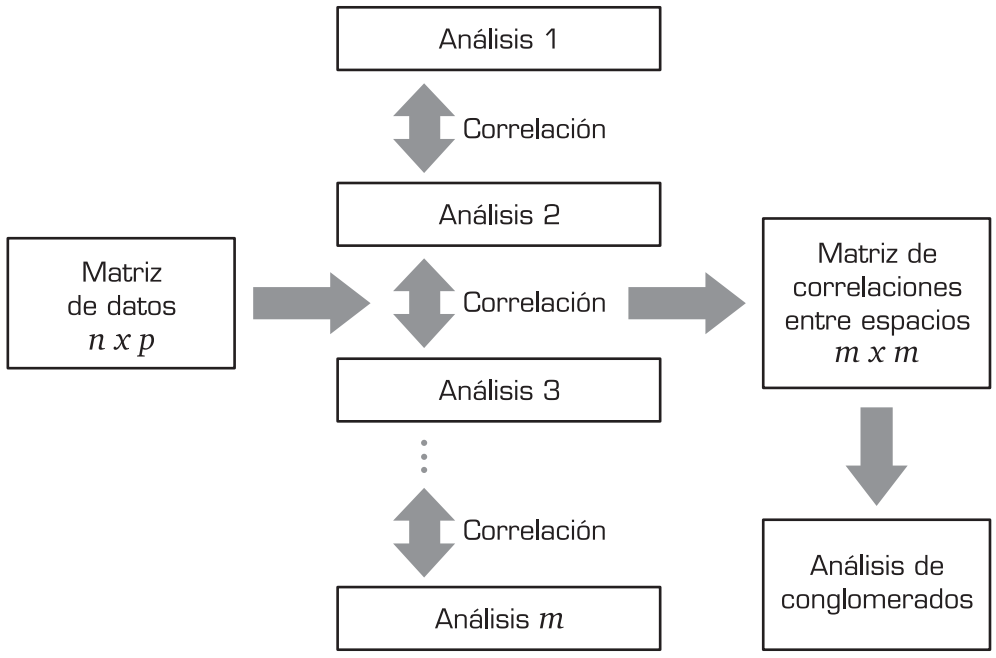
Hartman (1988) utilizó este procedimiento aunque realizó una ordenación sobre las matrices de correlaciones. En este trabajo se obtuvieron mejores resultados con el análisis de conglomerados.

RESULTADOS

Componentes Principales

Se observa en la tabla 1 que el centrado sin estandarización no permitió la identifica-

Esquema 4



ción de los 8 grupos. La estandarización ya sea por rango (todas las variables) o desvío estándar (todas las variables y sólo cuantitativas) permitió una correcta identificación de grupos. La estandarización por rango sólo de variables cuantitativas no resultó apropiada.

La tabla 2 muestra que la correlación entre los espacios euclidianos derivados a partir de datos estandarizados por desvío estándar y rango presentaron un valor de concordancia elevado (r = 0.97792) indicando que las ordenaciones obtenidas fueron muy similares cuando se estandarizó todas las variables. La ordenación con estandarización por rango sólo de variables cuantitativas es la que menor concordancia presentó con las demás. La relación entre espacios resultó fuertemente lineal para valores de r > 0.9. Como ejemplo se presenta un diagrama de Shepard en la figura 1.

La figura 2 presenta una comparación entre distancias Euclidianas en el espacio derivado bidimensional vs. las mismas distancias en el espacio *p*-dimensional original con estandarización de todas las variables. La figura mencionada muestra que los espacios original y derivado no manifestaron una relación lineal pero sí monótona, generándose una distorsión importante en distancias medias y bajas, aunque las relaciones de orden se mantuvieron, por lo que los valores de correlación entre espacios no representan, en este caso, un buen indicador de la calidad de las representaciones obtenidas. La gran estructura de agrupamientos (distancias grandes) está representada con precisión en la figura 5. Pero aquellos puntos que se encuentran próximos en el espacio *p*-dimensional, se verán más próximos en el espacio bidimensional de lo que estaban originalmente.

La figura 2 muestra que al estandarizar sólo las variables cuantitativas, por desvío estándar, la relación entre los espacios es aproximadamente lineal por lo que las dis-

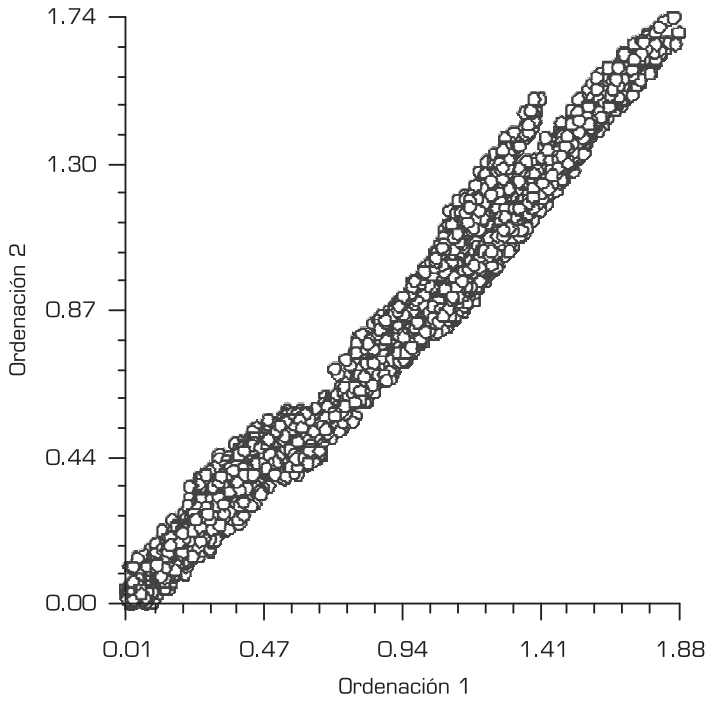


Figura 1. Diagrama de Shepard que muestra la concordancia entre la distancia Euclidiana en espacio bidimensional obtenida por componentes principales de una matriz de correlaciones (Ordenación 1) e igual distancia a partir de una matriz de datos estandarizados por rango (todas las variables). r = 0.97792

tancias se recuperaron con más exactitud en 2 dimensiones

Arce (2002, 2003) realizó un estudio similar al presente sólo con variables cuantitativas y encontró que las relaciones entre espacios fueron fuertemente lineales. La inclusión de variables cualitativas codificadas, tratadas como si se trataran de variables numéricas, seguramente está produciendo este efecto de no linealidad entre espacios.

Las correlaciones entre espacios originales y derivados en 3 dimensiones (tabla 1) resultaron en todos los casos mayores que en 2 dimensiones, ya que al ir aumentando dimensiones nos aproximamos cada vez más al espacio original *p*-dimensional.

Al analizar los porcentajes de varianza explicados por los tres primeros autovalores en la tabla 1 pudo verse que la estandarización produjo una reducción importante en esos valores. Sin embargo, esto no fue un factor determinante en la identificación de los ocho grupos en estudio. Esto está indicando que el porcentaje de varianza explicada no es, por sí mismo, un buen indicador de la calidad de las representaciones gráficas.

