DARWINIANA
NUEVA SERIE
www.ojs.darwin.edu.ar
ISSN 0011-6793 impresa - ISSN 1850-1699 en línea

DARWINIANA, nueva serie 11(1): 272-294. 2023
Versión de registro, efectivamente publicada el 30 de junio de 2023
DOI: 10.14522/darwiniana.2023.111.1086

ARTÍCULO INVITADO

ABORDAJE MULTIVARIADO EN ESTUDIOS BOTÁNICOS Y ECOLÓGICOS

Susana B. Perelman[1,2] & Laura E. Puhl[1]

[1] *Departamento de Métodos Cuantitativos y Sistemas de Información, Facultad de Agronomía, Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina; perelman@agro.uba.ar* (autora corresponsal).
[2] *IFEVA, Facultad de Agronomía, Universidad de Buenos Aires, CONICET, Ciudad Autónoma de Buenos Aires, Argentina.*

Abstract. Perelman, S. B. & L. E. Puhl. 2023. Multivariate approach in botany and ecology. *Darwiniana*, nueva serie 11(1): 272-294.

Interesting botanical and ecological studies used the multivariate approach to describe underlying patterns in large data sets and to answer questions about the structure of the studied systems at various scales. This work aims to encourage the correct use of the multivariate approach and offers guidelines for the appropriate choice of the analysis techniques according to the objectives of the study and the characteristics of the data. Some appropriate applications of principal components, multidimensional scaling, correspondence analysis, discriminant, and cluster analysis in articles published in scientific journals of these disciplines are showed. We used reduced versions of the data matrices from some of these papers to present the methods of analysis in a simple way. The focus is placed on the correct interpretation of the results and the biological questions that can be answered through multivariate analysis.

Keywords. Cluster analysis; correspondence analysis; discriminant analysis; multidimensional scaling; multivariate data; principal components.

Resumen. Perelman, S. B. & L. E. Puhl. 2023. Abordaje multivariado en estudios botánicos y ecológicos. *Darwiniana*, nueva serie 11(1): 272-294.

Interesantes estudios de botánica y ecología utilizaron un enfoque multivariado para describir patrones subyacentes en grandes conjuntos de datos. Mediante este abordaje lograron responder interrogantes acerca de la estructura de sistemas biológicos en diferentes escalas. Este trabajo se propone incentivar el uso correcto del enfoque multivariado y ofrecer pautas para la apropiada elección de las técnicas de análisis según los objetivos del estudio y las características de los datos. Contiene ejemplos de artículos publicados en revistas científicas con apropiadas aplicaciones de Componentes Principales, Escalamiento Multidimensional, Análisis de Correspondencia, Análisis Discriminante y Aglomeración Jerárquica. Utilizamos versiones reducidas de las matrices de datos de algunos de estos trabajos publicados para presentar de manera sencilla los métodos de análisis. El artículo pone en consideración las preguntas biológicas que los análisis multivariados permiten responder y la correcta interpretación de los resultados.

Palabras clave. Análisis de correspondencia; análisis discriminante; clasificación jerárquica; componentes principales; datos multivariados; escalamiento multidimensional.

INTRODUCCIÓN

Los botánicos y los ecólogos frecuentemente abordan temas de naturaleza multidimensional que requieren cuantificar simultáneamente varias características en los mismos objetos de estudio. Por ejemplo, pueden medir diferentes atributos en una planta, o varias características en una flor o ciertos descriptores del ambiente en los sitios de recolección. En general, las variables medidas en una misma planta, misma flor o mismo ambiente no son independientes entre sí, y esto

Original recibido el 28 de septiembre de 2022, aceptado el 8 de junio de 2023
Editor Asociado: Fernando Biganzoli

nos obliga a recurrir a metodologías especiales para analizarlas en conjunto. Además, el estudio de la variación conjunta entre variables aporta mucha riqueza a nuestro conocimiento de los sistemas naturales. Existen técnicas estadísticas de análisis multivariado que son poderosas herramientas para descubrir, resumir e interpretar la estructura subyacente en conjuntos complejos de datos multidimensionales. Sin embargo, la amplia paleta de opciones disponibles a la hora de elegir el método de análisis multivariado puede resultar bastante confusa y la interpretación de los resultados puede no ser muy obvia.

En este trabajo presentamos algunas pautas para la apropiada elección del método a utilizar según los objetivos del análisis y las características de los datos. Mostraremos los resultados de algunos trabajos con el foco puesto en las preguntas biológicas que intentan responder y en la correcta interpretación de los resultados. Con ese fin utilizaremos también ejemplos didácticos adaptados de conjuntos de datos mucho más extensos (Uribbe et al., 2020, Puhl et al., 2014, Yansen & Biganzoli, 2022) para ilustrar una explicación sencilla de los fundamentos de los métodos. Al mismo tiempo comentaremos resultados de aplicaciones interesantes de estos métodos en estos y otros trabajos publicados. Para descripciones más completas ya sea en la amplitud de metodologías abarcadas y/o en la profundidad de exploración de éstas, existen muy buenos libros (Legendre & Legendre, 2003; Husson & Pages, 2017; Palacio et al., 2020) y también cursos de posgrado dedicados íntegramente a estos procedimientos de análisis. A partir de aquí, a la unidad de muestreo (por ejemplo, planta, sitio, flor o parcela) sobre la que se miden las distintas variables, la nombramos en general como objeto, caso o individuo.

Pautas para seleccionar el método de análisis

El investigador decide qué es lo que desea indagar acerca del sistema bajo estudio y qué está buscando como resultado, y conoce qué tipo de datos tiene entre manos. Es capaz de responder preguntas como estas: ¿quiero mostrar la variación del conjunto total y detectar posibles patrones o quiero diferenciar entre grupos de objetos determinados *a priori?* Es decir, decide entre buscar el orden que muestra más variación o el que produce mayor discriminación entre los grupos. Por ejemplo, ordenar plantas de una misma especie y mostrar la máxima variabilidad entre ellas o, por el contrario, ordenar ejemplares recolectados en diferentes localidades (grupos) y representar la máxima separación entre grupos. Tal vez no quiere ordenar los objetos bajo estudio sino agruparlos para generar una tipología. En esa situación, ¿busca producir una jerarquía que informa acerca de la cercanía entre grupos a diferentes niveles o prefiere generar solo grupos internamente homogéneos sin establecer el grado de proximidad entre ellos? Por otra parte, puede conocer el tipo de datos que va a analizar: ¿son variables cuantitativas, o de presencia/ausencia o de categorías? ¿Todas son variables de respuesta o algunas son predictoras? Tanto los objetivos que persigue el investigador como las características de los datos definen cuál es el mejor método multivariado para cada circunstancia.

En este trabajo analizaremos dos grandes grupos de métodos de análisis multivariados, los métodos de ordenamiento y los métodos de clasificación (Caja 1). Existen otros métodos de análisis de datos con múltiples variables que no consideraremos, como los análisis específicamente diseñados para proponer relaciones de parentesco entre especies (análisis filogenéticos), el análisis de varianza multivariado (MANOVA) o la Regresión Múltiple. Los métodos de ordenamiento permiten representar en pocas dimensiones a los objetos bajo estudio. Por ejemplo, ordenarlos en un plano de dos dimensiones, de manera que los objetos más semejantes para los atributos observados ocupen posiciones más cercanas en el plano. Estas representaciones facilitan la interpretación de grandes matrices de datos y minimizan al mismo tiempo la pérdida de información que conlleva toda síntesis. También facilitan la interpretación porque proyectan al conjunto de las observaciones multivariadas en un espacio de fácil visualización (de sólo dos o tres dimensiones). La diferencia más importante entre los métodos de ordenamiento radica en el criterio de mejor solución que aplican para representar los objetos en pocas dimensiones. A través de los ejemplos en las próximas secciones veremos cómo estos criterios están asociados a diferentes preguntas del investigador.

Al mismo tiempo, estas técnicas difieren en el tipo de datos que pueden procesar (Caja 1): para Componentes Principales son variables cuantitativas, para Análisis de Correspondencia son tablas de contingencia, como frecuencias, porcentajes o medidas de presencia y ausencia de caracteres (con o sin pubescencia, por ejemplo). En cambio, para Escalamiento Multidimensional no hay restricción en el tipo de variables ya que este método permite elegir la medida de distancia que se va a utilizar en el ordenamiento. Esto lo hace muy versátil porque existe una gran variedad de medidas de distancia que admiten diferentes tipos de datos además de permitir ponderar diferentes aspectos de la semejanza/diferencia entre objetos (Caja 2). Para el Análisis Discriminante la matriz de datos podría ser idéntica a la de Componentes Principales, pero debe comprender además una variable de tipo categórica que indica el grupo de pertenencia para cada objeto. En un ejemplo de taxonomía, las variables podrían ser medidas morfométricas (variables cuantitativas) que serían utilizadas para la caracterización de diferentes taxones (variable categórica) con la finalidad de generar una regla automática para la identificación taxonómica de nuevos ejemplares (Christodoulou et al., 2020).

Por otra parte, si el propósito del investigador es el agrupamiento de los objetos para obtener una tipología o clasificación, como búsqueda de discontinuidades, que no necesariamente son discontinuidades naturales, podremos elegir entre diferentes maneras de agrupar observaciones semejantes. Las principales técnicas incluyen la Aglomeración Jerárquica (también denominada Análisis de Clusters) y las de Aglomeración No Jerárquica (por ejemplo, k-medias). En las técnicas jerárquicas se optimizan criterios asociados a cada paso de la aglomeración y se obtienen, además de los grupos, las relaciones entre grupos semejantes que conforman la jerarquía. En cambio, en las técnicas no jerárquicas se optimiza la estructura del grupo individual, que es todo lo homogénea que sea posible, no se busca generar relaciones entre los grupos. Su valor estriba en que son capaces de producir clasificaciones óptimas como consecuencia de la proximidad interna de los grupos y de la separación entre ellos.

Los ejemplos de métodos nombrados hasta ahora son los más utilizados en botánica y ecología y son los que presentaremos en detalle a continuación. Otras técnicas, como Correspondencia Múltiple, Correlación Canónica, Correspondencia Canónica, Clasificaciones supervisadas y Métodos de análisis filogenéticos, pueden encontrarse en los libros citados más arriba.

Análisis de Componentes Principales
(conocido como PCA por su sigla en inglés)

En este análisis la reducción de dimensión se alcanza construyendo nuevas variables capaces de abarcar la máxima proporción posible de la estructura de correlación de las variables originales. La estructura de correlación refiere a la fuerza y el sentido de la variación conjunta de las variables entre los objetos estudiados. Por ejemplo, en un muestreo de suelos, las variables pH y contenido de sodio podrían aumentar en la misma dirección y en sentido opuesto al aumento del porcentaje de arena y esto conformaría un aspecto de la estructura de correlación de ese conjunto de observaciones. El objetivo del análisis es tomar las p variables originales X_1, X_2, ..., X_p y encontrar combinaciones de ellas para producir índices Z_1, Z_2, …, Z_p que entre sí no estén correlacionados. La falta de correlación es una propiedad muy útil porque significa que cada uno de estos índices mide aspectos o dimensiones diferentes de los datos. Sin embargo, una propiedad aún más interesante de estos índices es que entre ellos la variación total se reparte de la manera más inequitativa posible, y al mismo tiempo, se ordenan de manera que Z_1 expresa la proporción más grande posible de la variación, Z_2 expresa la mayor proporción posible de la variación restante y así sucesivamente. Estos índices Z_j se denominan componentes principales. Quien conduce un análisis de componentes principales pretende capturar una parte muy importante de la variación total de los datos en las primeras componentes. De esta manera, la varianza contenida en el resto de las componentes no principales resultará casi insignificante.

