

ANÁLISIS MULTIVARIADO CON VARIABLES MIXTAS EN EL GERMOPLASMA DE ÑAME

Osmany Molina Concepción*, Marilys Milián Jiménez y Carmen C. Pons Pérez

Instituto de Investigaciones de Viandas Tropicales (INIVIT), Apartado 6, Santo Domingo, CP: 53 000, Villa Clara, Cuba.

* Autor para la correspondencia: taxonumeric@inivit.cu.

RESUMEN

Los métodos de clasificación son técnicas multivariadas comúnmente usadas en áreas tales como, la taxonomía numérica y el análisis genético para describir y analizar conjunto de datos. El presente trabajo tiene como objetivo clasificar el germoplasma de ñame (*Dioscorea* spp.) que se conserva en el INIVIT, a partir de los rasgos que los caracterizan, combinando variables cualitativas y cuantitativas. Se realizó un estudio con la función *dudi.mix* que permite trabajar con variables mixtas y reducir la dimensionalidad de la matriz de datos, como resultado se obtiene un menor número de variables, a esta matriz se le aplicó la distancia Euclidiana y posteriormente se usó un análisis de conglomerado jerárquico, con los métodos de agrupamiento de Ward y UPGMA. El desempeño de los métodos de aglomeración fue evaluado con el test de Mantel. Para los diferentes análisis se utilizaron funciones implementadas sobre la base del lenguaje de programación R. En esta investigación se determinó que ambos métodos mantienen una buena estabilidad en las estructuras obtenidas.

Palabras clave: métodos de agrupamiento, programación en R, taxonomía numérica, test de Mantel.

MULTIVARIATE ANALYSIS WITH MIXED VARIABLES IN THE YAM GERMPLASM

ABSTRACT

Classification methods are multivariate techniques commonly used in areas such as, the numerical taxonomy and the genetic analysis to describe and analyze data sets. The objective of this work is to classify the yam germplasm (*Dioscorea* spp.) that is conserved at the INIVIT, from the features that characterize it, combining qualitative and quantitative variables. A study was carried out with the *dudi.mix* function that allows to work with mixed variables and reduce the dimensionality of the data matrix, as a result, a smaller number of variables is obtained, this matrix was applied the Euclidean distance measure, and subsequently it was used a hierarchical conglomerate analysis with Ward's and UPGMA grouping methods. The performance of the agglomeration methods was evaluated with the Mantel's test. For the different analyses, functions implemented on the basis of the R programming language were used. Through this research, it was determined that both methods maintain a good stability in the structures obtained.

Keywords: grouping methods, R programming, numerical taxonomy, Mantel's test.

INTRODUCCIÓN

Con el uso de las computadoras, la taxonomía numérica, definida por Sokal y Sneath (1963) como la agrupación de unidades taxonómicas por métodos numéricos alcanza un importante crecimiento que ha permitido el uso de métodos estadísticos multivariados para la clasificación de los recursos genéticos.

Los métodos de clasificación (agrupamiento de entidades con similares patrones) y ordenamiento (descripción de la relación espacial entre entidades) son dos de las mejores técnicas multivariadas comúnmente usadas en áreas tales como la taxonomía numérica, análisis genético, cultivos de planta y biotecnología para describir y analizar datos multivariados. El análisis de patrones, que es el uso combinado del análisis de cluster y técnicas de ordenamiento, brinda una poderosa herramienta para examinar grandes conjuntos de datos.

Variables continuas y categóricas son evaluadas en cada accesión o cultivares de los bancos de germoplasmas, dificultando la elaboración de escalas numéricas que integren variables continuas, nominales u ordinales. Entre las alternativas metodológicas para abordar este problema, está la combinación de técnicas que permiten el análisis de variables mixtas.

Se han desarrollado en los últimos años diversos algoritmos de clasificación no supervisada capaces de realizar tal tarea: jerárquicos y no jerárquicos, los cuales dan solución a un número considerable de problemas en las ciencias biológicas a pesar de no producir muchas veces clasificaciones objetivas ni estables, ya que se pueden obtener diferentes agrupamientos a partir de la misma matriz de datos si se utilizan distintos algoritmos. En el caso de que nuevos caracteres fueran hallados, las distancias cambiarán y paralelamente, cambiará la clasificación previa.