El análisis de los porcentajes de varianza explicada y de las figuras 4 y 5 muestra que este no es un buen criterio para juzgar la calidad de las representaciones gráficas, como está ampliamente recomendado en la bibliografía. Las mejores ordenaciones se obtuvieron con proporciones menores de varianza explicada. Se debe tener en cuenta que, en los datos utilizados, las variables estuvieron medidas en diferentes escalas de medida y, por lo tanto, sus varianzas fueron muy distintas, alterando los resultados obtenidos. La estandarización apareció aquí como un elemento clave a tener en cuenta al realizar un análisis de componentes principales. Si todas las variables hubieran estado medidas en iguales unidades y, por lo tanto, sus varianzas hubieran sido similares, la mejor ordenación habría sido la de varianza–covarianza de datos no estandarizados ya

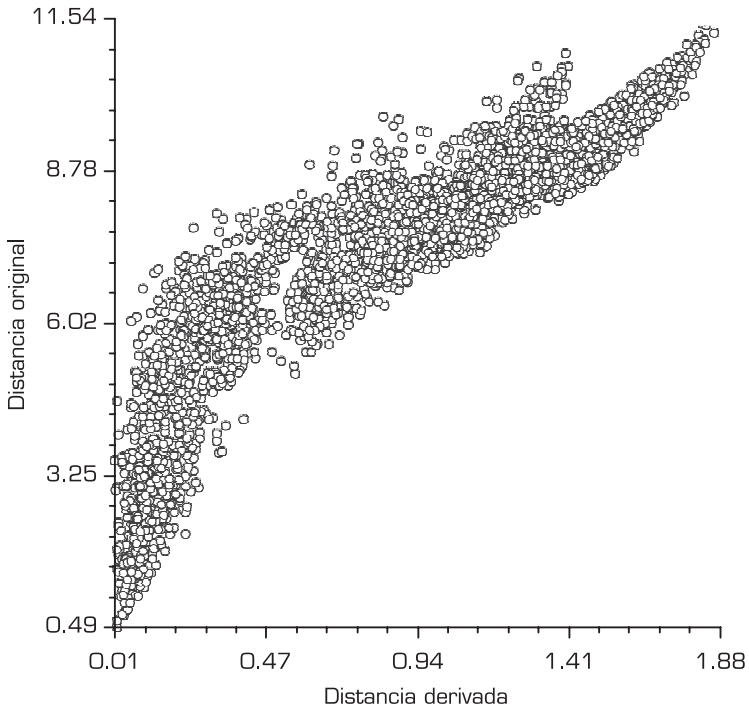


Figura 2. Distancia Euclidiana en espacio derivado bidimensional vs. igual distancia en espacio p-dimensional para componentes principales derivados a partir de una matriz de varianza–covarianza de datos estandarizados (todas las variables) por desvío estándar. r = 0.90527.

Codificación	Estandarización	Autovalores	% Varianza explicada	Correlación euc.–euc. en 2 y 3 dim.	Grupos identificados
eucne	ninguna	20852.6227 3285.9533 1989.8586	72.7269 84.1843 91.1245	0.97882 0.98947	0
eucde	desvío estándar	971.3848 467.5525 302.4450	36.3815 53.8928 65.2203	0.90257 0.93862	8
eucdec	desvío estándar sólo cuantitativas	782.4438 320.6091 229.4570	40.3805 56.9266 68.7684	0.92172 0.95762	8
eucrg	rango	160.9807 70.1990 44.1238	41.6909 59.8711 71.2983	0.91065 0.95244	8
eucrgc	rango sólo cuantitativas	572.0997 198.8215 153.8913	47.4834 63.9853 76.7580	0.90657 0.97570	10

Tabla 3. Autovalores, proporción de varianza explicada, correlaciones entre espacios euclidianos original y derivado en 2 y 3 dimensiones y número de grupos identificados al aplicar coordenadas principales a matrices de distancia Euclidiana. Palabras en cursiva corresponden a la codificación empleada en otras tablas y figuras.

que, como se observa en la tabla 1, es la condición que mejor preservó el espacio original (*r*'s próximos a 1).

Esto se nota claramente en las figuras 4 y 5. La ordenación a partir de varianza-covarianza de datos no estandarizados, que es la que mayor porcentaje de varianza explicada presentó, no logra separar los 8 grupos. La figura 4 muestra el efecto de la estandarización, en la cual se identifican los 8 grupos aún con menor porcentaje de varianza explicada.

Los árboles de distancia mínima superpuestos a las ordenaciones mostraron claramente la estructura de agrupamiento de las cinco poblaciones en las figuras 4 y 5 y permitieron detectar las distorsiones que se produjeron en las posiciones relativas de algunos puntos próximos.

Las estandarizaciones por desvío estándar y por rango (todas las variables) generaron ordenaciones casi idénticas (valores muy próximos a 1 en la tabla 2), al comparar los espacios derivados en dos dimensiones para opciones con desvío estándar vs. aquéllas con rango. Las opciones no estandarizadas presentaron configuraciones muy distintas a las de las opciones con estandarización ($r < 0.62$).

Coordenadas Principales

Distancia Euclidiana.— Existe dualidad entre los resultados obtenidos a partir de componentes principales y coordenadas principales sobre una matriz de distancias euclidianas. La comparación de las tablas 3 y 1 muestra lo mencionado. Los autovalores resultan diferentes pero los porcentajes de variabilidad explicada son los mismos y también las ordenaciones obtenidas. Por este motivo, todo lo enunciado en componentes principales es válido para este caso.

Las posiciones relativas de los objetos resultaron idénticas, no así sus posiciones absolutas. El gráfico de ordenación estuvo rotado 180º con respecto al de componentes principales. No se presenta el gráfico de la ordenación de coordenadas principales, pero sí el de disimilaridad Manhattan que fue similar. La comparación de las figuras 5 y 7 permite verificar esta situación.

Codificación	Estanda-rización	Autovalores	% Varianza explicada	Aut. negativos (autovalor más pequeño)	Correlación manh-euc en 2 y 3 dimensiones	Grupos identificados
manhne	ninguna	77.1890 45.5673 21.8463	47.3160 75.2483 88.6390	si (-6.6148)	0.91653 0.95017	0
manhde	desvío estándar	27.7460 10.5028 5.1645	54.6451 75.3302 85.5015	si (-0.8935)	0.94691 0.98749	8
manhdec	desvío estándar sólo cuantit.	18.8519 6.1158 3.4929	55.1748 73.0743 83.2973	si (-0.6585)	0.94022 0.96792	8
manhrg	rango	4.1112 1.4699 0.6892	57.9395 78.6547 88.3679	si (-0.1454)	0.94963 0.98074	8
manhrgc	rango sólo cuantit.	10.7183 2.9134 1.8804	60.3417 76.7435 87.3297	si (-0.4959)	0.94909 0.97164	8

Tabla 4. Autovalores, porcentaje de varianza explicada, presencia de autovalores negativos, correlación entre distancia original y derivada en 2 y 3 dimensiones, correlación entre distancia original y distancia Euclidiana derivada en 2 y 3 dimensiones y correlación entre distancia Euclidiana original y distancia Euclidiana derivada en 2 y 3 dimensiones, y número de grupos identificados en la representación bidimensional bajo distintas condiciones de estandarización. Palabras en cursiva corresponden a la codificación empleada en otras tablas y figuras.