Para centrarnos en el procedimiento de PCA utilizaremos de ejemplo didáctico una pequeña muestra extraída de un archivo de datos mucho más extenso. Se trata de un proyecto de investigación del

Tabla 1. Primeras filas de la matriz de datos del ejemplo de *Vriesea procera* (Bromeliaceae) adaptado de Uribbe et al. (2020). En 4 poblaciones se presentan mediciones de 7 variables en 10 plantas. La tabla completa utilizada en este ejemplo puede descargarse del Material suplementario I, https://www.ojs.darwin.edu.ar/index.php/darwiniana/article/view/1086/1300

Población	Ancho de vaina (AV)	Ancho de hoja (AH)	Longitud total inflorescencia (LTI)	Ancho de los sépalos (AS)	Ancho de los pétalos (AP)	Longitud de pétalos (LP)	Longitud de pistilos (LPS)
BA2	8,4	4,3	180	2,6	0,6	3,5	4,1
BA2	6,9	3,7	108	2,9	0,5	3,3	4,4
BA2	8,2	4	118	2,5	0,5	3,4	3,9
BA2	8	4,4	142,8	2	0,6	3,4	3,4
BA2	9,1	4,8	139	2,6	0,5	3,5	4,4
…							

patrón geográfico de variación del complejo *Vriesea procera* (Bromeliaceae) en un gradiente latitudinal que abarca 14 poblaciones espontáneas en la Mata Atlántica, Brasil (Uribbe et al., 2020). La pequeña matriz de datos que analizaremos aquí comprende solo 40 de las 271 plantas que los investigadores estudiaron (Tabla 1; Material suplementario I, https://www.ojs.darwin.edu.ar/index.php/darwiniana/article/view/1086/1300). Para construir esta matriz, elegimos 4 de las localidades que abarcaba el estudio original y tomamos al azar 10 plantas en cada una. Además, solo consideramos 7 de las 36 variables que ellos midieron: ancho de la vaina (AV), ancho de la hoja (AH), longitud total de la inflorescencia (LTI), ancho de sépalos (AS), ancho de pétalos (AP), longitud de pétalos (LP) y longitud de pistilos (LPS).

Consideraremos las siguientes preguntas que se le podrían formular a estos datos:

1. ¿Cómo se relacionan estas medidas entre sí?
2. ¿Cuál es la importancia relativa de las variables observadas para describir la heterogeneidad entre las plantas coleccionadas?
3. ¿Se podrían ordenar las plantas coleccionadas de acuerdo con algún índice que resuma su semejanza en estos caracteres?
4. ¿Existen grupos homogéneos de plantas dentro del conjunto observado? ¿o será que las variaciones entre ellas son graduales y continuas?

Los promedios y varianzas de los 7 caracteres (variables) medidos en este conjunto de datos se presentan en la Tabla 2. En esta tabla se observa que las varianzas de los caracteres difieren en varios órdenes de magnitud (por ejemplo, entre longitud total de la inflorescencia LTI y ancho de pétalos AP). Frente a una gran diferencia de varianzas o también si las variables son medidas en diferentes escalas (por ejemplo, altura de planta en centímetros y biomasa de hojas en gramos) es necesario estandarizar los datos. De esta forma, los distintos caracteres resultarán comparables entre sí.

Las relaciones entre pares de variables se resumen en general en la matriz de varianzas-covarianzas, y para datos estandarizados en la matriz de correlaciones. Las covarianzas miden la variación conjunta respecto a las respectivas medias entre cada par de variables en la escala original de medición, en cambio las correlaciones expresan la covariación en una escala estandarizada. En este conjunto de datos la matriz de correlaciones muestra algunas correlaciones positivas muy altas (por ejemplo, entre AV, AH y LTI o entre LP y LPS; Tabla 3) y algunas correlaciones negativas medianamente altas (AV con LP y LPS; Tabla 3), entre otras.

Tabla 2. Promedios y varianzas de las medidas de los caracteres que se presentaron en la Tabla 1. **AV,** ancho de la vaina; **AH,** ancho de la hoja; **LTI,** longitud total de la inflorescencia; **AS,** ancho de sépalos; **AP,** ancho de pétalos; **LP,** longitud de pétalos; **LPS,** longitud de pistilos.

	AV	AH	LTI	AS	AP	LP	LPS
Promedio	8,57	4,76	147,92	2,66	0,52	3,68	4,19
Varianza	4,08	2,21	2410,36	0,120	0,004	0,36	0,72

Tabla 3. Matriz de correlaciones de Pearson entre las medidas de los caracteres de la Tabla 1. **AV,** ancho de la vaina; **AH,** ancho de la hoja; **LTI,** longitud total de la inflorescencia; **AS,** ancho de sépalos; **AP,** ancho de pétalos; **LP,** longitud de pétalos; **LPS,** longitud de pistilos.

	AV	AH	LTI	AS	AP	LP	LPS
AV	1						
AH	0,90	1					
LTI	0,80	0,88	1				
AS	-0,52	-0,47	-0,37	1			
AP	0,24	0,23	0,27	0,05	1		
LP	-0,61	-0,44	-0,33	0,65	0,09	1	
LPS	-0,60	-0,42	-0,34	0,62	0,02	0,94	1

El procedimiento para obtener las componentes principales se basa en el cálculo de los autovalores y los autovectores de la matriz de covarianzas (o la matriz de correlaciones en caso de trabajar con variables estandarizadas). Los autovalores son las varianzas de las componentes principales y los respectivos autovectores contienen a los coeficientes de las combinaciones lineales que conforman las componentes principales. Comprender este significado de los autovalores y autovectores en el PCA, aun sin entender cómo se obtuvieron, alcanza para interpretar los resultados del análisis.

La primera componente tiene una varianza de 4 (Tabla 4). La varianza de las demás componentes es mucho más baja: 1,51; 0,76; 0,45; 0,15; 0,07 y 0,05. Esto indica que la primera componente es mucho más importante que las demás para representar la variación entre los 40 ejemplares para las 7 variables morfométricas registradas. El autovalor de cada componente permite calcular la proporción de la variación que es abarcada por esa componente. En este caso la primera componente explica el 57% de la variación total (Tabla 4) y en conjunto con la segunda acumulan el 79% de la variación observada. Se elige continuar el análisis con estas dos primeras componentes ya que esto permite abarcar casi el 80% de la variación total que contenían las 7 variables originales, mientras la tercera componente solo agrega un 11% de variación explicada y las siguientes mucho menos (Tabla 4). Es decir que al tomar en consideración las dos primeras componentes para interpretar las variaciones de morfometría entre estas bromelias, obtenemos un resumen con mínima pérdida de información. Logramos resumir en el plano generado por las dos primeras componentes gran parte de la información relativa a la semejanza/distancia entre estas plantas para los caracteres medidos.

Tabla 4. Autovalores de los 7 componentes principales o ejes para el ejemplo de la Tabla 1. Proporción de variación explicada por cada componente y proporción de variación acumulada. Las dos primeras componentes (cuya proporción de variación se encuentra resaltada en la tabla) explican el 79% de la variación contenida en las 7 variables originales.

Componente	Autovalor	Proporción	Proporción Acumulada
1	4	**0,57**	0,57
2	1,51	**0,22**	**0,79**
3	0,76	0,11	0,90
4	0,45	0,06	0,96
5	0,15	0,02	0,98
6	0,07	0,01	0,99
7	0,05	0,01	1

Los autovectores e1 y e2 (Tabla 5) contienen los coeficientes de las respectivas componentes principales. Los coeficientes son las ponderaciones de las variables en cada componente principal. La interpretación de estos coeficientes es una parte importante del análisis y permite comprender cómo se relacionan las variables entre sí. En particular, la interpretación asigna significado a las componentes principales, que generalmente representan a los principales gradientes subyacentes que ordenan a los objetos estudiados. En este ejemplo, la primera componente se calcula para cada ejemplar como:

$Z_1 = 0,46 * AV + 0,43 * AH + 0,39 * LTI – 0,36 * AS + 0,08 * AP – 0,40 * LP – 0,40 * LPS$

La segunda componente también se calcula para cada ejemplar como:

$Z_2 = 0,19 * AV + 0,31 * AH + 0,39 * LTI + 0,29 * AS + 0,55 * AP + 0,42 * LP + 0,39 * LPS$

La primera componente en este ejemplo resulta ser un contraste entre el ancho de vainas y hojas y longitud total de la inflorescencia versus ancho de sépalos, longitud de pétalos y de pistilos, mientras el ancho de pétalos no aporta a la variación en este eje (Tabla 5). En este gradiente (Fig. 1) se ordenan los ejemplares desde algunos ubicados a la izquierda del gráfico, como el 16, 13 o 20 que presentan flores con pistilos y pétalos más largos y con sépalos más anchos al mismo tiempo que tienen plantas con vainas y hojas más angostas e inflorescencias más cortas, hasta ejemplares como 31, 36 y 37 en el extremo derecho del gráfico, que son plantas con inflorescencias más largas, vainas y hojas más anchas y al mismo tiempo presentan flores con sépalos más angostos y pétalos y pistilos más cortos. La segunda componente en cambio resulta ser un gradiente de tamaño, es un promedio ponderado entre las variables: ancho de hoja, longitud total de la inflorescencia, longitud y ancho de pétalos (las dos últimas con mayor peso), de sépalos y longitud de pistilos, es decir que los ordena desde los más pequeños abajo hasta los más grandes arriba. Así quedan por ejemplo ordenados, para valores semejantes del contraste del primer eje, hacia arriba los ejemplares 35, 39 y 38 en promedio más grandes en esas variables y hacia abajo los ejemplares 24, 25 y 26 más pequeños.

Tabla 5. Autovectores e1 y e2 asociados a los componentes principales 1 y 2 para el ejemplo de la Tabla 1.

Variables	e1	e2
Ancho vainas (AV)	0,46	0,19
Ancho hojas (AH)	0,43	0,31
Longitud total inflorescencia (LTI)	0,39	0,39
Ancho sépalos (AS)	-0,36	0,29
Ancho pétalos (AP)	0,08	0,55
Longitud pétalos (LP)	-0,40	0,42
Longitud pistilos (LPS)	-0,40	0,39

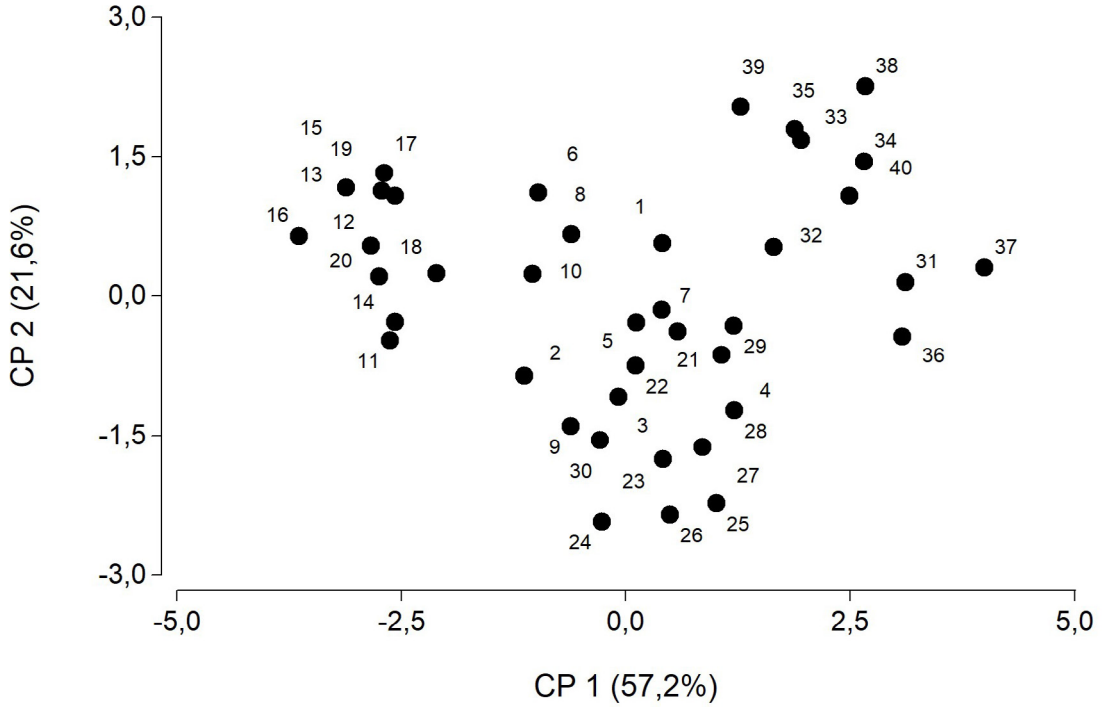

Fig. 1. Ordenamiento de los 40 individuos de *Vriesea procera* (Bromeliaceae) en el espacio de las dos primeras componentes principales (CP1 y CP2). Las componentes principales 1 y 2 explican respectivamente el 57,2 y 21,6 % de la variabilidad total. Ejemplo didáctico adaptado de Uribbe et al. (2020), datos previamente presentados en la Tabla 1.

