Centro Agrícola, 40(4): 7-10; octubre-diciembre, 2013
ISSN papel: 0253-5785 ISSN on line: 2072-2001
CE: 39,13 CF: cag034131939

EDITORIAL Felgo

ARTICULO GENERAL

Análisis taxonómico con variables mixtas en líneas de malanga (*Xanthosoma* spp.) y plátano (*Musa* spp.)

Taxonomic analysis with mixed variables in New cocoyam (*Xanthosoma* spp.) accessions and Banana and Plantain (*Musa* spp.)

Osmany Molina Concepción[1], Raisa García Rodríguez[1], Marilys Milián Jimenez[1], Lianet González Díaz[1], Carmen C. Pons Pérez[1] y Ricardo Grau Abalo[2]

[1] Instituto de Investigaciones de Viandas Tropicales (INIVIT), Apdo #6, Santo Domingo, Villa Clara, Cuba.
[2] Universidad Central "Marta Abreu" de Las Villas (UCLV), Carretera a Camajuaní km 5 1/2 , Santa Clara, Villa Clara, Cuba.

E-mail: osmany@inivit.cu

RESUMEN. El presente trabajo tuvo como objetivo comparar el desempeño de cuatro métodos de aglomeración jerárquicos en varias estructuras de datos, para determinar cuál método y topología son más consistentes al clasificar un grupo de genotipos de malanga (*Xanthosoma* spp.) y plátanos (*Musa* spp.) a partir de los rasgos que los caracterizan. En el primer procedimiento se realizó un estudio de la métrica de Gower para variables mixtas la cual se combinó con cuatro métodos de aglomeración y en el segundo se integraron las variables cualitativas transformadas a través del análisis de correspondencia múltiple con las cuantitativas normalizadas, en un análisis de conglomerados usando al igual que en el anterior, cuatro métodos de aglomeración: Promedio, Ward, agrupación de enlace completo y agrupación de enlace simple combinado a la vez con dos medidas de distancias (*Euclidiana y Manhattan*). Las diferentes estructuras fueron evaluadas con el coeficiente de correlación cofenética. Para los diferentes análisis se utilizaron funciones implementadas sobre la base del lenguaje de programación R. En está investigación se determinó que para ambos procedimientos con las medidas de distancias de Gower y Euclidiana el método de aglomeración Promedio es el que mantiene una mejor estabilidad en las estructuras obtenidas.

Palabras clave: conglomerados, aglomeración jerárquica, bancos de germoplasma, taxonomía numérica.

ABSTRACT. This study aims to compare the performance of four hierarchical clustering methods on several data structures to determine which method and topology is more consistent in classifying a taro (*Xanthosoma* spp.) genotype group and plantains (*Musa* spp.) from their characteristic traits. In the first method, a study is carried out on the Gower metric method for mixed variables which is combined with four other agglomeration methods and the second integrate transformed qualitative variables through multiple correspondence analysis with standardized quantitative variables, in a clustering analysis using, like the above, four agglomeration methods: average, Ward or minimum variance, Complete Linkage Agglomerative Clustering and Single Linkage Agglomerative Clustering combining at the same time with two distance measurements (*Euclidean* and *Manhattan*). Different structures were evaluated with the cophenetic correlation coefficient. For different analysis, implemented functions were used on the basis of the R programming language. In this research, it was determined that for both procedures with Gower and Euclidean distance measures, the average agglomeration method maintains better stability in the structures obtained.

Key words: clustering, hierarchical agglomeration, germplasm banks, numerical taxonomy.

INTRODUCCIÓN

Con el uso de las computadoras, la taxonomía numérica, definida por Sneath y Sokal (1973) 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 conjunto de datos multivariados. El análisis de patrones, que es el uso combinado del análisis de clúster 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 germoplasma, dificultando la elaboración de escalas numéricas que integren variables continuas, nominales u ordinales. Entre las alternativas metodológicas para abordar este problema, esta la combinación de técnicas que permiten el análisis de variables mixtas.

La selección correcta de la combinación de múltiples algoritmos de agrupamiento, diferentes representaciones de los datos y diferentes parámetros que son aplicados a los datos originales, depende en gran medida de la selección del método de validación para determinar que estructura es más consistente.

El presente trabajo tuvo como objetivo comparar el desempeño de cuatro métodos de aglomeración jerárquicos en varias estructuras de datos, para determinar cuál método y topología son más consistentes al clasificar un grupo de genotipos de malangas (*Xanthosoma* spp.) y plátanos (*Musa* spp.) a partir de los rasgos que los caracterizan.