Disimilaridad Manhattan o City Block.— La columna 6 de la tabla 4 indica la concordancia entre el espacio euclidiano original y la disimilaridad Manhattan en el espacio p-dimensional, que fue razonablemente buena ($r > 0.90$) para los casos con variables estandarizadas.

La tabla 5 muestra que las correcciones por autovalores negativos produjeron ordenaciones similares a las opciones no corregidas (r próximo a 1). En la tabla 4 se ve que los autovalores negativos tomaron valores muy pequeños en relación a los tres primeros valores propios positivos, razón por la cual los efectos de la corrección son mínimos.

Las figuras 6 y 7 ponen en evidencia el efecto de la estandarización en la identificación de los grupos.

Disimilaridad de Bray Curtis.— Se trabajó sólo con la transformación a raíz cuadrada, que tiene propiedades métricas (no genera autovalores negativos), y con datos no centrados. Esta medida de disimilaridad no tolera el centrado, produciendo resultados absurdos en ese caso tal como fue estudiado por Arce (2003). La ordenación sólo pudo identificar 2 grupos (tabla 6). La autonormalización incluida en este coeficiente no fue suficiente para lograr una correcta identificación de grupos, como se observa en la figura 8. La ordenación obtenida fue similar a la que se obtiene cuando no se estandarizan las variables.

Disimilaridad Canberra.— La disimilaridad Canberra, autonormalizada al igual que la disimilaridad de Bray Curtis, generó un espacio similar al de las variables estandarizadas. Para la identificación de los 8 grupos fue necesario usar 3 dimensiones (figuras 9 y 10).

El diagrama de Shepard correspondiente no se incluye por ser muy similar a los presentados anteriormente.

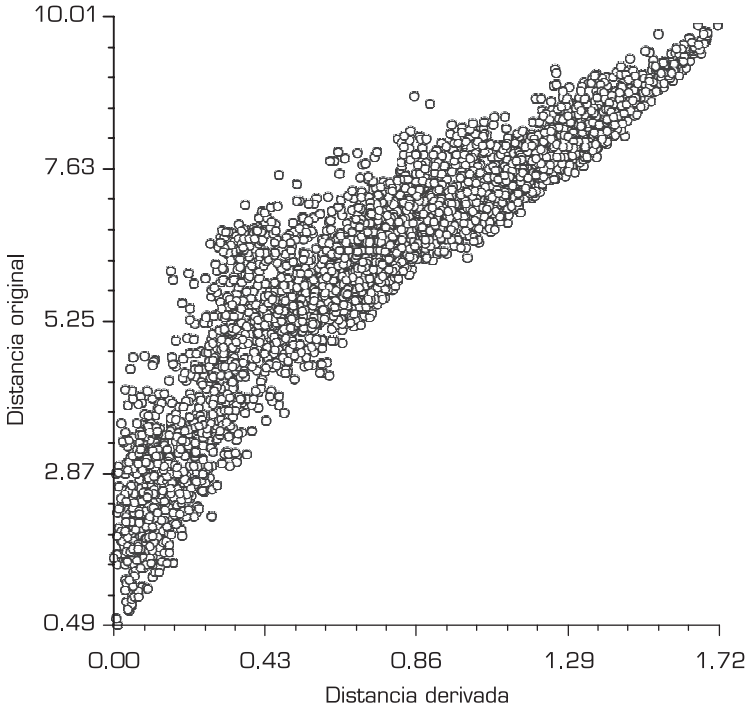


Figura 3. Distancia Euclidiana en espacio derivado bidimensional vs. igual distancia en espacio p-dimensional para componentes principales derivados a partir de una matriz de varianza–covarianza de datos estandarizados por desvío estándar sólo para variables cuantitativas. *r* = 0.92172.

De nuevo las correcciones por autovalores negativos no mostraron cambios importantes con respecto a la ordenación sin corrección (tabla 7).

Coeficiente de similaridad de Gower.— En la tabla 8 los valores de correlación negativos se deben a que se está comparando una medida de similaridad con una de disimilaridad. El coeficiente de Gower generó un espacio similar al de las variables estandarizadas.

Los diagramas de Shepard no se incluyen porque son similares a los presentados con anterioridad.

Al igual que en la disimilaridad Canberra,

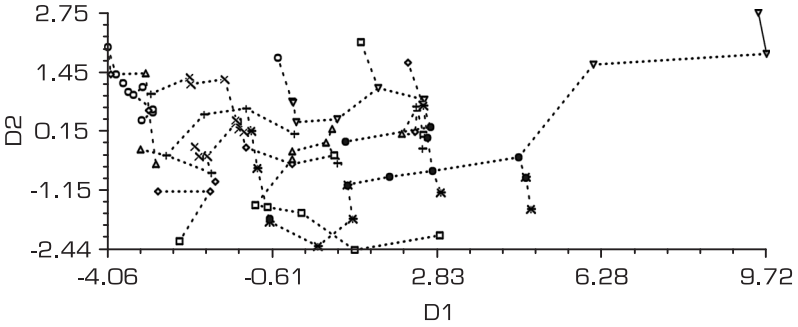


Figura 4. Ordenación de componentes principales obtenidos a partir de una matriz de varianza–covarianza de datos no estandarizados.

	manhde	*manhdec*	*manhrg*	*manhrgc*
manhde	1			
manhde_lg	0.99980			
mandeh_cl	0.98490			
manhdec		1		
manhdec_lg		0.99883		
manhdec_cl		0.99687		
manhrg			1	
manhrg_lg			0.99760	
manrgh_cl			0.97905	
manhrgc				1
manhrgc_lg				0.99962
manrghc_cl				0.99569

Tabla 5. Correlaciones entre espacios euclidianos bidimensionales obtenidos por coordenadas principales de una matriz de disimilaridad Manhattan sobre datos estandarizados por desvío estándar y rango, con y sin correcciones por autovalores negativos. Referencias en tabla 5. (*lg* = corrección de Lingoes; *cl* = corrección de Calliez).