Hasta aquí pudimos responder las tres primeras preguntas acerca de 1) las relaciones entre variables, 2) su importancia relativa en la explicación de la variación entre ejemplares. Por ejemplo, el ancho de pétalos no tiene relevancia en el gradiente asociado a la primera componente mientras el ancho de hojas y vainas y la longitud de pétalos y pistilos son muy importantes. En ejemplos reales de matrices más grandes también se pueden identificar variables con peso insignificante en ambas componentes, y 3) la identificación de componentes que resumen los principales sentidos de variación y permiten ordenar a los ejemplares por su similitud en esas combinaciones particulares de caracteres. En cambio, para contestar la cuarta pregunta referida a si existen agrupamientos naturales o si los cambios a lo largo de las componentes principales son continuos, debemos primero aclarar que el PCA solo ofrece respuesta a este tipo de preguntas cuando los agrupamientos naturales son muy obvios o cuando, por el contrario, la ausencia de discontinuidades es muy evidente. Esto se debe a que claramente el objetivo de este análisis no es agrupar observaciones sino ordenarlas siguiendo sus gradientes de máxima variación. En este ejemplo sólo se ve con cierta nitidez un agrupamiento a la izquierda del plano del ordenamiento (Fig. 1), limitado por los ejemplares 17, 18 y 11 y otro agrupamiento más disperso y mucho menos claro hacia la derecha del gráfico, limitado por los ejemplares 39, 32 y 36.

Se puede volcar sobre el mismo gráfico del ordenamiento información adicional. Por ejemplo, se puede incluir información que no haya participado del análisis (variables suplementarias) como alguna categoría de agrupamiento de las unidades conocida a priori (en este ejemplo localidades de latitudes diferentes en la Mata Atlántica; Fig. 2). Es importante que el lector esté advertido de cuál es la información superpuesta, que no es en sí misma resultado de PCA, porque el mismo patrón de ordenamiento puede aparecer muy diferente y resultar engañoso (basta comparar las figuras 1 y 2 para notarlo).

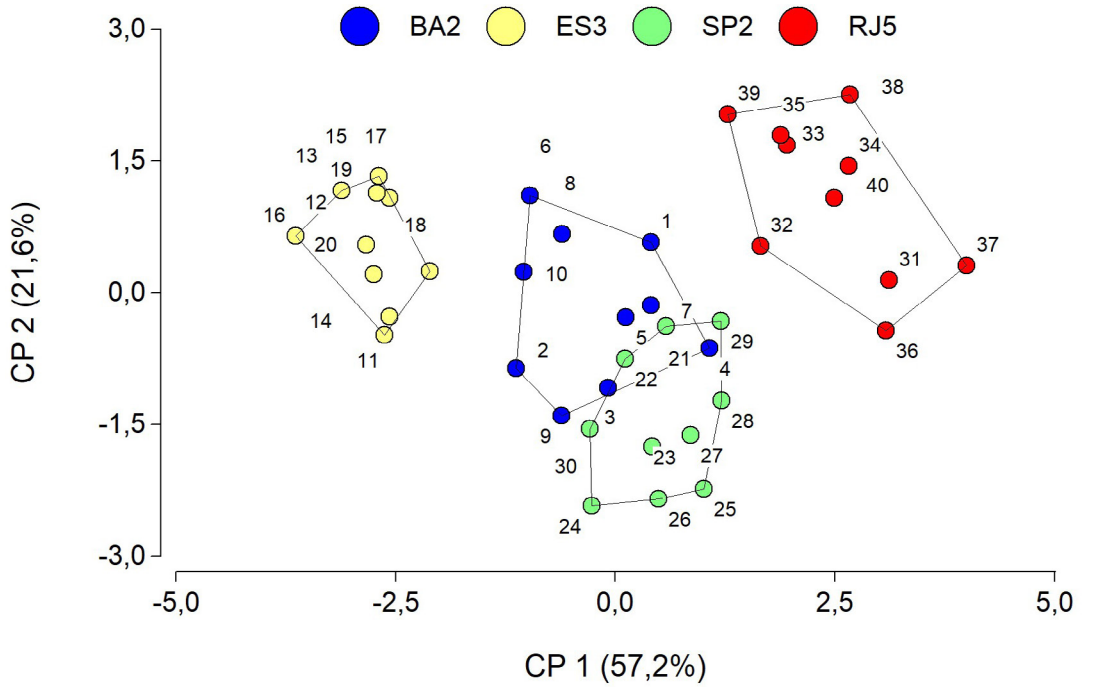

Fig. 2. Ordenamiento de los 40 individuos de *Vriesea procera* (Bromeliaceae) en el espacio de las dos primeras componentes principales (CP1 y CP2), exactamente el mismo ordenamiento de la Fig. 1. Aquí se incluye, además, la identificación de las poblaciones de origen diferenciadas por los colores y los polígonos. La población de origen es una variable categórica que no participó en el análisis de componentes principales. Figura en color en la versión en línea https://www.ojs.darwin.edu.ar/index.php/darwiniana/article/view/1086/1299

No siempre se logra concentrar en pocas combinaciones lineales (componentes) una proporción alta de la variación de las variables originales. Por ejemplo, en la situación extrema en que las variables originales están poco correlacionadas entre sí, el análisis no puede concentrar la representación de la información en pocas dimensiones. En cambio, los mejores resultados se alcanzan cuando las variables se encuentran altamente correlacionadas, ya sea positiva o negativamente. El análisis de componentes principales no se apoya necesariamente en el supuesto de distribución normal multivariada, sin embargo, en las situaciones en que este supuesto se cumple, el análisis se comporta mejor. En general se alcanza un buen desempeño si las variables presentan cierta simetría alrededor de la media con relaciones entre variables aproximadamente lineales. Por el contrario, si las variables presentan distribuciones de frecuencias muy asimétricas, o si las matrices de datos resultan poco densas (muchos ceros), se distorsionan los patrones que el investigador busca descubrir en el subespacio de dimensión reducida. Para ese tipo de datos se recomiendan otros métodos de ordenamiento como Análisis de Correspondencia y Escalamiento Multidimensional, que presentaremos más adelante. Eventualmente, también se pueden transformar algunas variables mediante una función como logaritmo o raíz cuadrada, para mejorar la simetría de la distribución.

En la bibliografía botánica y ecológica se encuentran muy buenos ejemplos aplicados del Análisis de Componentes Principales. Un trabajo muy interesante que estudió varias especies del género *Paspalum* mediante datos morfológicos y citológicos aplica esta técnica en dos etapas (Bonasora et al., 2015). Primero estudia separadamente mediante imágenes la forma del antecio superior de la espiguilla y condensa la información en el primer eje de un PCA que explica el 87% de variación. Esta nueva variable junto con el área medida en píxeles del antecio superior participa luego en otro análisis de PCA que involucra a otros 8 caracteres morfométricos registrados en los mismos ejemplares de *Paspalum* estudiados. El PCA de la matriz completa de datos muestra dos especies que se separan muy bien hacia ambos extremos del componente principal 1, que se relaciona principalmente con medidas de la espiguilla y que explica el 62% de la variabilidad. El mismo análisis permite a los autores describir una especie nueva que se separa en ambas componentes principales. Otros trabajos también combinan muy bien el PCA con técnicas de Aglomeración Jerárquica (p.e. Zapater et al., 2014), otros presentan de manera muy clara la interpretación de los autovectores y las componentes (p.e. Nagahama et al., 2012), y otros transforman caracteres cualitativos para combinarlos correctamente con los cuantitativos para realizar el PCA (p.e. Giussani et al., 2000). Uchida y Ushimaru (2015) construyeron modelos para identificar determinantes de la diversidad en paisajes seminaturales en Japón y utilizan PCA para sintetizar en una componente o a veces en dos componentes algunas de las variables registradas para caracterizar el uso del paisaje (por ejemplo, superficies dentro de un radio de 1 km alrededor de cada sitio que presentan cultivos intensivos, cultivos tradicionales, bosque o cultivos abandonados) y después utilizaron las componentes obtenidas con el PCA junto a otras variables predictoras en un modelo estadístico que predice la biodiversidad de artrópodos.

Escalamiento Multidimensional

El Escalamiento Multidimensional es un método que busca simplificar y revelar la estructura de un conjunto de observaciones con información multivariada mediante su ordenamiento en un espacio de dimensión reducida. Parte de una matriz de distancias entre los n objetos y produce un mapa donde las distancias entre ellos en el espacio reducido se aproximen lo más posible a las distancias originales en el espacio n dimensional. En muchas circunstancias, los ejes de este espacio de representación son interpretables y se pueden utilizar para comprender mejor los datos, pero lo que siempre es interpretable en el Escalamiento Multidimensional es la distancia relativa entre las observaciones en el mapa obtenido. Esto se asocia con la gran ventaja que presenta este método que es la versatilidad que ofrece para elegir la medida de distancia que resulte más apropiada para el tipo de datos que se está evaluando y para la/s pregunta/s que el investigador desea formular acerca del sistema bajo estudio.

Tabla 6. Primeras filas de la matriz de datos extraída de Puhl et al. (2014), reducida con fines didácticos. Cobertura de las especies en 5 sitios de ambientes mesofíticos (columnas comenzadas con A) y 5 sitios de ambientes halomórficos (columnas comenzadas con H), censados en 1968 y recensados 35 años después (indicado con asterisco). La tabla completa utilizada en este ejemplo puede descargarse del Material suplementario II, https://www.ojs.darwin.edu.ar/index.php/darwiniana/article/view/1086/1301

Especie	A38	A38*	A87	A87*	A217	A217*	A242	A242*	A271	A271*	H96	H96*	H246	H246*	H273	H273*	H288	H288*	H324	H324*
Paspalum dilatatum	3,0	1,0	3,0	2,0	2,0	3,0	1,0	2,0	1,0	2,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Apium leptophyllum	1,0	0,5	1,0	1,0	0,0	0,0	1,0	0,5	0,0	1,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Centaurium pulchellum	0,5	0,1	1,0	0,5	1,0	0,1	0,5	0,5	1,0	0,5	0,0	0,5	0,0	0,0	0,0	0,1	0,0	0,1	0,0	0,1
Cirsium vulgare	0,0	0,5	0,0	0,5	2,0	0,1	0,5	0,5	1,0	0,5	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Hypochoeris radicata	0,0	0,0	0,0	0,1	0,0	0,5	0,0	2,0	0,0	2,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
...																				

Las variables podrían ser cuantitativas continuas (por ejemplo, altura de planta), binarias (como presencia/ausencia), categóricas (alta/baja) o de conteo (Caja 1). Diferentes variantes de Escalamiento Multidimensional son el Escalamiento Multidimensional Métrico, a veces también denominado Análisis de Coordenadas Principales (PCoA por su sigla en inglés) y su pariente cercano, el Escalamiento Multidimensional No Métrico (NMDS por su sigla en inglés). Este último, en lugar de representar las distancias por su valor directo lo hace por el ranking de las distancias.

Como ejemplo de aplicación de Escalamiento Multidimensional se utilizará una versión reducida del conjunto de datos del trabajo de Puhl et al. (2014). En aquel trabajo se estudiaron los cambios ocurridos en la diversidad de especies de dos comunidades de pastizal de la Pampa Deprimida: la Pradera de Mesófitas y la Estepa de Halófitas (Perelman et al., 2001). La matriz de datos constaba de censos de vegetación realizados en 1968 en 51 sitios de pastizal de esas dos comunidades y nuevamente relevados de manera idéntica en 2003, 35 años después. En ambos momentos se registró la abundancia/cobertura de todas las especies presentes, las que totalizaron 286 especies. Con fines didácticos hemos tomado un muy pequeño extracto de aquella matriz original, el cual solo contiene 20 casos (10 sitios x 2 fechas) y 22 especies (Tabla 6; Material suplementario II, https://www.ojs.darwin.edu.ar/index.php/darwiniana/article/view/1086/1301). Aquí se utilizarán las distancias de Bray-Curtis entre pares de sitios para caracterizar las diferencias en composición florística entre ellos, tomando en cuenta la cobertura de las especies en cada par de sitios.

Consideraremos las siguientes preguntas que se le podrían formular a estos datos:

1. Las diferencias en composición de especies que se encontraron inicialmente entre las comunidades de mesófitas y de halófitas (las que sirvieron para diferenciarlas y caracterizarlas) ¿se mantienen aún después de 35 años de intervención antrópica?

2. ¿Cambió la composición florística de estos pastizales en el tiempo transcurrido entre relevamientos?

