Rev. Agricultura Tropical Vol. 1 No. 2:67-73, 2015
RNPS: 2397 (Versión electrónica)

MÉTODOS DE ESTANDARIZACIÓN DE VARIABLES CUANTITATIVAS EN COLECCIONES DE GERMOPLASMA VEGETAL

Osmany Molina Concepción*, Raisa L. García Rodríguez, Marilys Milián Jiménez, Lianet González Díaz 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 recursos fitogenéticos se han convertido en una prioridad científica, lo cual hace importante el análisis de esta diversidad mediante métodos cuantitativos que ayuden a agrupar poblaciones de un mismo género o especie. El banco de germoplasma del Instituto de Investigaciones de Viandas Tropicales (INIVIT) conserva accesiones de raíces, rizomas, tubérculos, plátanos y bananos procedentes de colectas e introducciones nacionales e internacionales, cuya variabilidad necesita de este tipo de análisis. Esta investigación tiene como objetivo estudiar la respuesta de cuatro métodos de aglomeración jerárquicos sobre una matriz de datos sin estandarizar y cuatro matrices con diferentes formas de estandarización correspondientes a datos cuantitativos de accesiones de ñame (*Dioscorea* spp.), malanga (*Xanthosoma* spp.) y de plátano (*Musa* spp.) que se conservan en colecciones de germoplasma del INIVIT. La fortaleza de los agrupamientos del conjunto de datos obtenidos por los métodos de aglomeración, se comprobó con el coeficiente aglomerativo y las diferentes estructuras de los conglomerados se evaluaron con el coeficiente de correlación cofenética. Para los diferentes análisis se utilizaron funciones implementadas en el paquete *clusterSim*, sobre la base del lenguaje de programación R. Las estrategias de análisis demostraron que los métodos de aglomeración de Ward y promedio, permiten obtener una mejor clasificación taxonómica de las accesiones presentes en las colecciones de germoplasma de ñame, malanga y plátano.
Palabras claves: conglomerados, clasificación, medidas de distancia, programación R.

QUANTITATIVE VARIABLES STANDARIZING METHODS IN VEGETAL GERMPLASM COLLECTIONS

ABSTRACT
Fitogenetic resources have become a scientific priority, which makes this diversity important by means of quantitative methods that help the joining of groupings of a same kind or species. Research Institute of Tropical Root and Tuber Crops, Bananas, and Plantains (INIVIT) germplasm bank keeps accessions of roots, tuber, bananas and plantains coming from collect and national and international introductions, which variability has a need for this kind of analysis. This research aims at studying the response of four hierarchical agglomeration methods on a data matrix without standardization and four matrices with different forms of standardization corresponding to quantitative data of yam accessions (*Dioscorea* spp.), cocoyan (*Xanthosoma* spp.) and plantain (*Musa* spp.) which are kept in INIVIT germplasm collections. The strength of the joining of data groupings obtained by agglomeration methods was verified by the

agglomerative coefficient and the different structures of the conglomerates were assessed by the coefficient of cophenetic correlation. For the different analysis *clusterSim* implemented functions in the package were used, based on R programming language. The analysis strategies demonstrated that the agglomeration methods Ward and average make possible to obtain a better taxonomic classification of the accessions present in the germplasm collections of yam, cocoyan and plantain.
Keywords: conglomerates, classification, distance measures, R programming.

INTRODUCCIÓN

Una de las prioridades dentro de los recursos fitogenéticos es el análisis de la diversidad mediante métodos cuantitativos que ayuden a agrupar poblaciones de un mismo género o especie.

Los métodos que se utilizan generalmente en el estudio de divergencias entre individuos siguen una aproximación fenética o numérica (Franco y Hidalgo, 2003). En este tipo de clasificación son extremadamente empleados los métodos de aglomeración jerárquica.

La mayoría de estos métodos requieren que las escala de medición de todas las variables sean iguales o similares (Jajuga y Walesiak, 2000; Aggarwal et al., 2015), por lo cual, es necesaria la estandarización de las variables en los casos donde las medidas de disimilaridad son sensibles a las diferencias de magnitud o escalas, tal como la distancia Euclidiana (Milligan y Cooper, 1985).

Muchas de las medidas de distancia varían respecto a la métrica de los datos, ya que las diferencias existentes entre las variables con puntuaciones altas pueden anular las diferencias existentes entre las variables con puntuaciones bajas.