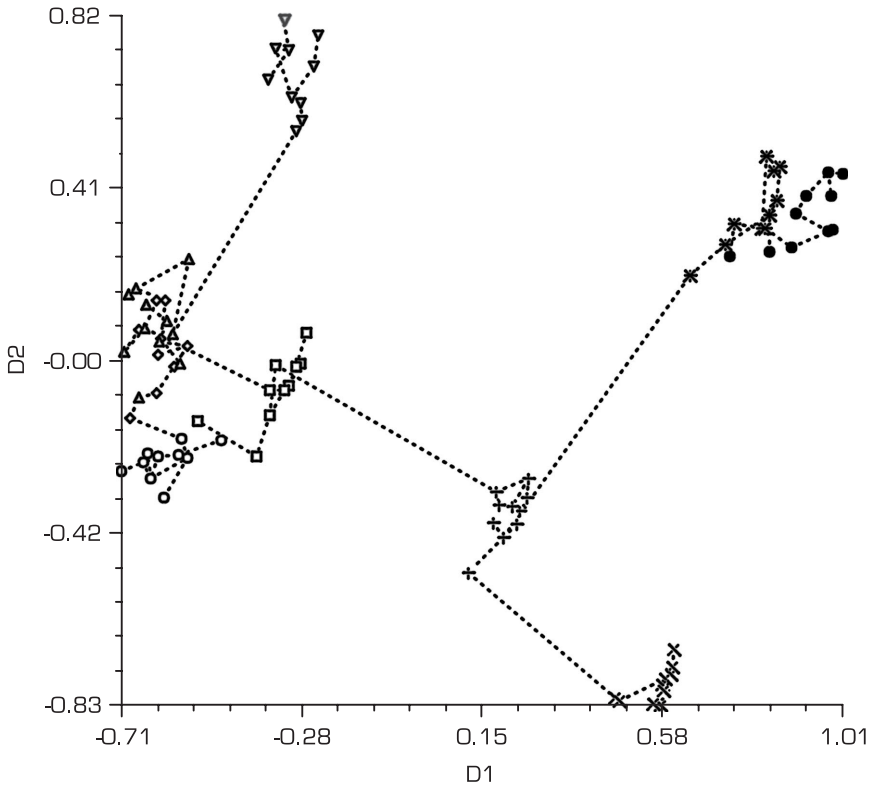


Figura 5. Ordenación de componentes principales obtenidos a partir de una matriz de varianza–covarianza de datos centrados y estandarizados por desvío estándar (sólo variables cuantitativas).

Codificación	Estandarización	Autovalores	% Varianza explicada	Aut. negativos	Correlación bray-euc en 2 y 3 dimensiones	Grupos identificados
rcbray	ninguna	24.7733 14.8198 9.4153	24.7783 39.5932 49.0084	no	0.85353 0.90017	2

Tabla 6. Autovalores, porcentaje de varianza explicada, presencia de autovalores negativos, correlación entre distancia original y derivada en 2 y 3 dimensiones, correlación entre distancia original y distancia Euclidiana derivada en 2 y 3 dimensiones y correlación entre distancia Euclidiana original y distancia Euclidiana derivada en 2 y 3 dimensiones, y número de grupos identificados en la representación bidimensional. Palabras en cursiva corresponden a la codificación empleada en otras tablas y figuras. (rcbray = √ Bray Curtis).

la adición de una tercera dimensión mejoró la representación gráfica (figuras 11 y 12).

Comparación de resultados de componentes principales y coordenadas principales

El dendrograma (figura 13) muestra claramente la poca concordancia entre configuraciones con variables estandarizadas y no estandarizadas, como se observa en el grupo de la parte superior (desde *varne* hasta *rcbray*). En este grupo está incluida la disimilaridad Bray Curtis, que pese a ser auto normalizada, generó resultados similares a los de opciones sin estandarizar. Todas las opciones en este grupo fracasaron en la separación de los grupos en estudio.

El grupo ubicado en el sector medio (des-

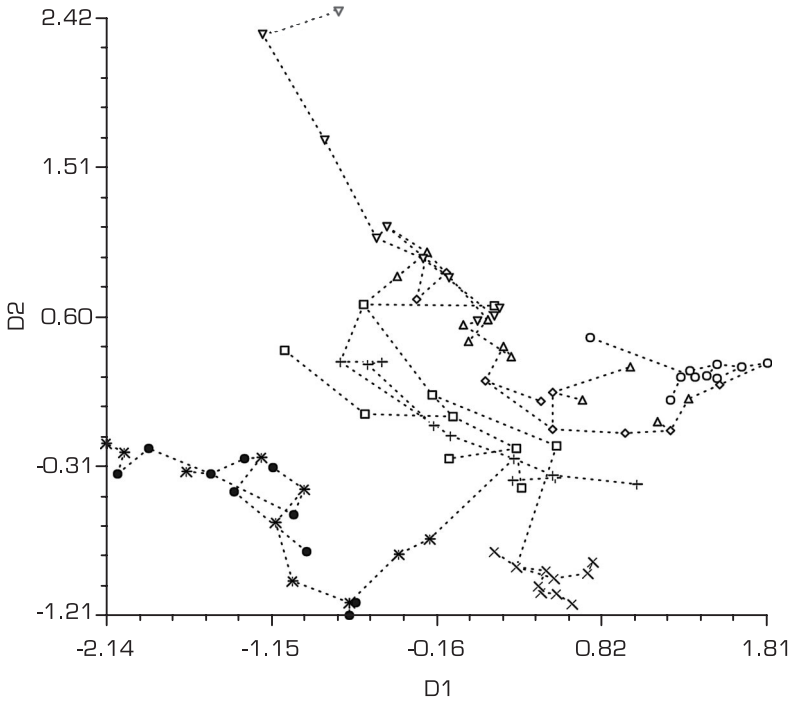


Figura 6. Ordenación de coordenadas principales a partir de una matriz de disimilaridad Manhattan de datos no estandarizados.

Codificación	Estanda-rización	Autovalores	% Varianza explicada	Aut. negativos (autovalor más pequeño)	Correlación can-euc en 2 y 3 dimensiones	Grupos identificados
can	ninguna	2.7156 1.1163 0..5990	55.2353 77.9412 90.1248	si (-2.1004)	0.93154 0.98410	8
canlg	ninguna correc. Lingoes	2.8160 1.2161 0.6994	20.3312 29.1157 34.1651	no	0.94475 0.97341	8
cancl	ninguna correc. Calliez	5.6082 2.5286 1.5509	33.0526 47.9553 57.0955	no	0.92897 0.98416	8

Tabla 7. Autovalores, porcentaje de varianza explicada, presencia de autovalores negativos, correlación entre distancia original y derivada en 2 y 3 dimensiones, correlación entre distancia original y distancia Euclidiana derivada en 2 y 3 dimensiones y correlación entre distancia Euclidiana original y distancia Euclidiana derivada en 2 y 3 dimensiones, y número de grupos identificados en la representación bidimensional. Palabras en cursiva corresponden a la codificación empleada en otras tablas y figuras.

de *varde* hasta *manhdec_cl*) muestra todas las opciones con resultados similares y que permitieron la identificación de los 8 grupos.

La disimilaridad Manhattan produjo resultados similares a los de distancia Euclídea.