3. Si hubo cambios en el tiempo ¿tuvieron diferente dirección y/o intensidad entre las comunidades?

Con el Escalamiento Multidimensional de la matriz de distancias se obtienen las coordenadas de los sitios en un espacio de dos dimensiones, donde las distancias son representadas con la mayor exactitud posible. Algunos autores denominan a este espacio "plano del escalamiento". Entonces, en el plano del escalamiento obtenido con los datos del ejemplo (Fig. 3) se interpretan distintos aspectos de las distancias entre los objetos:

1. El primer eje del escalamiento, que explica aprox. un 50% de la variación, está claramente asociado a las diferencias entre pastizales de las comunidades de mesófitas y de halófitas, es decir que la diferencia entre estos ambientes conforma el gradiente principal de la composición taxonómica en los sitios de pastizal estudiados y esto se verifica en ambas fechas de relevamiento.

2. El segundo eje, que explica aprox. un 12 % de la variación, se asocia con cambios en la composición florística en el tiempo, ya que en general en el

gráfico del ordenamiento, los censos de la segunda fecha de muestreo, indicados con asterisco, aparecen ubicados más abajo que sus pares originales.

3. Además, se observa mayor distancia entre los censos del tiempo 0 y el tiempo 35 para los censos de la comunidad de mesófitas ubicados a la derecha en el plano del escalamiento. En cambio, en ese mismo gradiente temporal representado en el eje 2, se observa que a la izquierda del gráfico la comunidad de halófitas presenta menor modificación de la composición en el tiempo.

En la literatura botánica y ecológica se encuentran otras muy buenas aplicaciones de Escalamiento Multidimensional y Escalamiento Multidimensional No Métrico. Son trabajos que utilizan medidas particulares de distancia apropiadas al tipo de datos y a la pregunta que se formula el proyecto de investigación. Por ejemplo, en un trabajo que evaluó la delimitación intraespecífica de dos variedades de *Deyeuxia velutina* (Ferrero et al., 2020), las autoras seleccionaron 61 variables morfométricas en 52 ejemplares y calcularon la distancia de Manhattan para hacer un Escalamiento Multidimensional (PCoA). Obtuvieron un ordenamiento que explicaba el 36% de la variación, donde se diferenciaban dos grupos de ejemplares. Luego, las autoras calcularon una regresión entre los ejes del escalamiento y la latitud y longitud del sitio de recolección de cada ejemplar que resultó significativa, el eje 2 que separaba los grupos se correlacionaba con el gradiente longitudinal. Chaneton et al. (2002), realizaron un escalamiento NMDS con distancia Bray-Curtis para mostrar que el pastoreo doméstico tiende a homogeneizar la vegetación en el paisaje.

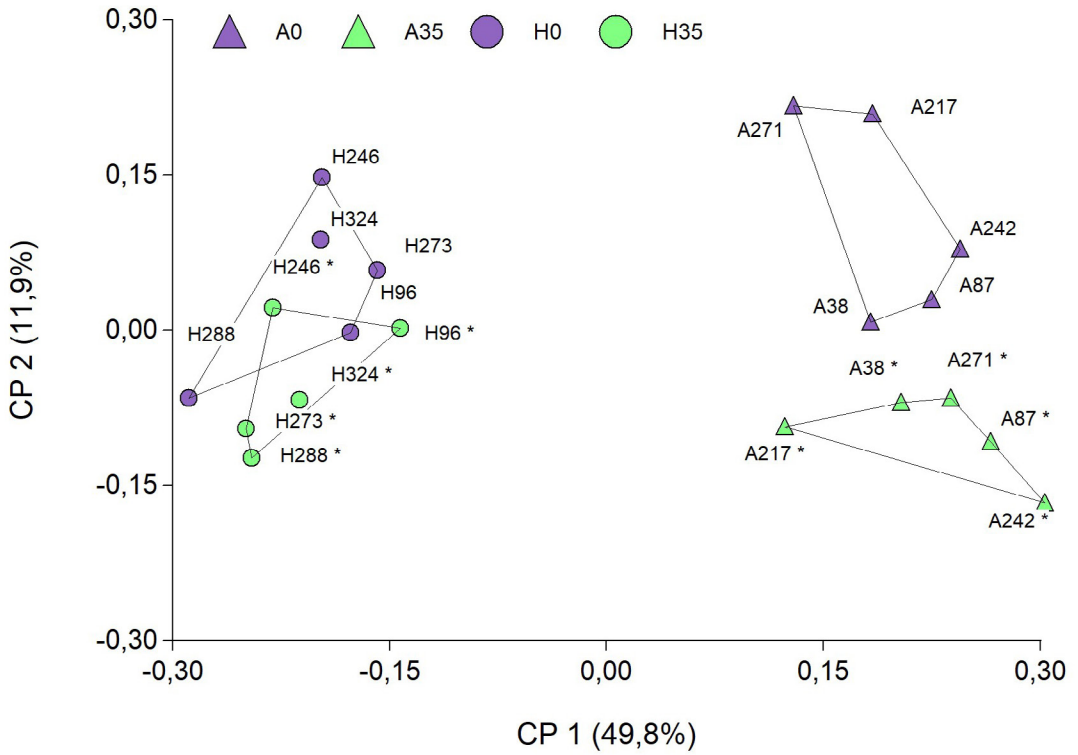

Fig. 3. Primeros dos ejes o coordenadas principales (CP1 y CP2) resultantes del Escalamiento Multidimensional de 20 sitios del pastizal de la Pampa Deprimida, pertenecientes a dos comunidades: la Pradera de Mesófitas (triángulos) y la Estepa de Halófitas (círculos). La medida de distancia utilizada es Bray-Curtis. El asterisco (*) después del número indica mismo sitio relevado 35 años después. Las componentes principales 1 y 2 explican el 49,8 y 11,9 % de la variabilidad total respectivamente. Ejemplo didáctico adaptado de Puhl et al. (2014). Figura en color en la versión en línea https://www.ojs.darwin.edu.ar/index.php/darwiniana/article/view/1086/1299

En el plano del ordenamiento se observaron mayores distancias entre ambientes frecuentemente inundados y aquellos que no se inundan para los sitios clausurados a herbívoros domésticos, mientras en los lotes pastoreados los sitios con diferente frecuencia de inundación aparecen superpuestos en el mismo plano del escalamiento. Otros trabajos combinan muy bien los escalamientos con Análisis de Aglomeración Jerárquica (Cluster Analysis), ya que realizan un agrupamiento jerárquico de las observaciones y luego los representan en un ordenamiento obtenido con Escalamiento Multidimensional con la misma medida de distancia utilizada en el agrupamiento, como Ulloa et al. (2011). En ese trabajo, el objetivo era detectar nuevas características de la morfología de la lema para clarificar 24 taxones en *Polypogon*. Los autores evaluaron 9 variables, una de las cuales es continua y el resto son multinomiales ordinales, utilizando la distancia de Gower y el método de Ward para armar grupos y luego proyectar las distancias en un Escalamiento Multidimensional (PCoA); indicando los grupos. También Zallochi et al. (1992), para diferenciar taxones del género *Macroptilium* con datos cromatográficos utilizaron datos de presencia y ausencia de 36 caracteres químicos en 51 individuos de 8 especies, calcularon la distancia de Manhattan entre ejemplares, construyeron grupos mediante un método de agrupamiento jerárquico y con la misma distancia del Escalamiento Multidimensional mostraron los grupos en 3 dimensiones que explicaban el 91,8% de la variación. Batista et al. (2014), también a continuación de una clasificación, utilizaron el Escalamiento Multidimensional para resumir diferencias y afinidades en composición florística entre comunidades vegetales del Parque Nacional El Palmar, con la particularidad de que en este ejemplo la medida de distancia se calculaba sobre las constancias de las especies en las comunidades (es decir, un resumen derivado de la clasificación previa de los sitios en comunidades). Con un experimento de largo plazo en microcosmos (Chase, 2010) se puso en evidencia mediante un Escalamiento Multidimensional no Métrico (NMDS), que en ambientes más productivos se potencia el componente estocástico en la determinación de la composición taxonómica de las comunidades. En el ordenamiento, los sitios de los ambientes más pobres aparecen muy concentrados ocupando muy pequeñas superficies (distancias pequeñas implican mayor semejanza en composición de especies). Al mismo tiempo los sitios de ambientes productivos aparecen mucho más dispersos porque la composición florística varía más entre ellos. Chase (2010) eligió utilizar con el NMDS una medida de distancia (índice de Raup-Crick, Legendre & Legendre, 2003), que permite comparar la composición de especies entre pares de sitios sin la influencia que ejerce sobre la mayoría de los índices el número total de especies. Esto es así porque el índice cuantifica el nivel de semejanza respecto a la esperada por azar para la misma riqueza de los sitios en cuestión. Estos ejemplos reflejan la ventaja que ofrece el Escalamiento Multidimensional por su versatilidad debida a las diferentes medidas de distancia.

Análisis de Correspondencia

(CA por su sigla en inglés)

Este análisis permite representar en un ordenamiento de pocas dimensiones la información contenida en una tabla de contingencia, en consecuencia, es muy útil para analizar variables categóricas. Las tablas de contingencia reúnen valores de conteos de casos en una clasificación cruzada, como los cuadros que se obtienen al organizar las respuestas a dos preguntas en una encuesta. Por ejemplo, se podrían observar dos caracteres en todas las especies leñosas exóticas de nuestro país (Yansen & Biganzoli, 2022) como el tipo de fruto (baya, legumbre, samara, drupa, etc.) y el tipo de dispersión predominante de sus semillas (anemocoria, zoocoria, hidrocoria, autocoria, etc.). Podemos contar luego cuántas especies coinciden en presentar cada combinación de forma de fruto y de tipo de dispersión y completar con esa información una tabla en la cual asignaremos, por ejemplo, las formas de frutos a las filas y los tipos de dispersión a las columnas (Tabla 7). En este ejemplo, los tipos de frutos o de dispersión que no sumaron en total al menos 5 especies, fueron eliminados del análisis, esta es una práctica recomendada en CA para evitar distorsiones.

Tabla 7. Número de especies de leñosas exóticas coincidentes con cada una de las categorías de tipo de fruto y forma predominante de dispersión de las semillas (adaptado de Yansen & Biganzoli, 2022).

	Dispersión			
Fruto	Anemocoria	Autocoria	Zoocoria	Total
Baya		1	7	8
Cápsula	22	1	3	26
Cono	13	1	1	15
Drupa			22	22
Legumbre	3	8	1	12
Pomo			7	7
Sámara	8			8
Total	46	11	41	98

Tabla 8. Porcentaje de especies de leñosas exóticas coincidentes con cada una de las categorías de tipo de fruto y forma predominante de dispersión de las semillas (datos originales en Tabla 7).

	Dispersión			
Fruto	**Anemocoria**	**Autocoria**	**Zoocoria**	Total
Baya	0	12,5	87,5	100
Cápsula	84,6	3,8	11,5	100
Cono	86,7	6,7	6,7	100
Drupa	0	0	100	100
Legumbre	25	66,7	8,3	100
Pomo	0	0	100	100
Sámara	100	0	0	100
Promedio	**46,9**	**11,2**	**41,8**	100

Por ejemplo, en estudios de comunidades, cuando una especie rara aparece sólo en un ambiente muy pobre en especies se recomienda que no sea incluida en el análisis (Greenacre, 2013), pero su omisión debe ser informada al igual que todo otro pretratamiento de los datos. La Tabla 7 muestra la cantidad de especies que coinciden en cada combinación de respuestas y nos permite estudiar si hay correspondencia entre las categorías de filas y las de columnas, o si por el contrario las formas de los frutos y el tipo de dispersión de las semillas son independientes entre sí.

La tabla de contingencia también nos ayuda a identificar cuáles son las combinaciones de categorías que muestran mayor correspondencia o alejamiento de la independencia. Esto se percibe más claramente cuando se expresa la cantidad de especies de cada clase de dispersión como porcentaje del total de especies que presentan cada tipo de fruto (Tabla 8), así se puede comparar el perfil de dispersión de cada fruto con el perfil total o promedio que se presenta en la última fila de la tabla. Si se trata de tablas muy grandes (muchas categorías de respuesta), no es fácil observar estas características del sistema que estamos estudiando con la sola inspección visual de la tabla y tampoco resulta sencillo ordenar las categorías de filas y columnas según los ejes de máxima correspondencia.