Para resolver este problema generalmente no se utilizan los valores directos de los descriptores evaluados en los bancos de germoplasma, sino los valores transformados a escalas del mismo rango. La estandarización de los descriptores resulta muy útil para eliminar su dependencia respecto a las unidades de medida empleadas.

Existen en la literatura estudios realizados por autores entre ellos Milligan y Cooper (1988) quienes muestran el efecto de la estandarización sobre la estructura de los conglomerados en configuraciones de datos.

Las variables medidas en diferentes escalas no contribuyen igualmente al análisis, por esta razón se realizó un estudio sobre tres bases de datos con el objetivo de observar que método de aglomeración jerárquico, de estandarización y que medida de distancia permiten obtener una mejor clasificación taxonómica de las accesiones presentes en los bancos de germoplasma.

MATERIALES Y MÉTODOS

Se usaron datos procedentes de un estudio de accesiones de ñame (*Dioscorea* spp.), malanga (*Xanthosoma* spp.) y plátano (*Musa* spp.) del Banco de Germoplasma que se conserva en el Instituto de Investigaciones de Viandas Tropicales (INIVIT).

La colección de ñame incluye 86 accesiones donde se evaluaron nueve variables cuantitativas incluidas en el Sistema de Descriptores Mínimos (Sánchez *et al*., 1995). En el germoplasma de malanga con 71 accesiones, se evaluaron 16 cuantitativas (Milián, 2008), y el de plátano con 131 accesiones en estudio, fueron evaluadas siete variables cuantitativas incluidas en el Sistema de Descriptores Mínimos (IPGRI - INIBAP/CIRAD, 1996).

A continuación se enumeran un grupo de técnicas necesarias para llevar a cabo el presente estudio:
Después de representados los datos, las medidas de distancia sensibles a las diferencias de escalas o de magnitudes hechas entre las variables cuantitativas deben estandarizarse (Milligan y Cooper, 1988) para llevarlas a unidades comparables (Jajuga y Walesiak, 2000).
Existen numerosas formas de estandarización, pero en el presente trabajo solo se tendrán en cuenta cinco de ellas, implementadas en la función *data.Normalization* (n0 (no estandarización), n1, n4, n6 y n7) del paquete *clusterSim* (Walesiak and Dudek, 2015), con las cuales se conforman cuatro matrices de distancia a partir de la distancia *Euclidiana*, teniendo en cuenta solo las variables cuantitativas. A estas cuatro matrices de distancia, para las tres bases de datos, se le aplican posteriormente cuatro métodos de agrupamiento jerárquicos.
En esta investigación se usan los métodos de aglomeración jerárquicos de Ward en sus dos implementaciones *ward.D* (Ward, 1963) y *ward.D2* (Murtagh y Legendre, 2014), Promedio (UPGMA) (Sneath y Sokal, 1973), Enlace Simple (Gower, 1967), Enlace Completo (Sorensen, 1948) implementados en la función *hclust*() en el paquete *stats* que forma parte de la librería básica de R que se instala por defecto.
Una vez obtenido el resultado del método de aglomeración, en el análisis taxonómico es importante determinar si el conjunto de datos muestra una tendencia a formar grupos, lo cual se determina a través del coeficiente aglomerativo (CA) con la función *coef.hclust*() en el paquete *cluster*. El coeficiente aglomerativo describe la fortaleza de la estructura de grupos que se ha obtenido, para comparar la calidad (desde el punto de vista del análisis de conglomerados) de diferentes conjuntos de datos, suponiendo que se usa el mismo algoritmo en cada conjunto (Rousseeuw, 1986).
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. El método más usado para verificar este hecho, o sea, para ver la relación entre el dendrograma y la matriz de proximidades original, es el Coeficiente de Correlación Cofenético (CCC) (Sokal y Rohlf, 1962).
Para procesar la información se utilizó el lenguaje de programación orientada a objetos, denominado R, versión 3.1.2 (R Development Core Team, 2014), el cual es un conjunto de programas integrados para análisis estadísticos y gráficos.