La autonormalización incluida en la di-

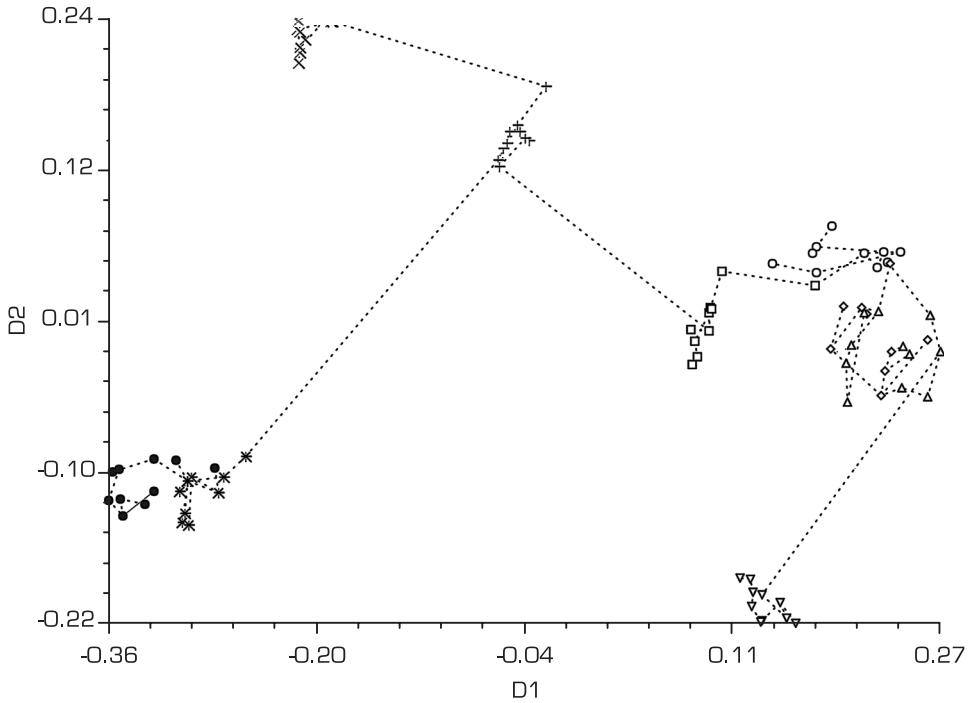


Figura 7. Ordenación de coordenadas principales a partir de una matriz de disimilaridad Manhattan de datos estandarizados por rango (todas las variables).

Codificación	Estanda-rización	Autovalores	% Varianza explicada	Autovalores negativos	Correlación gower-euc en 2 y 3 dimensiones	Grupos identificados
gower	ninguna	13.4788 6.1171 4.2227	34.5496 50.2294 61.0532	no	–0.92484 –0.95880	8

Tabla 8. Autovalores, porcentaje de varianza explicada, presencia de autovalores negativos, correlación entre distancia original y distancia Euclidiana derivada en 2 y 3 dimensiones y correlación entre distancia Euclidiana original y distancia Euclidiana derivada en 2 y 3 dimensiones, y número de grupos identificados en la representación bidimensional. Palabras en cursiva corresponden a la codificación empleada en otras tablas y figuras.

similaridad Canberra resultó equivalente a las estandarizaciones realizadas en otros coeficientes.

El coeficiente de similaridad de Gower, diseñado específicamente para variables mixtas, produjo ordenaciones similares a las de otros coeficientes con estandarización**.**

Se observa que las dos técnicas de corrección por autovalores negativos generaron ordenaciones casi idénticas entre sí y a la originada por la opción sin corrección. Pero se debe destacar que en los casos estudiados los valores propios negativos resultaron muy próximos a 0 o con valores absolutos muy pequeños en relación a los tres primeros con valores positivos

En el grupo inferior (desde *vardec* hasta *eucrgc*) se encuentran las opciones con estandarización solo de variables cuantitativas. Al aparecer juntas en el dendrograma indica

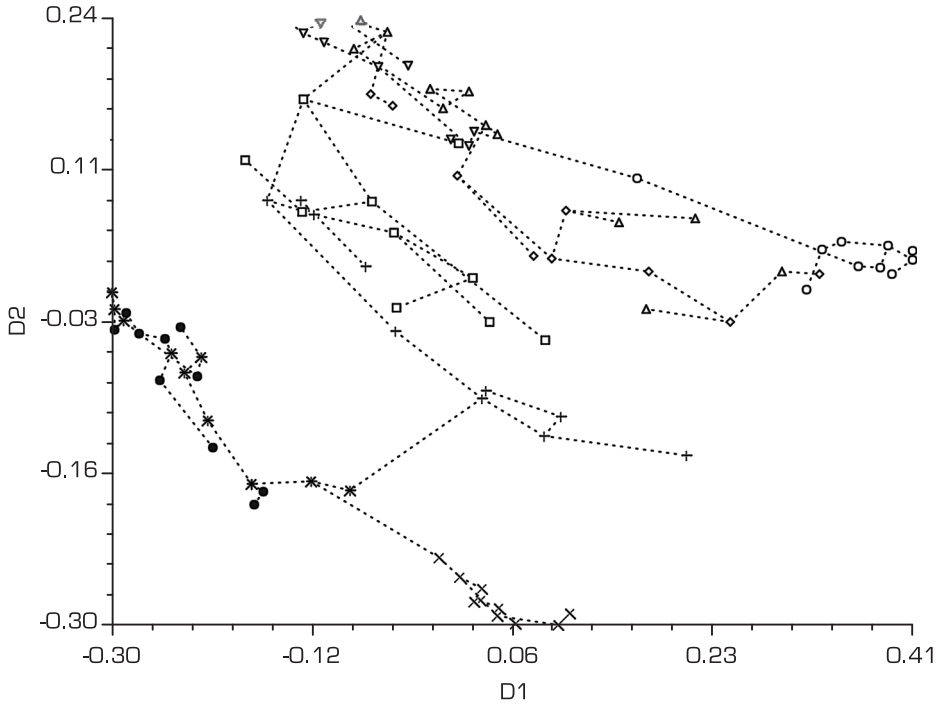


Figura 8. Ordenación de coordenadas principales a partir de una matriz de disimilaridad Bray Curtis.

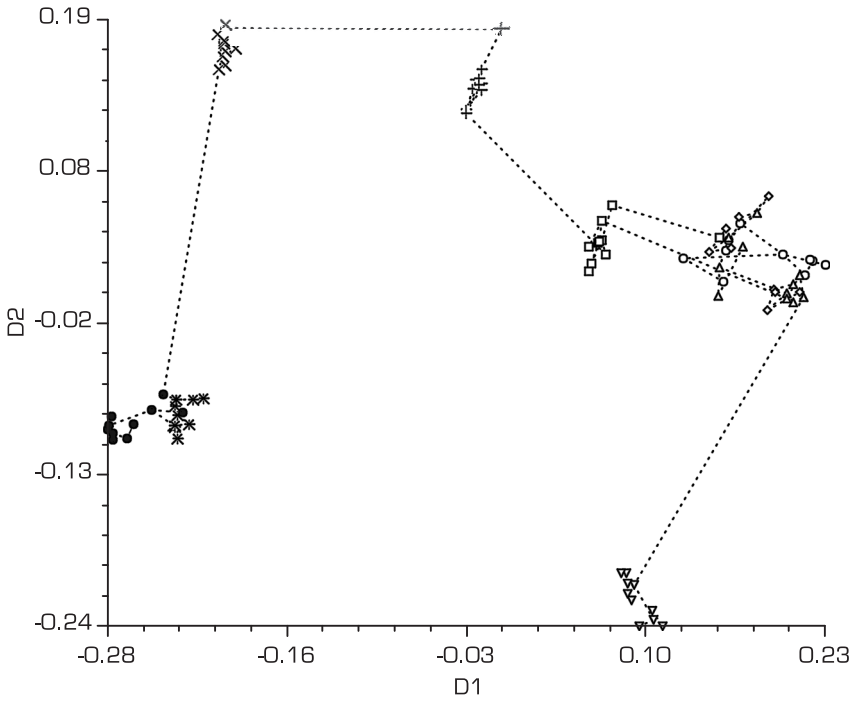


Figura 9. Ordenación de coordenadas principales, en 2 dimensiones, a partir de una matriz de disimilaridad Canberra.