MATERIALES Y MÉTODOS

Para realizar la investigación, se usaron datos procedentes de un estudio de accesiones de malangas (*Xanthosoma* spp.) y plátanos (*Musa* spp.) del Banco de Germoplasma, que se conserva en el Instituto de Investigaciones de Viandas Tropicales (INIVIT).

La colección de malangas (*Xanthosoma* spp.) analizadas estuvo conformada por 71 accesiones donde se evaluaron 20 variables cualitativas (nominales y ordinales) y 16 variables cuantitativas (Milián, 2008). La colección de plátanos (*Musa* spp.) incluyó 131 accesiones, con 20 variables cualitativas (nominales y ordinales) y 7 variables cuantitativas incluidas en el Sistema de Descriptores Mínimos (IPGRI-INIBAP/CIRAD, 1996). De esta forma quedaron conformadas dos matrices de datos por colección, una con las variables cualitativas y otra con las cuantitativas.

Al ser las medidas de distancia sensibles a las diferencias de escalas o de magnitudes hechas entre las variables cuantitativas, éstas fueron estandarizadas (Milligan y Cooper, 1988) para lo cual se usó la función *data.Normalization* (Paquete 'clusterSim') con transformación por puntuaciones Z (*z-score*).

Como medidas para evaluar las diferencias y similitudes entre objetos se usaron las distancias *Euclidiana* y *Manhattan* de la función *dist* (Paquete "stats").

Se usaron los métodos de aglomeración, Promedio o UPGMA (*unweighted pair-group using arithmeticAverages*) (Sneath y Sokal, 1973), Ward (Ward, 1963) o de varianza mínima, agrupación de enlace completo (*Complete linkage clustering*) (Sorensen, 1948) y agrupación de enlace simple (*Single linkage clustering)* (Gower, 1967), con la función *hclust* (Paquete "stats").

En el primer procedimiento se realizó un estudio de la métrica de Gower (Gower, 1971) para variables mixtas implementada en la función *daisy* (Paquete "cluster"). A la matriz de distancia se le aplicaron los cuatro métodos de aglomeración. En el segundo se integraron las variables cualitativas transformadas a través del análisis de correspondencia múltiple (Tenenhaus, 1985) con la función *MCA* (Paquete "FactoMineR"), para lo cual se tomaron las dimensiones que acumularon una varianza superior a uno con las cuantitativas normalizadas en un análisis de conglomerados con los cuatro métodos de agrupamiento y las dos medidas de distancia propuestas en el estudio, posteriormente se compararon las diferentes estructuras con el coeficiente de correlación cofenética (Sokal y Rohlf, 1962), el cual es muy utilizado por los taxonomistas (Cuadras, 1981) para determinar la calidad de la clasificación jerárquica obtenida.

Para procesar la información se utilizó un lenguaje de programación, orientado a objetos denominado R (R Development Core Team, 2009); 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 métrica de Gower para variables mixtas a la matriz de datos de variables cualitativas y cuantitativas a las colecciones de malanga (*Xanthosoma* spp.) y plátanos (*Musa* spp.) se obtuvo una matriz de distancia entre las accesiones, a la cual se le aplicó los tres métodos de aglomeración seleccionados para el estudio. En las Tablas 1 y 2 se observa que en el método UPGMA o Promedio fue superior el coeficiente de correlación cofenética con respecto a los demás métodos en las colecciones de malangas y plátanos respectivamente. Esta medida indica el grado de buena clasificación (Cuadras, 1981).

En el segundo al integrar las variables cualitativas transformadas (tomando 31 dimensiones con una varianza acumulada de 91,175 para malanga y 26 con varianza acumulada de 81,11 para plátanos) a través del análisis de correspondencia múltiple con las cuantitativas normalizadas en un análisis de conglomerados, los tres métodos de agrupamiento y las dos medidas de distancias propuestas en el estudio, se observa un coeficiente de correlación cofenética ligeramente superior con la distancia *Euclidiana* respecto a *Manhattan* en los cuatro métodos de agrupamiento para ambas colecciones (Tabla 1 y 2).

Tabla 1. Coeficiente de correlación cofenética de los procedimientos I y II de la colección de malanga (*Xanthosoma* spp.)

Métodos	Malanga					
	Procedimiento I		Procedimiento II			
			Euclidean		Manhattan	
Ward	4	0.5759	4	0.823	4	0.480
Promedio	1	0.7488	1	0.993	1	0.941
Simple	2	0.6983	2	0.991	2	0.932
Completo	3	0.5908	3	0.987	3	0.838

Tabla 2. Coeficiente de correlación cofenética de los procedimientos I y II de la colección de plátanos (*Musa* spp.)