RESULTADOS Y DISCUSIÓN

Como ya se enunció, se usaron datos procedentes de un estudio de accesiones de ñame, malanga y plátano, a estas bases de datos como proceso preliminar se le eliminaron los descriptores con rasgos uniformes.
En una primera estructura de los datos no se considera la estandarización de los mismos (n0). A su vez se analizaron otras cuatro estructuras de datos, realizándose la estandarización con la función *data.Normalization* (n1, n4, n6 y n7). Como se observa en la Tabla 1, el método de *Ward.D* muestra un coeficiente aglomerativo superior con respecto a los demás métodos de agrupación, para las cinco estructuras de datos, y entre las diferentes bases de datos, expresado en estructuras bien definidas, y del menor valor de Coeficiente de Correlación Cofenético con relación a los demás métodos, y UPGMA fue superior en el Coeficiente de Correlación Cofenético, conclusiones similares fueron planteadas por Blackburn *et al.* (2005), por lo cual es el

que mejor se ajusta a la matriz de distancia original (Sokal y Rohlf ,1962; Mohammadi y Prasanna, 2003: Podani y Schmera, 2006). La respuesta es similar para las tres colecciones en estudio. Además, se puede observar que los métodos de estandarización n1 y n6 tienen comportamiento similar para los coeficiente aglomerativo y Coeficiente de Correlación Cofenético al igual que n4 y n7, para todos los métodos en las tres bases de datos.

Tabla 1. Resumen de los coeficiente aglomerativo y Coeficiente de Correlación Cofenético para las tres matrices de datos en estudio.

Bases de datos	Métodos de agrupamiento	Est. n0		Est. n1		Est. N4		Est. N6		Est. n7	
		CA	CCC	CA	CCC	CA	CCC	CA	CCC	CA	CCC
Malanga (*Xanthosoma* spp.)	Ward.D	0,986	0,670	0,893	0,383	0,922	0,548	0,893	0,383	0,922	0,548
	Ward.D2	0,956	0,681	0,830	0,292	0,843	0,557	0,830	0,511	0,843	0,557
	Promedio	0,790	0,772	0,566	0,756	0,539	0,746	0,566	0,756	0,539	0,746
	Simple	0,524	0,619	0,475	0,726	0,389	0,63	0,475	0,726	0,389	0,630
	Completo	0,879	0,681	0,693	0,693	0,679	0,546	0,693	0,693	0,679	0,546
Ñame (*Dioscorea* spp.)	Ward.D	0,954	0,500	0,926	0,39	0,938	0,419	0,926	0,39	0,938	0,419
	Ward.D2	0,894	0,571	0,853	0,411	0,869	0,448	0,853	0,411	0,869	0,448
	Promedio	0,622	0,676	0,616	0,673	0,614	0,64	0,616	0,673	0,614	0,64
	Simple	0,492	0,549	0,538	0,641	0,516	0,583	0,538	0,641	0,516	0,583
	Completo	0,720	0,547	0,726	0,45	0,707	0,524	0,726	0,45	0,707	0,524
Plátano (*Musa* spp.)	Ward.D	0,998	0,670	0,990	0,648	0,994	0,819	0,990	0,648	0,994	0,819
	Ward.D2	0,991	0,718	0,964	0,692	0,974	0,851	0,964	0,692	0,974	0,851
	Promedio	0,964	0,892	0,843	0,787	0,843	0,877	0,843	0,787	0,843	0,877
	Simple	0,923	0,659	0,738	0,698	0,790	0,819	0,738	0,698	0,790	0,819
	Completo	0,978	0,854	0,901	0,684	0,907	0,808	0,901	0,684	0,907	0,808

CA: coeficiente aglomerativo CCC: Coeficiente de Correlación Cofenético.

Los mejores valores de coeficiente aglomerativo del método de ward.D se reflejan para los datos no estandarizados (n0) en las bases de datos de malanga Xanthosoma y ñame, por otro lado los valores más bajos de Coeficiente de Correlación Cofenético para este método se obtienen con n1 y n6.

Sin embargo para la base de datos de plátano los valores de coeficiente aglomerativo y Coeficiente de Correlación Cofenético tienen una mayor estabilidad con respecto a los métodos de estandarización empleado.

Adicionalmente, se puede observar que en las figuras 1, 2 y 3 los dendrogramas con el método de ward.D son diferentes y se observa una contracción en el dendrograma de la figura 1.