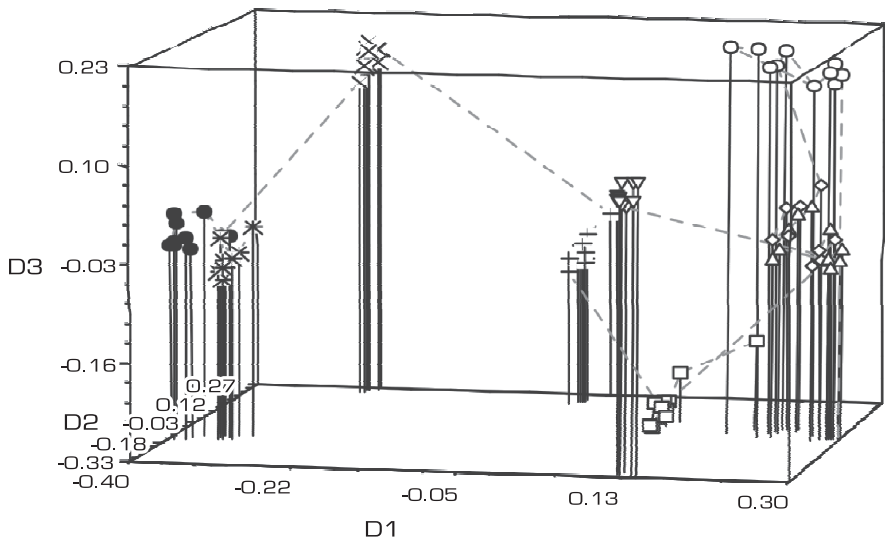


Figura 10. Ordenación de coordenadas principales, en 3 dimensiones, a partir de una matriz de disimilaridad Canberra.

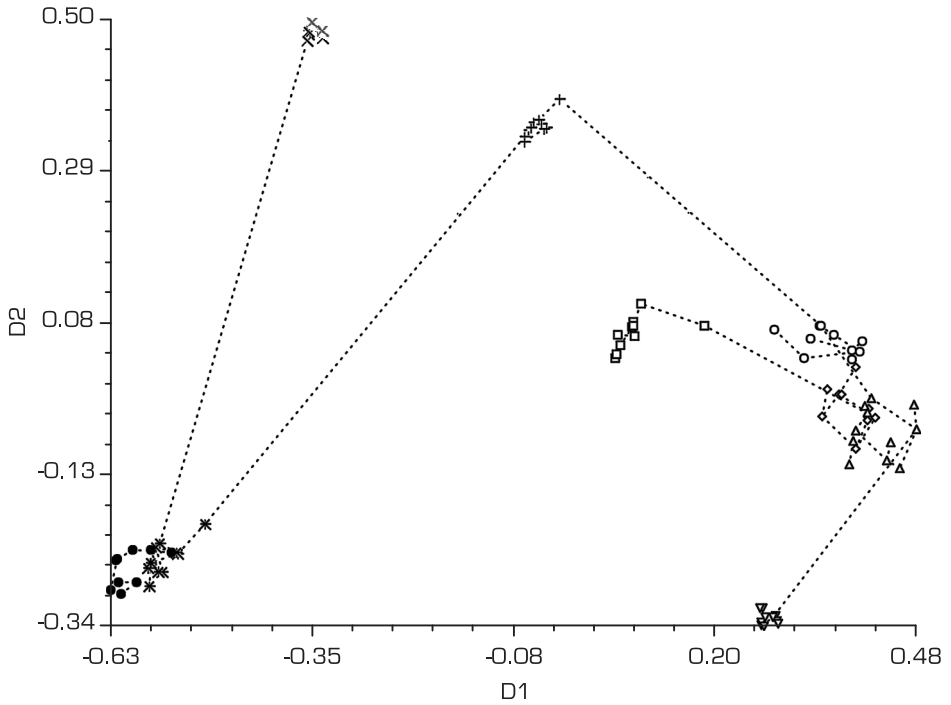


Figura 11. Ordenación de coordenadas principales, en 2 dimensiones, a partir de una matriz de coeficiente de Gower.

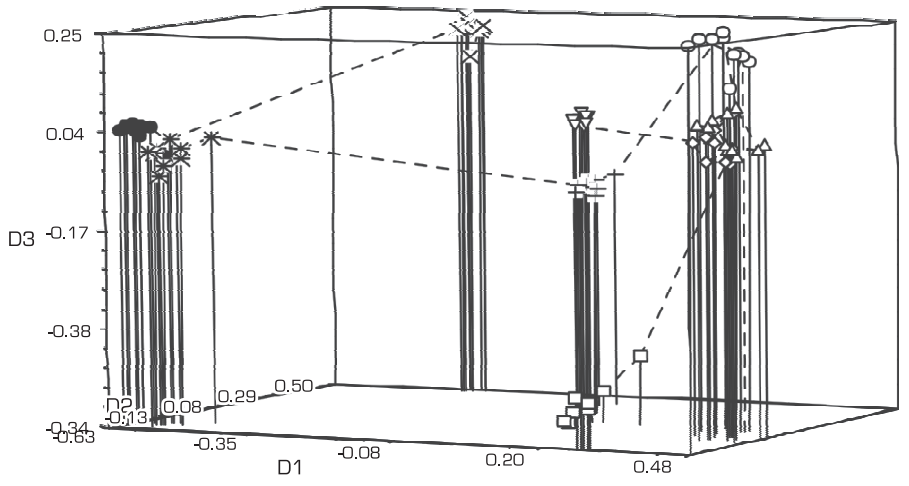


Figura 12. Ordenación de coordenadas principales, en 3 dimensiones, a partir de una matriz de coeficiente de Gower.

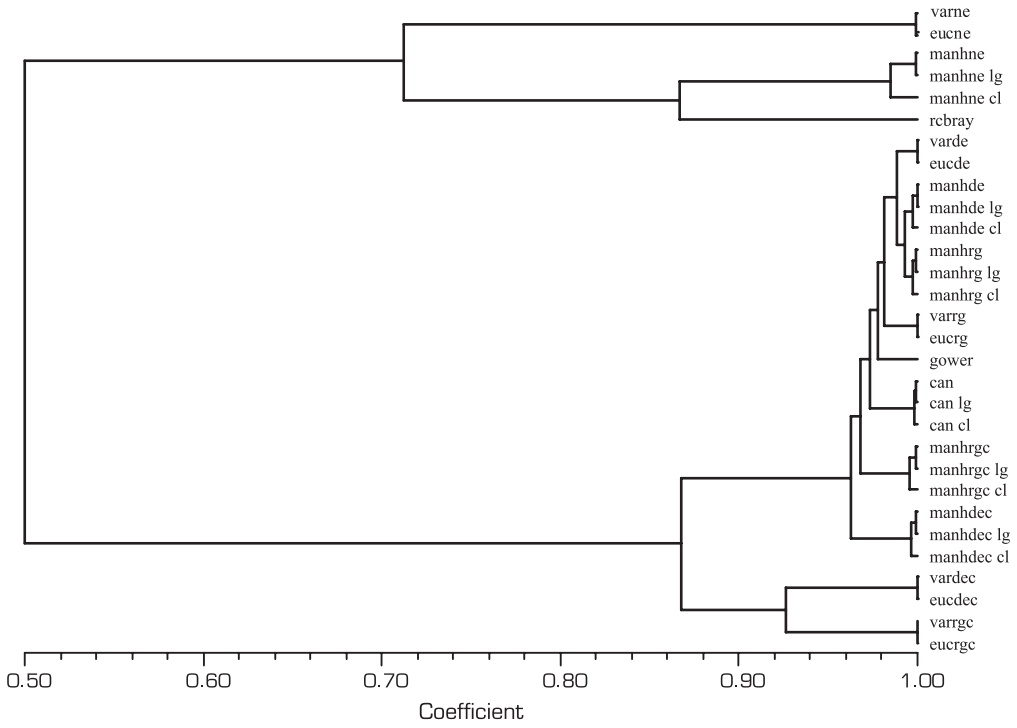


Figura 13. Análisis de agrupamientos por ligamiento promedio realizado para la matriz de correlaciones entre los espacios euclidianos derivados bidimensionales para todos los coeficientes, estandarizaciones y las dos técnicas en estudio. Referencias en tablas 1, 3, 4, 6, 7 y 8.