Consideraremos las siguientes preguntas que podemos formular a partir de la información contenida en la tabla de contingencia (Tabla 7):

1. ¿El tipo de dispersión predominante de las semillas se encuentra asociado con el tipo de fruto entre las especies leñosas exóticas de Argentina?
2. ¿Se pueden ordenar los tipos de dispersión según el perfil de frecuencias que presentan entre tipo de frutos?
3. ¿Se pueden identificar gradientes principales de asociación entre tipos de frutos y forma de dispersión predominante?

El CA es la herramienta adecuada para resolver estas cuestiones, ya que genera un ordenamiento que muestra la correspondencia entre las variables categóricas consideradas. Al igual que en el PCA, en el CA se obtienen autovalores y autovectores, pero en este análisis existen dos conjuntos de autovectores que corresponden a las filas y a las columnas, respectivamente. Los autovectores contienen los coeficientes de las combinaciones lineales que permiten calcular las coordenadas para la representación simultánea de filas y columnas en los ejes de máxima correspondencia (Tabla 9). La "inercia total" de la tabla de contingencia, que mide el grado de falta de independencia entre las categorías de filas y columnas, se calcula como un coeficiente Chi cuadrado, es decir

una suma de las diferencias entre los conteos observados y los esperados bajo independencia en cada combinación de categorías, doblemente relativizados por los totales de filas y columnas. Los autovalores se interpretan de manera similar al PCA, aunque se expresan en términos de "inercia" o falta de independencia y se comunica el % de inercia explicado por los sucesivos ejes principales. Los primeros ejes deberían explicar un porcentaje elevado de la inercia total de la matriz de datos.

La representación conjunta de casos y variables (filas y columnas) en el CA permite interpretar las relaciones entre ellos. Las filas y columnas que aparecen cercanas en el plano del ordenamiento indican que las variables presentan en esos casos valores mayores a los que se predicen bajo condición de independencia (Fig. 4).

Tabla 9. Autovalores con el porcentaje de inercia asociado a cada uno, autovectores de filas y de columnas para los dos primeros ejes resultantes del análisis de correspondencia para las categorías de tipo de fruto y forma de dispersión de semillas de las especies leñosas exóticas (datos originales en Tabla 7).

	Eje 1	Eje 2
Autovalor	0,78	0,43
% inercia	64,3	35,7

Autovectores de columnas

	Eje 1	Eje 2
Anemocoria	0,820	-0,341
Autocoria	0,418	1,825
Zoocoria	-1,032	-0,106

Autovectores de filas

	Eje 1	Eje 2
Baya	-0,963	0,205
Cápsula	0,669	-0,351
Cono	0,758	-0,276
Drupa	-1,168	-0,161
Legumbre	0,450	1,705
Pomo	-1,168	-1,168
Samara	0,925	-0,412

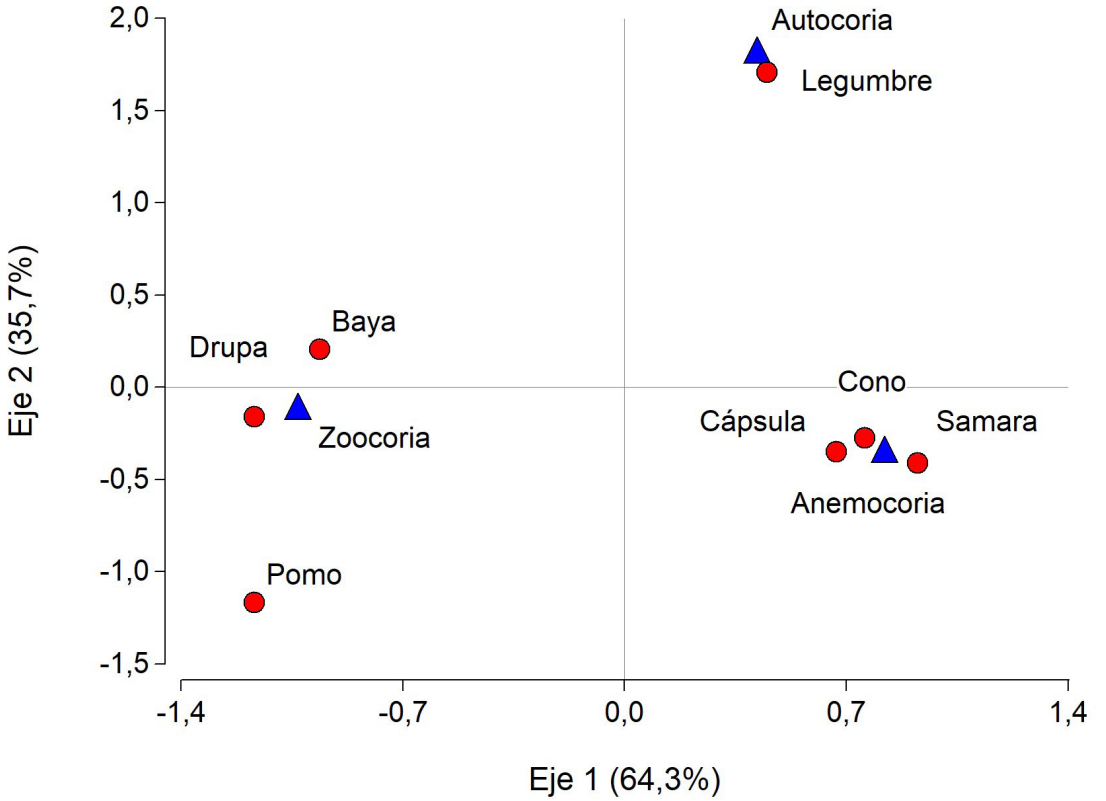

Fig. 4. Coordenadas en el plano de los dos primeros ejes del Análisis de Correspondencia de las categorías forma de dispersión (triángulos) y tipo de fruto (círculos) de las especies leñosas exóticas de Argentina (Yansen & Biganzoli, 2022). El eje 1 y 2 explican el 64,3 y 35,7% de la inercia total respectivamente. Figura en color en la versión en línea https://www.ojs.darwin.edu.ar/index.php/darwiniana/article/view/1086/1299

El ordenamiento en dos dimensiones explica una proporción muy alta de la falta de independencia (64,3% y 35,7% de inercia, respectivamente) entre los dos criterios de clasificación asignados a estas especies. Se observa que el primer eje separa los frutos con predominante zoocoria de los de predominante anemocoria mientras el segundo separa a estos dos grupos respecto al de las legumbres, en las que predomina la autocoria en las especies exóticas de Argentina (Fig. 4). Asimismo, permite ordenar los tipos de frutos según sus perfiles de dispersión predominante (Fig. 5).

Por extensión el CA se puede aplicar a matrices que no son exactamente tablas de contingencia pero que presentan datos no negativos, con todas las variables medidas en la misma escala y para los cuales interesa comparar perfiles. Es decir, datos para los cuales tenga sentido su expresión relativa a los totales de filas o columnas. En ecología de comunidades este método ha sido ampliamente usado para estudiar bases de datos que tienen en filas las especies identificadas en cada sitio y en columnas los sitios relevados, los datos de las celdas de la matriz pueden ser de presencia/ausencia (1 y 0 respectivamente) o de abundancia (densidad o cobertura) como los de la Tabla 6, la versión reducida de los datos de Puhl et al. (2014) que ya analizamos con Escalamiento Multidimensional. El CA en estos estudios muestra la correspondencia entre sitios y especies, las relaciones entre especies que coinciden en habitar ambientes semejantes y las relaciones entre sitios que coinciden en compartir las mismas especies. Ordena simultáneamente a los sitios y las especies en ejes principales de máxima correspondencia, ejes que en general coinciden con los principales gradientes ambientales responsables de la distribución de las especies en los sitios estudiados. Por ejemplo, en Burkart et al. (1998) los ejes de CA permitieron reconocer los gradientes del paisaje que operan como principales determinantes de la composición florística de los pastizales en una subregión ubicada al sur de la Pampa Deprimida (partidos de Ayacucho, Maipú y Dolores). En Perelman et al. (2001) el ordenamiento de 749 censos de vegetación correspondientes a relevamientos que abarcaron toda la extensión latitudinal de los pastizales de la Pampa Deprimida mostró que en los primeros dos ejes de CA (que explican el 61,8% de la inercia) la composición florística responde

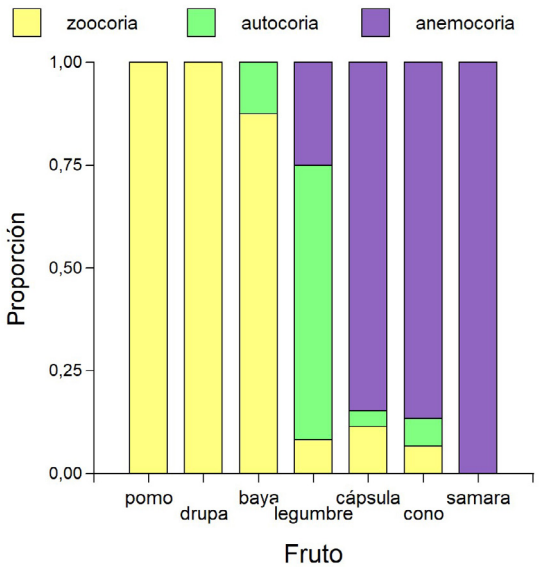

Fig. 5. Proporción del tipo predominante de dispersión de las especies leñosas exóticas para cada una de las categorías de tipo de fruto. Representa las filas de la Tabla 8 ordenadas por el resultado de CA. Figura en color en la versión en línea https://www.ojs.darwin.edu.ar/index.php/darwiniana/article/view/1086/1299

principalmente a los gradientes topográfico y halomórfico, mientras sólo en el tercer eje (19,6%) se pone de manifiesto la respuesta al gradiente regional de latitud. Es decir que los principales determinantes de la heterogeneidad de la vegetación en la escala de paisaje trascienden a la escala regional donde superan en importancia a los factores que operan a mayor escala, asociados a la latitud. En Puhl et al. (2014) el CA permitió comparar la trayectoria de pastizales de diferentes comunidades del norte de la Pampa Deprimida después de 35 años de uso antrópico. Perelman et al. (2003a) utilizaron CA con datos de presencia/ausencia de las especies en sitios dominados por una gramínea cespitosa de gran porte (*Paspalum quadrifarium*), especie ingeniera del ecosistema que modifica fuertemente la fisonomía de los sitios donde domina. El ordenamiento permite reconocer que más allá de la dominancia de la gramínea cespitosa, se pueden identificar tres comunidades vegetales que responden al ambiente abiótico de los diferentes sitios. A partir de allí se pudo estudiar el impacto de la especie ingeniera en cada una de las comunidades. Cué-Hernández et al. (2022) estudiaron la composición y la abundancia

Caja 1. Comparación de los métodos multivariados tratados en este trabajo con los principales usos y tipo de datos.

Método	Objetivo	Datos	Medida de distancia
Análisis de componentes principales	Ordenar los casos en ejes de máxima variación	Variables continuas, simétricas y relacionadas linealmente	Euclídea
Escalamiento multidimensional	Representar en pocas dimensiones las distancias entre los casos	Variables continuas, discretas, binarias o multinomiales	Elegida por el usuario
Análisis de correspondencia	Ordenar las categorías de filas y columnas en ejes de máxima correspondencia	Tablas de contingencia con conteos o porcentajes para las categorías de filas y columnas o variables binarias	Chi-cuadrado
Análisis discriminante	Ordenar los grupos de casos en ejes de máxima separación. Clasificar casos en los grupos	Variables continuas y una variable categórica que diferencia entre grupos de casos.	Mahalanobis
Análisis de conglomerados o *Cluster*	Descubrir grupos de casos o variables	Variables continuas, discretas, binarias o multinomiales	Elegida por el usuario

de visitantes florales en dos variedades de *Phaseolus coccineus* cultivadas bajo sistemas de manejo diferentes en el altiplano de Puebla, México. Realizaron observaciones simultáneas coincidentes con el pico de floración, durante ocho días, contabilizando el número de visitantes florales y de visitas y aplicaron un análisis de correspondencia para estudiar la importancia relativa del manejo agrícola y del color de las flores sobre la cantidad y diversidad de visitantes florales.