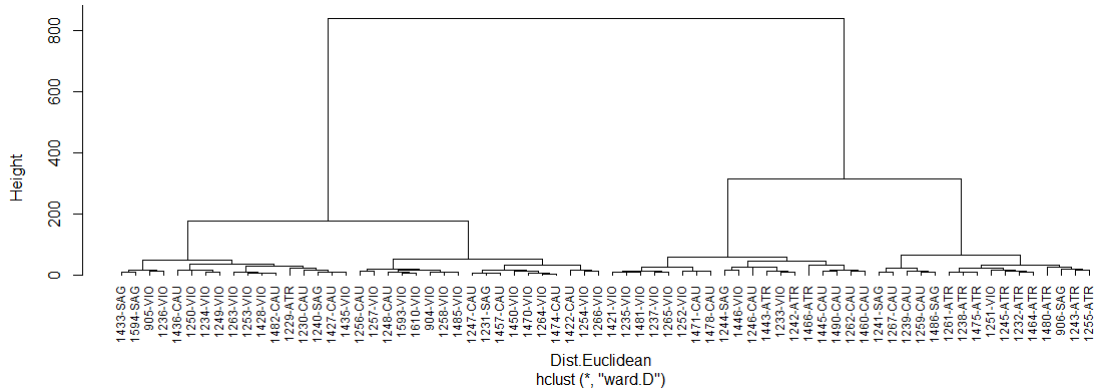

Figura 1. Dendrograma con el método de ward.D a partir de la distancia Euclidiana con los datos sin estandarización.

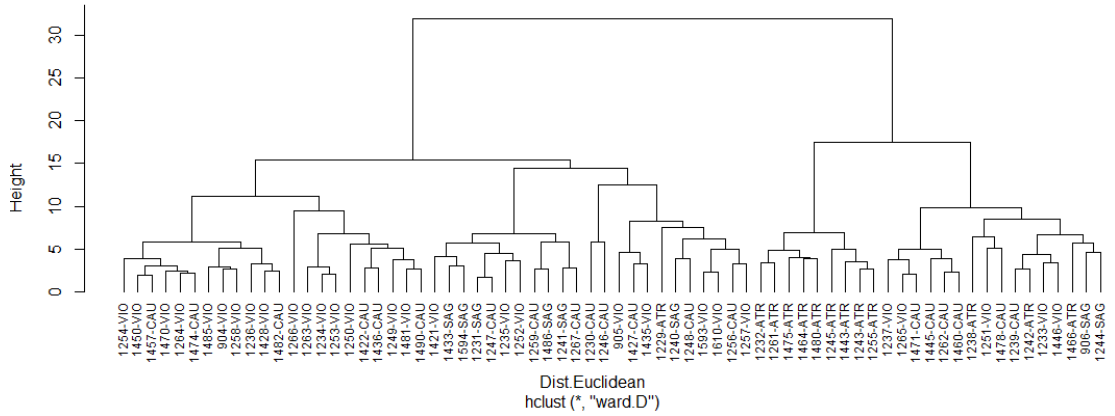

Figura 2. Dendrograma con el método de ward.D a partir de la distancia Euclidiana con estandarización n1.