que fueron similares, sin embargo como se vio anteriormente la estandarización por rango solo de variables cuantitativas condujo a una sobreidentificación de grupos.

También se observa la dualidad de componentes principales y coordenadas principales con distancia Euclidiana.

DISCUSIÓN

La mayoría de la bibliografía referida a análisis estadístico multivariado presenta las dos técnicas mencionadas para situaciones de aplicación en las que se dispone de variables de un solo tipo (Peña, 2002; Hair *et al.*, 1999; Gnanadesikan, 1997; Jobson, 1992; Johnson y Wichern, 1992; Jolliffe, 1986; Anderson, 1984; Dillon y Goldstein, 1984; Seber, 1984; Karson, 1982; Mardia *et al.*, 1979; Morrison, 1967). Se asume que la matriz básica de datos está constituida por variables cuantitativas, en el caso de componentes principales, o por variables binarias, categóricas multiestado o cuantitativas en el caso de coordenadas principales.

Numerosos trabajos de aplicación de estas técnicas pueden encontrarse en la literatura científica (Hartmann, 1988; Pimentel, 1981; Thorpe, 1980; Crisci y López Armengol, 1983; Sneath y Sokal, 1973; Moss, 1968). En la mayoría de ellos se aplican las técnicas multivariadas sin tener en cuenta la naturaleza mixta de los datos. La interpretación de los ejes, en este caso, resulta poco clara ya que las correlaciones entre variables numéricas y categóricas codificadas o entre los códigos numéricos de las variables categóricas carecen de sentido.

Con respecto a la rotación en 180º de las ordenaciones de componentes principales y coordenadas principales, Legendre y Legendre (1998) dicen lo siguiente; “las elecciones

de signos de los autovectores son arbitrarias durante la ejecución de los algoritmos de computadora". Por lo tanto, esta rotación no es una característica propia de la técnica sino del algoritmo incluido en el paquete con el que se está trabajando En el caso de NTSys la multiplicación de los dos primeros autovectores de coordenadas principales por (-1) generará ordenaciones con posiciones similares a las de componentes principales.

En este trabajo se pretendió comparar las ordenaciones obtenidas usando el coeficiente de Gower, apropiado para variables mixtas, y evaluar la robustez de las técnicas de componentes principales y coordenadas principales cuando se aplican coeficientes que no son apropiados para matrices con variables de distintos tipos.

El uso de los diagramas de Shepard (McCune y Grace, 2002; Gnanadesikan, 1997; Everitt y Dunn, 1991; Legendre, 1998; Kruskal, 1964) resultó muy útil para evaluar las relaciones entre espacios originales y derivados, ya que permitieron observar el tipo, dirección y magnitud de las distorsiones obtenidas. Estas relaciones son lineales cuando se trabaja con variables cuantitativas solamente (Arce y Santillán, 2002; Arce, 2003) Es posible que la inclusión de variables no numéricas codificadas haya producido las relaciones no lineales observadas entre los espacios original y derivados. Por este motivo cabe esperar mayores distorsiones al trabajar con datos mixtos que cuando sólo se utilizan variables cuantitativas. En los trabajos mencionados con anterioridad no se hace mención a este efecto de no linealidad. La inclusión en NTSys de estos gráficos resulta muy apropiada, como así también el cálculo de correlación entre matrices.

La superposición de árboles de recorrido mínimo a las ordenaciones (Gower y Ross, 1969) resultó muy útil para detectar distorsiones en la representación de los puntos en el espacio de dimensión reducida. Bramardi (2000) también analiza en su trabajo la utilidad de esta herramienta de evaluación de ordenaciones en este sentido.

El efecto de la autovalores negativos ha sido tratado por Gower y Legendre (1986) y resuelto computacionalmente por Legendre y Anderson (1998). No se encontró un efecto importante de las correcciones porque el valor absoluto de los autovalores negativos resultó muy pequeño en relación a los primeros tres autovalores positivos. Pero se trabajó empíricamente con un solo conjunto de datos. Esto no necesariamente será siempre así por lo que debería verificarse el efecto de las correcciones siempre que se obtienen autovalores negativos

Se intentó presentar una metodología que incluyera varias técnicas para juzgar ordenaciones. Se observó en los resultados que el uso de la más usada (porcentaje de varianza explicada) no es suficiente por sí misma y que incluso puede conducir a le elección de una ordenación inapropiada.

El uso del coeficiente de Gower debiera extenderse. Se ha visto empíricamente en este trabajo, que su uso en la matriz de datos de *Echinochloa* resultó satisfactorio. Pimentel (1981) y Bramardi (2000) mencionan también su efectividad en este tipo de matrices. Este coeficiente no se encuentra como una alternativa disponible en el software de uso común. En NTSys no está incluido pero se lo puede calcular de manera sencilla usando las operaciones de matrices incluidas en el módulo TRANSF. Recientemente ha sido incluido en el paquete "cluster" de R (R Development Core Team, 2009), mediante el procedimiento DAISY (Kaufman y Rousseeuw, 1990).

También se observó que el uso de componentes principales y coordenadas principales, al ser empleados sin considerar la naturaleza mixta de las variables, presentó robustez, es decir, ordenaciones apropiadas. Pero en este caso se trabajó con una sola matriz de datos por lo que los resultados obtenidos no deberían generalizarse para toda matriz de datos mixtos.

CONCLUSIONES

En la matriz de datos estudiada se conocía la estructura de agrupamientos con anterioridad a su análisis estadístico, es decir, el modelo taxonómico era conocido. La utili-

zación de ambas técnicas, bajo diferentes condiciones de aplicación, permitió verificar si el modelo taxonómico conocido podía ser reproducido o no en espacios de dimensión reducida. El número de grupos identificados en algunos casos fue nulo, en otros fue el correcto y en unos pocos resultó menor que el número original. Esto indica que algunas condiciones de aplicación pueden llevar al investigador a obtener conclusiones taxonómicas erróneas.

Excepto en los casos especiales de dualidad mencionados, el análisis de componentes principales no tiene una relación directa con el de coordenadas principales.

El análisis de las dos técnicas mostró que ambas son robustas, ya que en la mayoría de los casos produjeron resultados muy similares y concordantes con el modelo taxonómico conocido.