El presente trabajo tiene como objetivo combinar variables cualitativas y cuantitativas para determinar si ofrece una topología consistente al clasificar un grupo de genotipos de ñame (*Dioscorea* spp.) para determinar los grupos de accesiones que tienen un patrón de expresión común.

MATERIALES Y MÉTODOS

Para realizar la investigación, se usaron datos procedentes de un estudio de los diferentes genotipos de la colección de prueba de ñame (*Dioscorea* spp.) (86 accesiones donde se evaluaron 43 variables cualitativas y nueve variables cuantitativas incluidas en el Sistema de Descriptores Mínimos (Sánchez *et al.*, 1995)), que se conserva en el Instituto de Investigaciones de Viandas Tropicales (INIVIT).

Las variables cualitativas (nominales y ordinales) y cuantitativas se integraron con la función *dudi.mix()* en el paquete *Ade4* (Chessel *et al.*, 2004; Dray y Dufour, 2007; Dray *et al.*, 2007; Bougeard y Dray, 2018), tomando las dimensiones que acumularon una varianza superior a uno, posteriormente se obtuvo una matriz con la distancia Euclidiana a través de la función *dist()* implementada en el paquete *stats*. Finalmente se realizó un análisis de conglomerados con los métodos de agrupamiento Ward o de varianza mínima (*Minimum Variance Clustering* (Ward, 1963) en sus dos variantes Ward.D y Ward.D2 y Promedio ó UPGMA (*Unweighted Pair-Group Method using*

Arithmetic Averages) (Sneath y Sokal, 1973) implementados en la función *hclust()* desarrollada en el paquete *stats*.

Por otra parte, los métodos jerárquicos imponen cierta estructura sobre los datos y es necesario con frecuencia considerar si es aceptable o si se introducen distorsiones inaceptables en las relaciones originales. Para verificar este hecho se puede obtener una significación estadística aplicando el test de Mantel (Mantel, 1967), implementado en la función *mantel()* del paquete *vegan* (Oksanen *et al.*, 2011).

Cuando se emplean técnicas de aglomeración jerárquicas, el investigador no siempre está interesado en la jerarquía completa, sino en un subconjunto de particiones obtenidas a partir de ella. Las particiones se obtienen cortando el dendrograma o seleccionando una de las soluciones en la sucesión de conglomerados que comprende la jerarquía.

Por ello, en esta investigación para validar la partición de los datos, obtenida con un algoritmo de agrupamiento, se emplea el índice de Dunn implementado en la función *dunn()* del paquete *clValid* (Brock, 2008).

Para procesar la información se utilizó un lenguaje de programación, orientado a objetos denominado R (R Development Core Team, 2018); el cual es un conjunto de programas integrados para análisis estadísticos y gráficos. R es un *software* libre, por lo cual la implementación de cualquier técnica en este lenguaje le dará mayor potencialidad e independencia.

RESULTADOS Y DISCUSIÓN

Al aplicar la función *dudi.mix()* a la base de datos de variables cualitativas y cuantitativas obtenidas a partir de la caracterización morfoagronómica de la colección de ñame (*Dioscorea* spp.) se seleccionan 29 dimensiones con valores superiores a uno que explica el 98,4 % de la variabilidad.

Al aplicar la distancia Euclidiana sobre la matriz conformada por las 29 dimensiones se obtiene una matriz de distancia, a la cual se le aplicó los métodos de Ward.D, Ward.D2 y promedio.

El test de Mantel para el método de promedio con 0,968 es el mayor, seguido de 0,867 para Ward.D2 y 0,830 Ward.D, lo cual se puede observar en las buenas estructuras de los dendrogramas.