En sistemas donde el gradiente principal es tan importante que eclipsa fuertemente a los gradientes secundarios, el CA puede presentar un segundo eje que es casi una función cuadrática (una parábola) del primer eje. Esto se denominó en la bibliografía "efecto de arco" y algunos lo consideraron un artefacto matemático que debía ser corregido. Así surgió un método denominado DCA (*detrended correspondence analysis*, Hill & Gauch, 1980) específicamente desarrollado para corregir ese supuesto defecto del CA. El DCA fue muy popular en la ecología de comunidades durante las décadas de 1980 y 1990, no así en otras ramas de la ciencia, como la economía, la sociología o la lingüística, que también aplican con frecuencia el CA. Al mismo tiempo, algunos autores alertaron acerca de las distorsiones que puede producir el procedimiento de "detrending" sobre resultados del CA potencialmente reveladores y especialmente advirtieron que la condición de óptimos que tienen los ejes de CA obtenidos según el procedimiento algebraico original se pierde en el DCA (Digby & Kempton, 1987; Jackson & Sommers, 1991; Kenkel et al., 2002).

Análisis Discriminante

(DA por su sigla en inglés)

Este análisis sólo se aplica cuando los objetos bajo estudio pertenecen a grupos preexistentes, por ejemplo, son plantas de diferentes especies (grupo = especie), observaciones en diferentes localidades (grupo = localidad) o macetas que recibieron diferentes tratamientos de humedad (grupo = tratamiento). Entonces la matriz de datos debe contener además de las variables cuantitativas que caracterizan a cada objeto, una variable de tipo categórica que indica la clase o grupo de pertenencia para ese objeto. El propósito principal del DA es describir la separación entre los grupos, para eso debe encontrar combinaciones lineales de las variables, las cuales determinen ejes sobre los que se puedan discriminar los grupos lo mejor posible. Mediante este análisis, el espacio p-dimensional (donde p es el número de variables originales) se reduce a un subespacio de menos dimensiones (1 o 2 ejes) formado por las combinaciones lineales de las variables que mejor explican la separación de los grupos. Una vez encontradas esas combinaciones, denominadas funciones discriminantes, se puede realizar la clasificación de un nuevo objeto en el mismo subespacio. La regla de predicción de ser miembro de un determinado grupo es en muchas circunstancias de gran utilidad. Por ejemplo, si se trata de tejidos que provienen de organismos enfermos con diferentes patologías (grupos = enfermedades), una nueva muestra de tejido en la que se miden las mismas variables puede servir para diagnosticar la enfermedad.

De manera semejante si se trata de clasificaciones taxonómicas de plantas u otros organismos a partir de variables morfométricas. El DA permite entonces, tanto describir las diferencias entre grupos en términos de las variables que más contribuyen a la separación, como predecir la pertenencia de una nueva observación a un grupo.

Aquí retomaremos el ejemplo de Uribbe et al. (2020), el cual habíamos analizado previamente con PCA, análisis que nos permitió responder una serie de preguntas. En el contexto de DA podemos responder otra pregunta de interés que aún no había sido formulada:

5. ¿Será posible identificar para estas bromeliáceas las variables o combinaciones lineales de variables que mejor discriminen entre las poblaciones naturales de diferentes latitudes (grupos) en la Mata Atlántica?

Con este fin tomaremos en consideración una variable que no participó del análisis de PCA, se trata de la variable categórica que identifica la población natural a la que corresponde cada observación (columna denominada Población en la Tabla 1). Se obtienen como resultado las funciones discriminantes que cuantifican el peso de cada variable en las combinaciones lineales de máxima discriminación entre grupos (Tabla 10) y la representación gráfica de las observaciones en el plano determinado por las dos primeras funciones discriminantes (Fig. 6). A diferencia de la representación que generó el PCA para los mismos individuos (Figs. 1 y 2), aquí vemos a los grupos muy bien separados en el plano con grandes discontinuidades entre ellos. Mediante la función discriminante se asignan los individuos a los grupos y se puede evaluar el margen de error en la clasificación. La primera función discriminante en este ejemplo separa fuertemente a las plantas de la población de Itagacú (ES3) respecto al resto y también diferencia a las poblaciones de Itamarajú (BA2) y Ubatuba (SP2) entre sí. Esta función discriminante es una combinación lineal que representa un contraste entre ancho de sépalos, vainas y hojas versus longitud de pistilos y de pétalos (Tabla 10) y puede expresarse de la siguiente manera:

$F_1 = -0,34 * AV - 0,26 * AH - 0,03 * LTI - 0,94 * AS - 0,03 * AP + 0,38 * LP + 1,12 * LPS$

Tabla 10. Coeficientes de las funciones discriminantes correspondientes a los dos primeros ejes para el ejemplo de *Vriesea procera* (Bromeliaceae) adaptado de Uribbe et al. (2020).

	Eje 1	Eje 2
Ancho vainas (AV)	-0,34	-0,37
Ancho hojas (AH)	-0,26	0,95
Longitud total inflorescencia (LTI)	-0,03	0,51
Ancho sépalos (AS)	-0,94	-0,28
Ancho pétalos (AP)	-0,03	0,11
Longitud pétalos (LP)	0,38	0,24
Longitud pistilos (LPS)	1,12	0,42

La segunda función discriminante separa claramente a las plantas de Angra do Reis (RJ5) respecto a las otras poblaciones. Representa un contraste entre ancho de hojas, longitud de las inflorescencias, longitud de pétalos y longitud de pistilos versus ancho de vainas y ancho de sépalos. Se calcula para cada planta mediante la siguiente expresión:

$F_2 = -0,37 * AV + 0,95 * AH + 0,51 * LTI - 0,28 * AS + 0,11 * AP + 0,24 * LP + 0,42 * LPS$

Las Tablas 5 y 10 presentan diferentes coeficientes de ponderación para las mismas variables de la misma matriz de datos (Tabla 1) y generan en consecuencia diferentes ordenamientos (Figs. 1 y 2 y Fig. 6). El primer ordenamiento (Figs. 1 y 2), obtenido con PCA, expresa la máxima variación entre individuos para las variables medidas, el segundo, obtenido con DA, genera la máxima discriminación entre individuos de diferentes poblaciones de origen.

El Análisis Discriminante tiene supuestos de normalidad y de homogeneidad en la estructura de variación y covariación entre grupos, ya que trabaja con una matriz de varianzas-covarianzas común a todos los grupos. La violación de supuestos se puede corregir con transformaciones de escala de las variables. El método es más robusto a alejamientos del supuesto de normalidad multivariada que a la falta de cumplimiento del supuesto de homogeneidad de matrices de varianzas-covarianzas. Existe un análisis discriminante no paramétrico para situaciones donde no se puede asumir distribución normal y

también un análisis discriminante no lineal que trabaja con matrices de varianzas-covarianzas por grupo, cuando no se puede asumir homogeneidad en la estructura de variación y covariación entre los grupos. Este último método se conoce como Análisis Discriminante Cuadrático y puede consultarse en Palacio et al. (2020).

Las aplicaciones del Análisis Discriminante en botánica no requieren gran presentación ya que en la cuna de este método estuvo presente la morfometría, con un famoso ejemplo en el que se clasificaron 150 ejemplares entre tres especies de *Iridaceae* según los caracteres morfológicos medidos en sus flores. El ejemplo, basado en los datos colectados en la Península Gaspé (este de Quebec, Canadá) por el botánico Edgar Anderson del Jardín Botánico de Missouri, fue utilizado por Fisher en 1936 para presentar el método de análisis (Legendre & Legendre, 2003) y luego fue mencionado en numerosos libros y tutoriales de programas estadísticos. Dado este prestigioso antecedente el lector puede encontrar fácilmente muchas buenas aplicaciones de Análisis Discriminante en la literatura botánica. Aquí solo nombraremos un par de ejemplos cercanos en la geografía y en el tiempo. Villalobos et al. (2019) realizaron un estudio morfométrico para delimitar las especies de *Anthoxanthum* nativas de Chile; utilizaron 261 especímenes de herbario en los que midieron caracteres morfológicos y anatómicos, junto con el estudio micromorfológico de la epidermis abaxial de la lema y la epidermis adaxial de las hojas vistas bajo microscopio electrónico de barrido. Para delimitar los taxones, se realizó un análisis discriminante de 16 variables (2 cualitativas y 14 cuantitativas). Previamente, los análisis de coordenadas principales y de conglomerados permitieron definir claramente a *A. juncifolium* y *A. pusillum*, que fueron excluidas de los análisis posteriores, de manera de centrar el estudio en las especies del complejo *A. redolens*: *A. altissimum*, *A. gunckelii*, *A. redolens* y *A. utriculatum*.

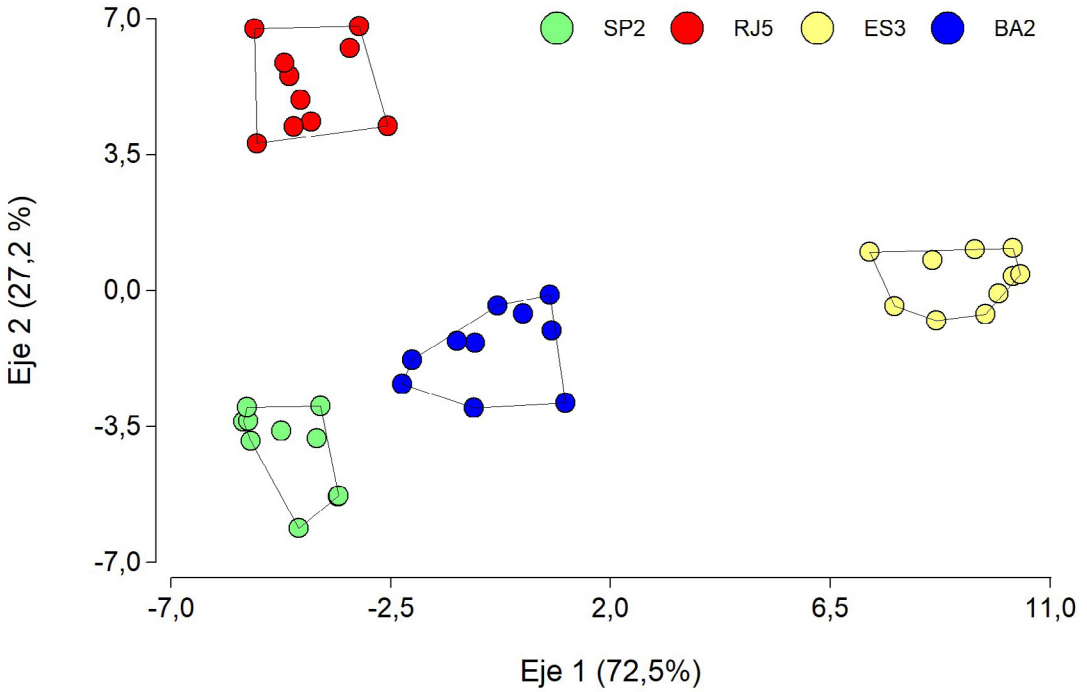

Fig. 6. Ordenamiento de los individuos de *Vriesea procera* (Bromeliaceae) en el espacio determinado por las dos primeras funciones discriminantes. Los colores indican localidades diferentes en la Mata Atlántica: **ES3,** Itaguacú; **BA2,** Itamarajú; **SP2,** Ubatuba; **RJ5,** Angra do Reis. Esta variable categórica denominada "población" en la Tabla 1, participa activamente en el análisis discriminante. Figura en color en la versión en línea https://www.ojs.darwin.edu.ar/index.php/darwiniana/article/view/1086/1299

Caja 2. Medidas de Distancia. El mundo de las medidas de distancia es muy vasto y no pretendemos recorrerlo todo en este breve cuadro, solo presentamos algunos ejemplos vinculados con los métodos aquí descriptos. Los coeficientes de distancia o de similitud cuantifican respectivamente las diferencias o semejanzas entre casos. Incluso, a partir de medidas de similitud se pueden calcular medidas de distancia. A la hora de elegir la medida de distancia adecuada, ésta debe representar lo mejor posible las diferencias biológicas que se desean estudiar. La elección se fundamenta en dos aspectos principales: 1) los tipos de variables medidas y 2) las características de la estructura subyacente que deseamos representar en relación con nuestra pregunta o hipótesis.