Métodos	Plátano					
	Procedimiento I		Procedimiento II			
			Euclidean		Manhattan	
Ward	4	0.786	4	0.814	4	0.405
Promedio	1	0.886	1	0.991	1	0.973
Simple	3	0.843	2	0.979	2	0.961
Completo	2	0.868	3	0.972	3	0.798

Teniendo en cuenta los resultados obtenidos podemos afirmar que para ambos procedimientos con las medidas de distancias de Gower y euclidiana el método de aglomeración UPGMA es el que mantiene una mejor estabilidad en las estructuras obtenidas, lo cual constituye una herramienta fundamental para el análisis morfológico y la correcta identificación de los materiales utilizados en este estudio ya que permite determinar topologías consistentes.

Dentro de las ventajas de usar estas técnicas multivariadas está la posibilidad de convertir datos cualitativos en cuantitativos lo que permite a los métodos numéricos una mayor capacidad de resolución en la separación de taxones (Alfaro, 2000).

Estas mezcla de análisis es aplicada por primera vez para la clasificación taxonómica de las colecciones cubanas de germoplasma de plátanos y malangas.

En este estudio nunca se tuvo en cuenta las relaciones evolutivas entre los caracteres descritos, y se le asignó igual importancia a cada uno de ellos, o sea, las clasificaciones establecidas no son filogenéticas.

Los autores concuerdan en que no existe un método de agrupamiento que sea capaz de encontrar todos los tipos de conglomerados posibles en cualquier

conjunto de datos (Jain, 1999).
Kleimberg (2002) demuestra la imposibilidad de diseñar un algoritmo de agrupamiento perfecto, y distintos algoritmos serán adecuados para encontrar distintos tipos de agrupamientos en los datos.

CONCLUSIONES

Los análisis de conglomerados jerárquicos dan la posibilidad de utilizar distintos tipos de medidas para estimar la distancia existente entre los casos, la posibilidad de transformar la métrica original de las variables y de seleccionar de entre una gran variedad de métodos de aglomeración. Pero no existe ninguna combinación de estas posibilidades que optimice la solución obtenida. Por lo cual se recomienda valorar distintas soluciones para elegir la más consistente.

BIBLIOGRAFÍA

1. Alfaro, Yanely; V. Segovia: Maíces del sur de Venezuela clasificados por taxonomía numérica. I. Caracteres de la planta. *Agronomía Tropical* 50(3): 413-433; 2000.

2. Cuadras, C. M.: Métodos de Análisis Multivariante. Editorial *EUNIBAR*, Barcelona. España, 1981.

3. Gower, J. C.: A general coefficient of similarity and some of its roperties. *Biometrics* (27): 857–874; 1971.

4. Gower, J. C.: A comparison of some methods of cluster analysis. *Biometrics* (23): 623-628; 1967.

5. IPGRI-INIBAP/CIRAD: Descriptores para el banano (*Musa* spp.) Instituto Internacional de Recursos Fitogenéticos. Roma, Italia. Red Internacional para el mejoramiento del banano y el plátano, Montpellier, Francia y el Centre de Cooperation Internationale. En: Recherche Agronomique pour le Development, Montpellier, Francia, 1996.

6. Kleimberg, J.: An impossibility theorem for clustering. In: Proc. of the 16th conference on Neural Information Processing Systems (15): 446–453; 2002.

7. Jain, A. K. y col.: Data clustering: a review. *ACM Computing Surveys* 31(3): 264–323; 1999.

8. Milligan, G. W.; M. C. Cooper: A study of standardization of variables in cluster analysis. *Journal of Classification* (5):181-204; 1988.

9. Milián, Marilys: Caracterización de la variabilidad de los cultivares de la colección cubana de germoplasma del género *Xanthosoma* (Araceae). Tesis para aspirar al grado científico de Doctor en Ciencias Agrícolas, INIVIT, Santo Domingo, Villa Clara, Cuba, 2008.

10. R Development Core Team: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN: 3-900051-07-0; 2009.

11. Sneath, P. H.; R. R. Sokal: Numerical Taxonomy. W. H. Freeman. San Francisco, USA, 1973, 458 p.

12. Sokal, R. R.; F. J. Rohlf: The comparisons of dendrograms by objetive methods. *Taxon* (11):33-40; 1962.

13. Sorensen, T. A.: A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to alalysies of vegetation on Danish commons. *Biologiske Skrifter* (5):1-34; 1948.

14. Tenenhaus, M.; F. W. Young: An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis ans other methods for quantifying categorical multivariate data. *Psychometrika* 50(1): 91-119; 1985.

15. Ward, J. H. Jr.: Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc*. (58):236–244; 1963.

Recibido: *03/05/2013*
Aceptado: *14/07/2013*