En los dendrogramas de la figura 1, obtenidos por el método de Ward en sus dos variantes se observa que tienen estructuras bien definidas, aunque el método *Ward.D* tiene una mayor contracción sobre el eje x que el método *Ward.D2*. El índice de Dunn con un valor de 0,554 define para ambos métodos un corte en cuatro conglomerados, el grupo uno (G1) tiene las 72 accesiones de la especie *D. alata* (ALA). El grupo dos (G2) lo componen dos de la especie *D. esculenta* (ESC), el grupo tres (G3) cinco de la especie *D. cayenensis* (CAY) y tres de la especie *D. rotundata* (RTA), el grupo cuatro (G4) tres de la especie *D. bulbifera* (BUL).

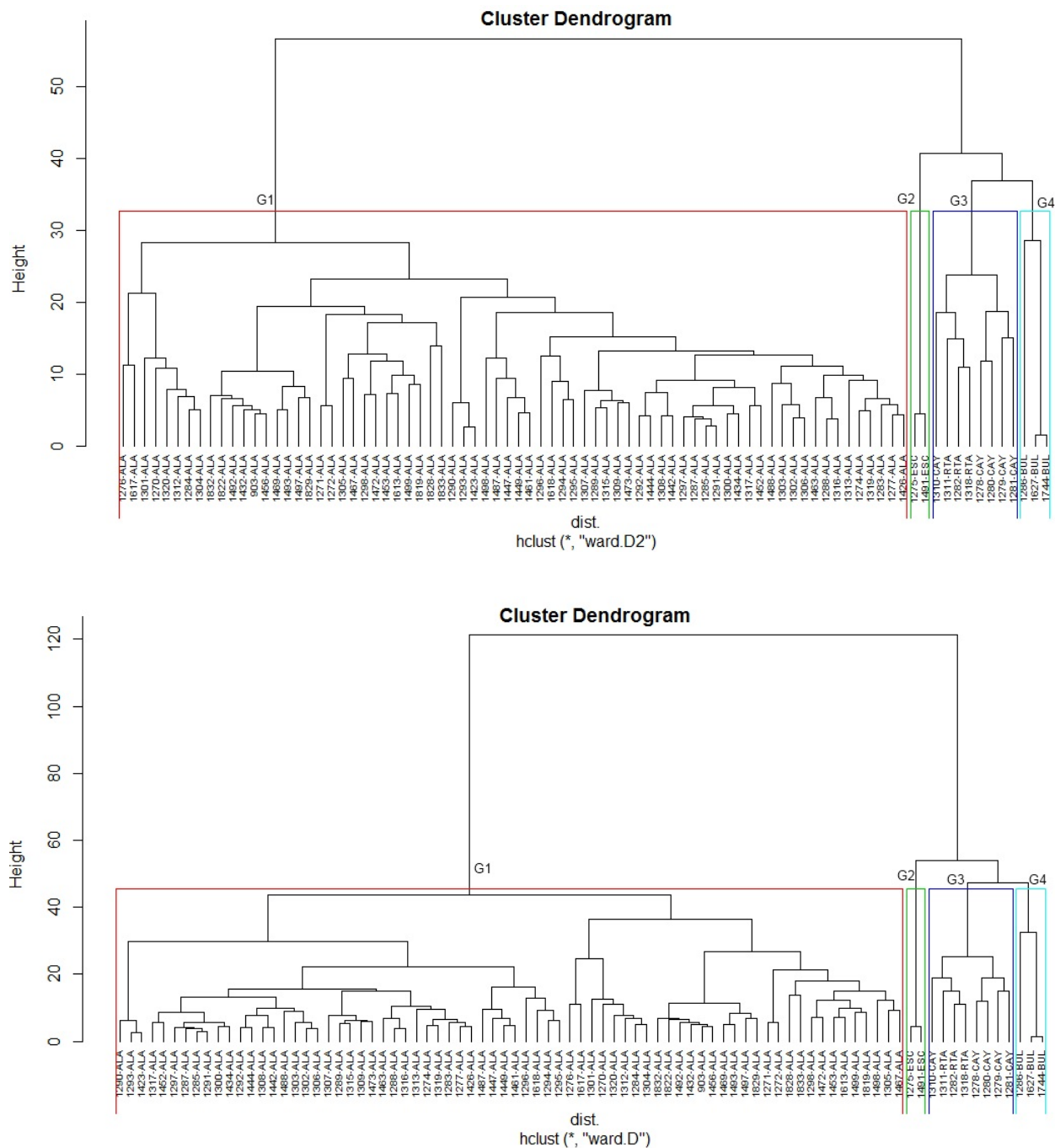


Figura 1. Dendrogramas del método de agrupamiento Ward variante Ward.D2 y Ward.D. Colección de ñame (*Dioscorea* spp.).

El dendrograma obtenido por el método promedio (Figura 2) tiene una estructura bien definida. Para un índice de Dunn de 0,877 se establecen tres grupos lo cual no describe la estructura de la base de datos en estudio; a criterio del especialista del cultivo el mejor corte se realiza para seis grupos, el grupo uno (G1) está formado por dos de la especie *D. esculenta*, el grupo dos (G2) una de la especie *D. bulbifera*, el grupo tres (G3) dos de la especie *D. bulbifera*, grupo cuatro (G4) tiene las 72 accesiones de la