Medidas de distancia	Tipo de variable	Características
Euclídea	Cuantitativas, continuas o discretas.	Escalas similares, sin dobles ceros, simétricas
Mahalanobis	Cuantitativas, continuas o discretas	Independiente de la escala, penaliza las variables con alta correlación y varianza
Chi-cuadrado	Binarias o frecuencias	Penaliza variables de mayor magnitud
Bray-Curtis	Binarias o frecuencias	Ignora dobles 0
Gower	Cuantitativas continuas, discretas y binarias	Combina variables continuas y categóricas

Se llevaron a cabo distintos análisis discriminantes, combinando distintos grupos a priori, con el fin de evaluar la delimitación de las especies del complejo *A. redolens* y evaluar la validez de los caracteres tradicionalmente utilizados para delimitar los taxones. Se observó que los caracteres utilizados para discriminar permanecieron constantes a lo largo del perfil latitudinal y altitudinal en los especímenes estudiados. De este modo, se descartaron factores ambientales, asociados a la latitud y/o altitud, que pudieran influir en su variabilidad. El segundo ejemplo es el trabajo ya citado más arriba de Ferrero et al. (2020) que evalúa la delimitación intraespecífica de *Deyeuxia.* Las autoras, después del escalamiento ya comentado, conducen un DA con variables cuantitativas no correlacionadas y obtienen un valor p del eje, la tasa de error en la clasificación y las variables asociadas al eje discriminante.

Aglomeración jerárquica

(Análisis de Conglomerados o *Clusters*)

El propósito de este método es descubrir grupos de objetos semejantes en relación con el conjunto de variables que se les midieron. A diferencia del análisis discriminante, donde los grupos están definidos de antemano, con este método se los construye agrupando los objetos. El análisis se realiza con alguna de las medidas de distancia (Caja 2) combinada con uno de los distintos algoritmos de fusión (Legendre & Legendre, 2003; Husson & Pages, 2017; Palacio et al., 2020). Se comienza agrupando objetos y luego grupos de objetos hasta que todos quedan finalmente agrupados formando una jerarquía. La representación habitual de los resultados de la aglomeración jerárquica es un dendrograma o gráfico de árbol que muestra la secuencia de fusiones en una escala creciente de heterogeneidad interna de los grupos, hasta abarcar al conjunto total que sería el tronco del árbol en la representación (Fig. 7). El investigador elige la medida de distancia (ver Caja 2), el criterio (algoritmo) de fusión y el nivel de heterogeneidad donde frenar la clasificación, esto último es equivalente a decidir número de grupos que serán retenidos e interpretados. Entre la cantidad de algoritmos de fusión disponibles en la bibliografía, el método de Ward es muchas veces preferido por su propiedad de minimizar en cada paso de fusión el incremento de la suma de las varianzas totales de los grupos, es decir que busca formar grupos con gran homogeneidad interna.

En el siguiente dendrograma (Fig. 7) se muestra como ejemplo el resultado de agrupar las especies (filas) de la Tabla 6 (adaptada de Puhl et al., 2014), en grupos florísticos de ocurrencia conjunta. Se trata del mismo conjunto de datos con el que realizamos anteriormente el ordenamiento de censos con Escalamiento Multidimensional. Aquí las especies se clasificaron mediante la distancia de Bray-Curtis y el algoritmo de fusión de Ward.

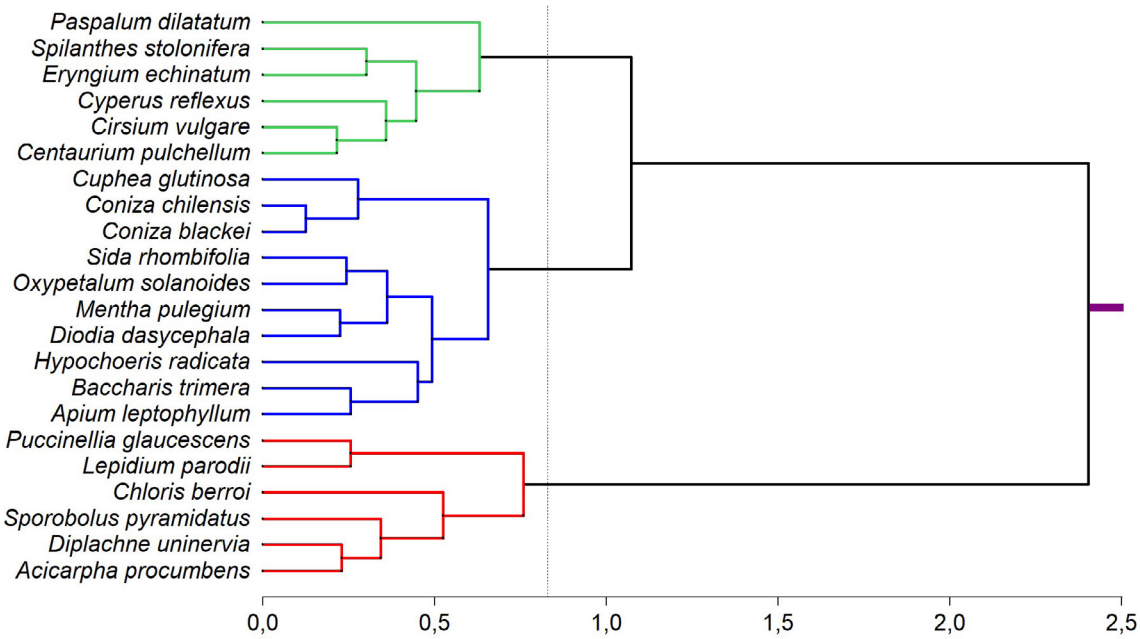

Fig. 7. Dendrograma resultante de la clasificación de las especies de la Tabla 6 con distancia de Bray-Curtis y criterio de aglomeración de Ward. La línea vertical discontinua indica el punto de corte de la clasificación que retiene tres grupos de especies, identificados con diferentes colores en el gráfico. Figura en color en la versión en línea https://www.ojs.darwin.edu.ar/index.php/darwiniana/article/view/1086/1299

La distancia entre objetos o grupos y por lo tanto el nivel de heterogeneidad en el que ocurren las fusiones, crece en el gráfico de izquierda a derecha. La línea vertical discontinua indica el punto de corte de la clasificación que retiene tres grupos de especies, identificados con diferentes colores en el gráfico. Se observa que los grupos verde (de *Paspalum dilatatum*) y azul (de *Cuphea glutinosa*) son los más semejantes entre sí ya que se unen primero entre ellos, antes de fusionarse con el grupo rojo (de *Puccinellia glaucescens*) que es el más diferente respecto a los demás ya que solo se une con ellos al final de la jerarquía. El dendrograma podría girar sus ramas como si fuera un colgante sostenido desde el tronco (segmento coloreado a la derecha del gráfico) y seguiría representando idéntica configuración, porque son importantes las uniones entre las ramas y su nivel en la jerarquía, pero no la ubicación relativa de las ramas del mismo nivel. Entonces, el grupo azul y el verde pueden girar e intercambiar posiciones adentro del grupo mayor que los une, porque la representación de la jerarquía no informa si alguno de ellos es más cercano al grupo rojo que el otro. Es decir que la ubicación de los sub-grupos dentro de una llave (grupo) dada es irrelevante porque no sabemos las distancias entre estos grupos y el resto del dendrograma, solo sabemos la distancia del grupo completo a su grupo vecino en la jerarquía.

El análisis de conglomerados responde a nuestra indicación de conformar grupos y siempre los encuentra, más allá de que existan o no agrupaciones naturales en el conjunto de objetos bajo estudio. Por eso es bueno completar el análisis con una prueba de diferencias entre grupos (como *Multi Response Permutation Procedure*, MRPP; Zimmerman et al., 1985; Pillar, 2013) y/o tablas resumen de la información que muestren las características de los grupos, como esta que presentamos a continuación (Tabla 11). También se podría completar la interpretación de las relaciones entre grupos con un gráfico de ordenamiento. En nuestro ejemplo, el dendrograma (Fig. 7) junto con la tabla de promedios (Tabla 11) permiten identificar e

interpretar los grupos de especies de respuesta similar a las condiciones de los sitios relevados. El dendrograma muestra que la diferencia más importante ocurre entre el grupo rojo y los otros dos. La razón de este agrupamiento, como se observa en la tabla, es que las especies del grupo rojo solo están presentes en los ambientes de suelos halomórficos (H0 y H35) y están ausentes en los ambientes de suelos profundos en ambos momentos de muestreo (A0 y A35). Las diferencias entre las especies de los otros dos grupos son más sutiles. Si bien todas están presentes en la comunidad A35, las del grupo verde pueden también estar presentes en las comunidades H0 y H35 y además presentan promedios de cobertura más altos en el tiempo cero de muestreo (A0) que en el posterior (A35). En cambio, las especies del grupo azul no aparecen nunca en las comunidades de suelos halomórficos y pueden presentar coberturas más altas en el segundo momento de muestreo en la comunidad de suelos profundos (A35).

Los métodos jerárquicos han sido muy utilizados por taxónomos y ecólogos principalmente por dos razones. La primera es el fuerte paralelismo con las teorías evolutivas (aunque las medidas de distancia presentadas no tengan interpretación evolutiva; recordar que hay métodos específicos para ello). La segunda es su capacidad para organizar el conocimiento, ya que la jerarquización de los grupos permite establecer las relaciones de similitud que existen entre ellos. Algunos ejemplos interesantes son los trabajos que ya fueron citados más arriba en el contexto del escalamiento multidimensional, donde representaron los resultados del agrupamiento con análisis de conglomerados, como Ulloa et al. (2011), que utilizó la distancia de Gower (ya que utilizó simultáneamente variables cuantitativas y cualitativas), el algoritmo de fusión de Ward para armar grupos cuyo objetivo era clarificar la taxonomía en *Polypogon* a través de detectar nuevas características de la morfología de la lema. También Zallochi et al. (1992), para diferenciar taxones del género *Phaseolae* con datos cromatográficos construye grupos mediante un método de agrupamiento jerárquico con la distancia de Manhattan. Batista et al. (2014), para caracterizar comunidades vegetales en el Parque

Tabla 11. Promedios de cobertura de las especies en las comunidades del pastizal de la Pampa Deprimida. **A0,** pastizales en suelos profundos bien drenados (Pradera de Mesófitas), censos de vegetación en el tiempo cero; **A35,** mismos ambientes censados 35 años después; **H0,** pastizales en suelos halomórficos (Estepa de Halófitas), censos de vegetación en el tiempo cero; **H35,** mismos ambientes censados 35 años después. Los grupos de especies corresponden a la clasificación presentada en la Fig. 7. Tabla en color en la versión en línea https://www.ojs.darwin.edu.ar/index.php/darwiniana/article/view/1086/1299

	Especie	A0	A35	H0	H35
	Paspalum dilatatum	2,00	2,00	0	0
	Spilanthes stolonifera	0,30	0,18	0,40	0,16
	Eryngium echinatum	0,90	0,24	0,40	0,12
	Cyperus reflexus	0,60	0,34	0,20	0,16
	Cirsium vulgare	0,70	0,42	0	0
	Centaurium pulchellum	0,80	0,34	0	0,16
	Cuphea glutinosa	0,50	0,24	0	0
	Coniza chilensis	0,40	0,24	0	0
	Coniza blackei	0,10	0,22	0	0
	Sida rhombifolia	0,40	0,14	0	0
	Oxypetalum solanoides	0	0,16	0	0
	Mentha pulegium	0,20	0,52	0	0
	Diodia dasycephala	0,30	0,40	0	0
	Hypochoeris radicata	0	0,92	0	0
	Baccharis trimera	0	0,80	0	0
	Apium leptophyllum	0,60	0,60	0	0
	Puccinellia glaucescens	0	0	0	0,52
	Lepidium parodii	0	0	0,10	0,50
	Chloris berroi	0	0	0,60	0,14
	Sporobolus pyramidatus	0	0	1,80	1,50
	Diplachne uninervia	0	0	0,80	0,50
	Acicarpha procumbens	0	0	0,80	0,90

Nacional El Palmar realizaron una clasificación preliminar de los censos mediante un método de aglomeración jerárquico con la distancia Euclídea y con el criterio de fusión de Ward (mínimo incremento de la suma de las varianzas totales de los grupos en cada paso). Esta clasificación jerárquica produjo un dendrograma donde identificaron siete grupos de censos. Luego, reclasificaron los censos en siete grupos mediante un procedimiento no jerárquico de k-medias para minimizar la suma de las varianzas totales de los k grupos (MacQueen, 1967), utilizando como clasificación inicial, o semilla, los centroides de los grupos obtenidos con el método jerárquico. Obtuvieron así una clasificación definitiva de los censos en 7 comunidades combinando ventajas de ambos métodos de clasificación.