La condición fundamental para obtener ordenaciones que reprodujeron el modelo taxonómico original fue la estandarización de los datos y no la selección de coeficientes determinados.

No debiera utilizarse componentes principales con datos de naturaleza mixta debido a la no linealidad que se genera entre los espacios original y derivado, que produce distorsiones en la representación bidimensional.

Se observó que coordenadas principales fue capaz de forzar un modelo euclidiano a una matriz cualquiera de disimilaridades o similaridades, lo que indicaría que se puede aplicar a cualquier tipo de datos usando medidas de di/similaridad apropiadas. El coeficiente de Gower surge como una alternativa interesante.

BIBLIOGRAFÍA

Anderson, T. 1984. An introduction to multivariate statistical analysis. 3rd edition. New York, Wiley, 752 pp.

Arce, O y M. Santillán. 2002. A comparative study of two ordination techniques based on simulated multivariate normal data. Biocell, 26 (1): 159.

Arce, O. 2003. Componentes principales y coordenadas principales: estudio comparativo con aplicaciones a la taxonomía numérica. Tesis de Maestría en Estadística Aplicada. Tucumán, Facultad de Ciencias Económicas, Universidad Nacional de Tucumán, 149 pp.

Bramardi, S. 2000. Estrategias para el análisis de datos en la caracterización de recursos fitogenéticos. Tesis doctoral. Valencia, Universidad Politécnica, 390 pp.

Cailliez, F. 1983. The analytical solution of the additive constant problem. Psychometrika, 48: 305-308.

Crisci, J. y M. F. López Armengol. 1983. Introducción a la teoría y práctica de la taxonomía numérica. Monografía Nº 26. Washington, Organización de Estados Americanos, 132 pp.

De La Sota, E. 1982. La taxonomía y la revolución en las ciencias biológicas. Monografía Nº 3. Washington, Organización de Estados Americanos, 86 pp.

De Marco, N. 2006. *Echinochloa*. En A. Molina y Z. R. de Agrasar (editores). Colección científica del INTA 23. INTA, Buenos Aires, pp. 493-510.

De Marco, N. 2002. Estudio sistemático y fitogeográfico de las especies del género *Echinochloa (Poaceaea: Panicoideae: Paniceae*) para la Argentina. Tesis doctoral. Tucumán, Facultad de Agronomía y Zootecnia, Universidad Nacional de Tucumán, 178 pp.

Dillon, W y M. Goldstein. 1984. Multivariate analysis. Methods and applications. New York, Wiley, 287 pp.

Everitt, B. y G. Dunn. 1991. Applied multivariate data analysis. London, Arnold, 304 pp.

Gnanadesikan, R. 1997. Methods for statistical data analysis of multivariate observations. 2 nd edition. New York, Wiley, 353 pp.

Gower, J. 1966. Some distance properties of latent roots and vector methods used in multivariate analysis. Biometrika, 53: 325-338.

Gower, J. 1971. A general coefficient of similarity and some of its properties. Biometrics, 27: 857-74.

Gower, J. 1985. Measures of similarity, dissimilarity and distance. Encyclopedia of statistical sciences, 5: 397-405.

Gower, J. y N. Digby. 1981. Expressing complex relationships in two dimensions. En V. Barnett (editor). Interpreting multivariate data . Wiley, UK, pp 83-118.

Gower, J. y P. Legendre. 1986. Metric and euclidean properties of dissimilarity coefficients. Journal of Classification, 3: 5-48.

Gower, J. y G. Ross. 1969. Minimum spanning trees and single linkage cluster analysis. Applied statistics, 18: 54-64.

Hair, J., R. Anderson, R. Tatham y W. Black. 1999. Análisis multivariante. 5ª edición. Madrid, Prentice Hall Iberia, 799 pp.

Hartman, S. 1988. Evaluation of some alternative procedures used in numerical systematic. Systematic zoology, 37 (1): 1-18.

Jobson, J. 1992. Applied multivariate data analysis. Volume II: Categorical and multivariate methods. New York, Springer-Verlag, 768 pp.

Johnson, R y W. Wichern.1992. Applied multivariate statistical analysis. 3rd edition. New Jersey, Prentice Hall, 642 pp.

Jolliffe, I. 1986. Principal component analysis. Heidelberg, Springer-Verlag, 271 pp.

Karson, M. 1982. Multivariate statistical methods. An introduction. Iowa, Iowa State University Press, 307 pp.

Kaufman, L y P. Rousseeuw. 1990. Finding groups in data. An introduction to cluster analysis. New York, Wiley-Interscience, 342 pp.

Kruskal, J. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 29 (1): 1-27.

Legendre, P. y J. Anderson. 1998. Program DistPCoA. User's manual. Montreal: Université de Montreal, Département de Sciences Biologiques. http://www.bio.umontreal.ca/casgrain/en/telecharger/index.html#DistPCoA (consultado el 21 de octubre de 2009).

Legendre, P. y L. Legendre. 1998. Numerical ecology. 2nd Edition. Amsterdam, Elsevier, 853 pp.

Lingoes, J.C. 1971. Some boundary conditions for a monotone analysis of symmetric matrices. Psychometrika, 36: 195-203.

McCune, B y Grace, J. 2002. Analysis of ecological communities. Oregon, MJM, 300 pp.

Mardia, K., J. Kent y J. Bibby. 1979. Multivariate analysis. London: Academic Press, 521 pp.

Morrison, D. 1967. Multivariate statistical methods. New York, McGraw Hill, 409 pp.

Moss, W. 1968. Experiments with various techniques of numerical taxonomy. Systematic Zoology, 17 (1): 31-47.

Peña, D. 2002. Análisis de datos multivariantes. Madrid, MacGraw Hill Interamericana de España, 539 pp.

PimenteL, R. 1981. A comparative study of data and ordination techniques based on a hybrid swarm of sand verbenas *(Abronia* Juss.). Systematic zoology, 30 (3): 250-267.

R Development Core Team. 2009. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org (consultado el 21 de octubre de 2009).

Rohlf, F. 1972. And empirical comparison of three ordination techniques in numerical taxonomy. Systematic zoology, 21 (3): 271-280.

Rohlf, F. 1990. Numerical taxonomy system of multivariate statistical programs. Version 1.8. New York: State University at Stony Brook.

Rohlf, F., 2009. Numerical taxonomy system of multivariate statistical programs. Getting started guide. Version 2.2. New York: State University at Stony Brook, 43 pp.

Seber, G. 1984. Multivariate observations. New York, Wiley and Sons, 686 pp.

Sneath, P. y R. Sokal. 1973. Numerical taxonomy. The principles and practice of numerical classification. San Francisco, Freeman, 573 pp.

Thorpe, R. 1980. A comparative study of ordination techniques in numerical taxonomy in relation to racial variation in the ringed snake *Natrix natrix* (L.). Biological journal of the Linnean Society, 13: 7-40.

Zuloaga, F. O., E. G. Nicora, Z. E. Rúgolo de Agrasar, O. Morrone, J. Pensiero y A. M. Cialdella. 1994. Catálogo de la Familia *Poaceae* en la República Argentina. Monographs in Systematic Botany. Missouri Botanical Garden, 47: 1-178.