especie *D. alata* (ALA), el grupo cinco (G5) cuatro de la especie *D. cayenensis* y el grupo seis (G6) una de la especie de la especie *D. cayenensis*, tres de la especie *D. rotundata*.

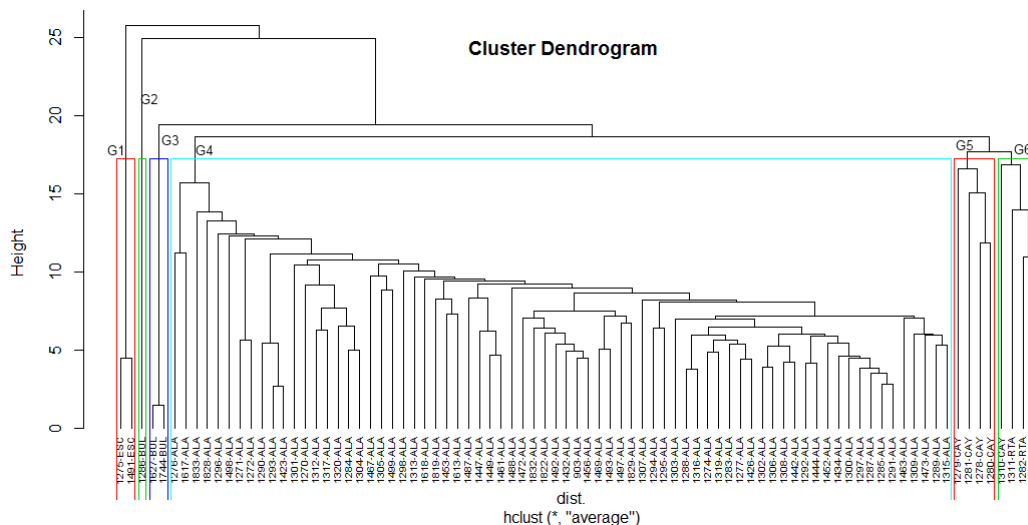


Figura 2. Dendrograma del método de agrupamiento Promedio. Colección de ñame (*Dioscorea* spp.).

Comparar los méritos relativos de cada método es difícil, pues cuando aplicamos algoritmos de agrupamiento a un conjunto de datos, los resultados difieren.

A pesar de las limitaciones del análisis de conglomerados, esta es una técnica ampliamente utilizada en muchas ramas de la ciencia por ser eminentemente exploratoria, puesto que la mayor parte de las veces, no utiliza ningún tipo de modelo estadístico para llevar a cabo el proceso de clasificación. Estos métodos son muy adecuados para extraer información de un conjunto de datos sin imponer restricciones previas en forma de modelos estadísticos, al menos de forma explícita y, por ello, puede llegar a ser muy útil como una herramienta de elaboración de hipótesis acerca del problema considerado, sin imponer patrones o teorías previamente establecidas.