Comentarios finales

Hasta aquí estuvimos revisando qué es lo que permiten hacer estos métodos de análisis multivariado para poder elegir mejor cuál aplicar en cada circunstancia e interpretar correctamente los resultados que producen, así se trate de la conducción de nuestros propios trabajos o de la comprensión de los trabajos publicados por otros científicos. En ese sentido, también vimos que debemos poner atención en diferenciar los resultados genuinos del análisis respecto de la información suplementaria que a veces se presenta sobre los mismos gráficos. Encontramos publicados ejemplos interesantes donde los resultados del análisis multivariado sirven para otros análisis, ya sean ejes de ordenamientos que funcionan como variables predictoras de modelos posteriores o bien como variables respuesta de síntesis en otros modelos. También vimos en los trabajos publicados combinaciones de diferentes métodos aplicados al mismo conjunto de datos.

Los análisis estadísticos que presentamos fueron realizados con el paquete InfoStat (Di Rienzo et al., 2020), que posee como características distintivas la gran facilidad de uso que presenta su interfase para el usuario ajeno a la informática y la buena calidad de representación de los resultados. Los mismos análisis, así como otros de mayor complejidad y muy buenas combinaciones de métodos y excelentes representaciones están disponibles para usuarios que se sientan cómodos en el ambiente R. El ambiente R es un lenguaje de programación que la comunidad científica adoptó como herramienta computacional de vanguardia en las últimas décadas, www.r-project.org. En particular encontrarán en R el paquete *FactoMineR* (Lê et al., 2008; Husson & Pages, 2017) que se puede complementar también con algunas aplicaciones del paquete *Vegan* (Oksanen et al., 2018).

Como consideración final, queremos subrayar la importancia de alimentar a los métodos de análisis con buenos datos: primero formular claramente las preguntas que deseamos responder acerca del sistema bajo estudio y discutirlas en el grupo de trabajo para guiar la elección de las variables relevantes y la identificación de los casos donde van a ser medidas. Los mejores métodos de análisis multivariado no encuentran los patrones que buscamos, aunque existan en la naturaleza, si no los proveemos de buena información y si no abarcamos la variabilidad necesaria en el material recolectado. En ocasiones es preferible implementar algún tipo de muestreo estratificado para asegurar que todas las situaciones (o las que se justifiquen como importantes) estén representadas en la matriz de datos, independientemente de su abundancia en el sistema estudiado. Por otra parte, también puede ocurrir lo contrario y conviene estar atentos, ya que podemos encontrar patrones inesperados en bases de datos que estaban bien estructuradas para responder otras preguntas de interés (Perelman et al., 2003b).

Con la sola excepción del Análisis Discriminante, todas las técnicas presentadas en este trabajo son descriptivas, ya que no involucran inferencia estadística desde la muestra hacia una población de referencia. En estos proyectos, la base de datos que tenemos entre manos constituye el universo bajo estudio. En general, son bases de datos muy grandes que justifican esta afirmación y cuya sola descripción aporta importante conocimiento nuevo. Es por eso que no tienen requerimientos particulares de muestreo aleatorio ni de cumplimiento de supuestos, salvo para el Análisis Discriminante lineal, como se comentó antes, y sobre todo cuando la aplicación de este análisis involucra un proceso de inferencia al asignar nuevos individuos a los grupos preexistentes. En circunstancias diferentes, cuando lo que interesa es la comparación entre grupos, sí es necesario encarar un muestreo aleatorio, pero si se trata de trabajos descriptivos es más eficiente un muestreo dirigido a abarcar la necesaria heterogeneidad para que los patrones buscados puedan manifestarse y también dirigido a evitar esas situaciones muy atípicas que cuando aparecen inevitablemente despiertan la curiosidad de todo naturalista. Los casos raros (*outliers*) provocan importantes distorsiones en los análisis multivariados, por eso se recomienda que ante la curiosidad que despiertan en medio de la campaña, queden registrados en la libreta de campo como disparador de un próximo proyecto: solicitar financiamiento para buscar en otra campaña varios casos más así de raros y estudiarlos para entender esa rareza.

AGRADECIMIENTOS

Agradecemos los comentarios de tres revisores anónimos que contribuyeron a mejorar este trabajo.

BIBLIOGRAFÍA

Batista, W.; A. Rolhauser, F. Biganzoli, S. Burkart, L. Goveto, A. Maranta, A. Pignataro, N. Morandeira & M. Rabadán. 2014. Las comunidades vegetales de la sabana del parque nacional El Palmar (Argentina). *Darwiniana*, nueva serie 2(1): 5-38. DOI: https://doi.org/10.14522/darwiniana.2014.21.569

Bonasora, M. G., M. T. Pozzobon, A. I. Honfi & G. H. Rua. 2015. *Paspalum schesslii* (Poaceae, Paspaleae), a new species from Mato Grosso (Brazil) with an unusual base chromosome number. *Plant Systematics and Evolution* 301: 2325-2339.

Burkart, S. E.; R. J. C. León, S. B. Perelman & M. Agnusdei. 1998. The grasslands of the flooding pampa (Argentina): Floristic heterogeneity of natural communities of the southern Rio Salado basin. *Coenoses* 13: 17-27.

Chaneton, E. J.; S. B. Perelman, M. Omacini & R. J. C. León. 2002. Grazing, Environmental Heterogeneity, and Alien Plant Invasions in Temperate Pampa grasslands. *Biological Invasions* 4: 7-24.

Chase, J. M. 2010. Stochastic community assembly causes higher biodiversity in more productive environments. *Science* 5984: 1388-1391.

Christodoulou, M. D.; J. Y. Clark & A. Culham. 2020. The Cinderella discipline: morphometrics and their use in botanical classification. *Botanical Journal of the Linnean Society* 194: 385-396.

Cué-Hernández, K. A., Gil-Muñoz, A., Aguirre-Jaimes, A., López P.A., & Rey Taboada-Gaytán, O. (2022). Floral visitors in the crop Phaseolus coccineus (Fabaceae) on the Altiplano of Puebla, Mexico: importance of agricultural management and flower color. *Acta Botánica Mexicana* 128: e1924. DOI: https://doi.org/10.21829/abm129.2022.2054

Digby, P. G. N. & R. A. Kempton. 1987. Population and Community Biology Series: Multivariate Analysis of Ecological Communities. Chapman and Hall, London.

Di Rienzo J. A.; F. Casanoves, M. G. Balzarini, L. Gonzalez, M. Tablada & C. W Robledo. InfoStat versión 2020. Centro de Transferencia InfoStat, FCA, Universidad Nacional de Córdoba, Argentina. https://www.infostat.com.ar

Ferrero, M. A.; A. B. Sassone, L. Giussani & A. S. Vega. 2020. Distribution models and morphometric analyses as additional tools for the study of diversification in *Deyeuxia velutina*, an Andean grass species. *Darwiniana*, nueva serie 8(2): 509-524. DOI: https://doi.org/10.14522/darwiniana.2020.82.894

Giussani, L. M.; E. G. Nicora & F. A. Roig. 2000. *Poa durifolia* and its relationship with the phenetic pattern of *Poa* section Dioicopoa (Poaceae). *Darwiniana* 38(1-2): 47-57. DOI: https://doi.org/10.14522/darwiniana.2000.381-2.161

Greenacre, M. 2013. The contributions of rare objects in correspondence analysis. *Ecology* 94: 241- 249.

Hill, M. O. & H. G. Gauch, Jr. 1980. Detrended Correspondence Analysis: an improved ordination technique. *Vegetatio* 42: 47-58.

Husson, F., S. Lê & J. Pagès. 2017. Exploratory Multivariate Analysis by Example Using R. 2nd edition. Chapman & Hall/CRC Computer Science & Data Analysis.

Jackson, D. A. & K. M. Somers. 1991. Putting things in order: the ups and downs of detrended correspondence analysis. *American Naturalist* 137: 704-712.

Kenkel, N. C.; D. A. Derksen, A. G. Thomas & P. Watson. 2002. Review: Multivariate analysis in weed science research. *Weed Science* 50: 281-292.

Lê, S.; J. Josse & F. Husson. 2008. FactoMineR: an R package for multivariate analysis. *Journal of Statistical Software* 25(1): 1-18.

Legendre, P. & L. Legendre. 2003. *Numerical Ecology*, 3rd English edition. Amsterdam: Elsevier, 853p.

MacQueen, J. B. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1: 281-297.

Nagahama, N.; A. M. Anton, M. I. Hidalgo & G. A. Norrmann, 2012. Naming hybrids in the *Andropogon lateralis* complex (Poaceae, Andropogoneae) after multivariate analyses. *Darwiniana* 50(1): 114-123. DOI: https://doi.org/10.14522/darwiniana.2012.501.436

Oksanen J. F.; G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs & H. Wagner. 2018. Vegan: community ecology package. Paquete de R versión 2.5-3. https://CRAN.R-project.org/package=vegan

Palacio, F.; M. Apodaca & J. V. Crisci. 2020. Análisis multivariado para datos biológicos: Teoría y su aplicación utilizando el lenguaje R. Fundación Félix de Azara; pp. 265.

Perelman, S.; R. J. C. León & M. Oesterheld. 2001. Cross-scale vegetation patterns of Flooding Pampa grasslands. *Journal of Ecology* 89(4): 562-577.

Perelman, S. B.; S. E. Burkart & R. J. C. León. 2003a. The role of a native tussock-grass *(Paspalum quadrifarium)* in structuring plant communities in the Flooding Pampa grasslands, Argentina. *Biodiversity and Conservation* 12: 225-238.

Perelman, S. B.; M. A. Mazzella, J. Muschietti, T. Zhu & J. J. Casal. 2003b. Finding unexpected patterns in microarray data. *Plant Physiology* 133(4): 1717-1725.

Pillar, V. P. 2013. How accurate and powerful are randomization tests in multivariate analysis of variance? *Community ecology* 14(2): 153-163.

Puhl, L. E.; S. B. Perelman, W. B. Batista, S. E. Burkart & R. J. C. León. 2014. Local and regional long-term diversity changes and biotic homogenization in two temperate grasslands. *Journal of Vegetation Science* 25: 1278-1288.

R Core Team. 2018. R: a language and environment for statistical computing. https://www.r-project.org/

Uchida, K. & A. Ushimaru. 2015. Land abandonment and intensification diminish spatial and temporal beta-diversity of grassland plants and herbivorous insects within paddy terraces. *Journal of Applied Ecology* 52: 1033-1043.

Ulloa, W.; C. M. Baeza, V. L. Finot, A. Marticorena & E. Ruiz. 2011. Micromorfología de la lemma de los géneros *Polypogon*, x *Agropogon* y *Agrostis* (Poaceae) en Chile. *Journal of the Botanical Research Institute of Texas* 5(1): 237-253.

Uribbe, F. P.; B. Neves, S. S. A. Jacques & A. F. Costa. 2020. Morphological Variation in the *Vriesea procera* complex (Bromeliaceae, Tillandsioideae) in the Brazilian Atlantic Rainforest, with Recognition of New Taxa. *Systematic Botany* 45(1): 53-58.

Villalobos, N. I.; V. L. Finot, E. Ruiz, P. Peñailillo & G. Collado. 2019. Morphometric and taxonomic study of the native Chilean species of genus Anthoxanthum (Poaceae, Pooideae, Poeae, Anthoxanthinae). *Darwiniana*, nueva serie 7(1): 93-136. DOI: https://doi.org/10.14522/darwiniana.2019.71.822

Yansen, M. V. & F. Biganzoli. 2022. Las especies arbóreas exóticas en Argentina: caracterización e identificación de las especies actual y potencialmente problemáticas. *Darwiniana*, nueva serie 10(1): 80-97. DOI: https://doi.org/10.14522/darwiniana.2022.101.1001

Zapater, M. A.; P. S. Hoc, E. C. Lozano & S. S. Sühring, 2014. Delimitación de las especies argentinas del género *Inga* (Mimosoideae) mediante técnicas numéricas. *Darwiniana*, nueva serie 2(2): 248-259. DOI: https://doi.org/10.14522/darwiniana.2014.22.614

Zallocchi, E. M.; A. B. Pomilio & R. A. Palacios. 1992. Chemotaxonomical study of the subtribe Phaseolinae (Phaseoleae-Papilionoideae-Leguminosae) I: Chromatography of flavonoids from argentinian species of the genus Macroptilium*, Darwiniana* 31: 299-313.

Zimmerman, G. M.; H. Goetz & P.W. Mielke Jr. 1985. Use of an improved statistical method for group comparisons to study effects of prairie fire. *Ecology* 66: 606-611.