Estudios sobre integración de variables cualitativas y cuantitativas se pueden revisar en trabajos realizados sobre bancos de germoplasma vegetal en Molina *et. al* (2013a, 2013b, 2017 y 2018).

CONCLUSIONES

La obtención de dendrogramas por los métodos de Ward y Promedio a partir de la integración de variables mixta a través de la función *dist.mix* logran describir la estructura de la base de datos de ñame (*Dioscorea* spp.) en estudio.

BIBLIOGRAFÍA

BOUGEARD, S. and S. DRAY. 2018. Supervised Multiblock Analysis in R with the ade4 Package. *Journal of Statistical Software*, 86(1):1–17. doi: 10.18637/jss.v086.i01.

- BROCK, G.; V. PIHUR and S. DATTA. 2008. clValid: An R Package for Cluster Validation. *Journal of Statistical Software*, 25(4):1-22. URL <http://www.iostatsoft.org/v25/i04/>.
- CHESSEL, D.; A. DUFOUR and J. THIOULOUSE. 2004. The ade4 Package – I: One-Table Methods. *R News*, 4(1):5–10. <https://cran.r-project.org/doc/Rnews/>.
- DRAY, S. and A. DUFOUR. 2007. The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, 22(4):1–20. DOI: 10.18637/jss.v022.i04.
- DRAY, S.; A. DUFOUR and D. CHESSEL. 2007. The ade4 Package – II: Two-Table and K-Table Methods. *R News*, 7(2):47–52. <https://cran.r-project.org/doc/Rnews/>.
- MANTEL, N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2):209-220.
- OKSANEN, J.; F. G. BLANCHET; R. KINDT; P. LEGENDRE; R. B. O'HARA; G. L. SIMPSON; P. SOLYMOS; M. H. H. STEVENS and H. WAGNER. 2011. vegan: Community Ecology Package. R package version 1.17-9.
- R DEVELOPMENT CORE TEAM. 2018. R: A Language and Environment for Statistical Computing (version 3.5.1). R Foundation for Statistical Computing. Vienna, Austria. Fecha de consulta: 7 septiembre de 2018, Disponible en: <http://www.r-project.org/>.
- SÁNCHEZ, I.; M.D. MILIÁN; A. RAYAS y S. RODRÍGUEZ. 1995. Lista de descriptores y caracterización de la colección cubana de ñame (*Dioscorea* spp). Instituto de Investigaciones de Viandas Tropicales (INIVIT), Santo Domingo, Villa Clara, Cuba.
- SOKAL, R.R. and P.H.A. SNEATH. 1963. Principles of numerical taxonomy San Francisco & London: Freeman and Cía.
- MOLINA O.; R. GARCÍA; M. MILIÁN; L. GONZÁLEZ; C. PONS y R. GRAU. 2013a. Clasificación y determinación del número óptimo de conglomerados en bancos de germoplasma. *Centro Agrícola* 40(4):19-24.
- MOLINA O.; R. GARCÍA; M. MILIÁN; L. GONZÁLEZ; C. PONS y R. GRAU. 2013b. Análisis taxonómico con variables mixtas en accesiones de malanga (*Xanthosoma* spp.) y plátano (*Musa* spp.). *Centro Agrícola* 40(4):7-10; octubre-diciembre, 2013.
- MOLINA O.; R. GARCÍA; M. MILIÁN; L. GONZÁLEZ; C. PONS y R. GRAU. 2017. Taxobanger v.1.0: Aplicación informática en R para el análisis taxonómico en bancos de germoplasma vegetal. *Revista Cubana de Ciencias Informáticas*, Vol. 11(3): 159-173. <http://rcci.uci.cu>.
- MOLINA O.; M. MILIÁN; L. GONZÁLEZ y C. PONS. 2018. Análisis de variables cuantitativas en bancos de germoplasma vegetal. *Agricultura Tropical*, 4(2): 58-64. <http://ojs.inivit.cu>.
- WARD, J.H., JR. 1963. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58:236-244.