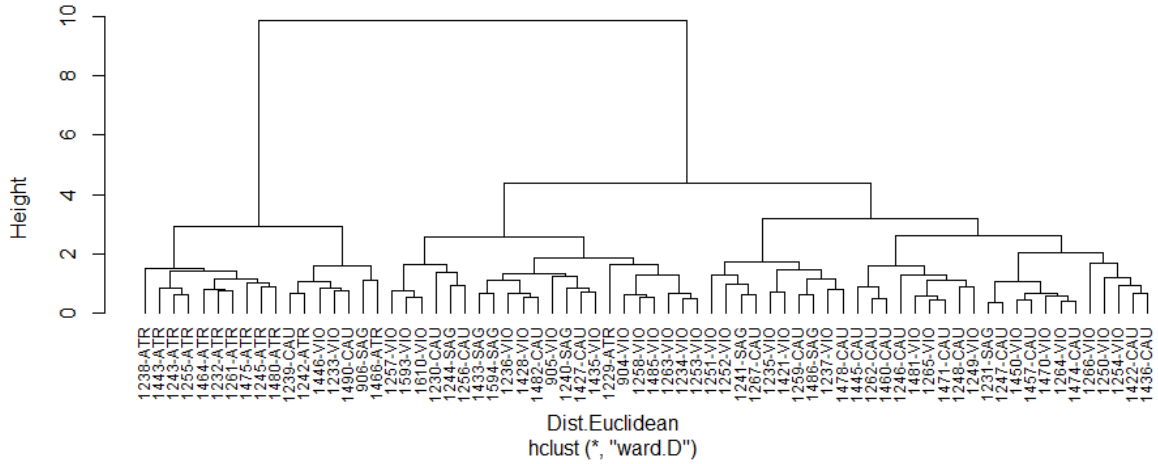

Figura 3. Dendrograma con el método de ward.D a partir de la distancia Euclidiana con estandarización n4.

Dado los resultados para la base de datos de plátano se debe estandarizar con los métodos n1 y n4 atendiendo a que presentan una mejor representación gráfica y mejor coeficiente aglomerativo y Coeficiente de Correlación Cofenético. En el caso de las base de datos de malanga y ñame los mejores resultados se obtienen sin estandarizar, pero el método de ward.D*,* como ya se señaló, genera un dendrograma con una contracción sobre el eje y. Además, la medida de distancia Euclidiana es sensible a las diferencias de escalas o de magnitudes hechas entre las variables.
Con la estandarización n1 el método de ward.D y ward.D2 obtiene los peores resultados de coeficiente aglomerativo y Coeficiente de Correlación Cofenético con respecto a los demás métodos estudiados.
Se han realizado estudios taxonómicos en estas colecciones de germoplasma utilizando diferentes métodos de aglomeración jerárquicos a partir de la distancia de Gower; sin embargo no existen estudios de estandarización anteriores sobre estas colecciones.

CONCLUSIONES

Se dispone de un conocimiento detallado de los métodos de estandarización que influyen mejor en los resultados de la clasificación del germoplasma en estudio.
Por su flexibilidad, la concepción de este análisis puede ser aplicada a otros estudios de clasificación en bancos de germoplasma vegetal.

BIBLIOGRAFÍA

AGGARWAL, S.; P. PHOGHAT and S. MAITREY. 2015. Hierarchical Clustering- An Efficient Technique of Data mining for Handling Voluminous Data*. Int. J. Comput. Appl.,* 129(13):31–36.

FRANCO, T.L. y R. HIDALGO (eds). 2003. Análisis Estadístico de Datos de Caracterización Morfológica de Recursos Fitogenéticos. *Boletín técnico IPGRI*, vol 8. Instituto Internacional de Recursos Fitogenéticos (IPGRI), Cali, Colombia.

GOWER, J.C. 1967. A comparison of some methods of cluster analysis*. Biometrics*, 23:623–628.

IPGRI-INIBAP/CIRAD. 1996. 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; Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Montpellier, Francia. ISBN 92-9043-307-8.

JAJUGA, K. and M. WALESIAK. 2000. Standardization of data set under different measurement scales, in: DECKER, R. and GAUL, W. (Eds.), *Classification and Information Processing at the Turn of the Millennium*. Springer-Verlag, Berlin, Heidelberg, p. 105–112.

MILIÁN, M.J. 2008. Caracterización de la variabilidad de los cultivares de la colección cubana de germoplasma del género *Xanthosoma (Araceae).* Tesis para optar por el grado de Doctor en Ciencias Biológicas. Ciudad de La Habana, 123 p.

MILLIGAN, G.W. and M.C. COOPER. 1985. An examination of procedures for determining the number of clusters in data set*. Psychometrika*, 50:159–179.

MILLIGAN, G.W. and M.C. COOPER. 1988. A study of standardization of variables in cluster analysis*. J. Classif.* 5:181–204.

MOHAMMADI, S.A. and B.M. PRASANNA. 2003. Analysis of genetic diversity in crop plants - salient statistical tools and considerations*. Crop Sci.,* 43:1235-1248.

MURTAGH, F. and P. LEGENDRE. 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *J. Classif.,* 31(3):274-295.

PODANI, J. and D. SCHMERA. 2006. On dendrogram-based measures of functional diversity*. Oikos*, 115:179-185.

R DEVELOPMENT CORE TEAM. 2014. R: A language and environment for statistical computing, R Foundation for Statistical Computing. Vienna, Austria.

ROUSSEEUW, P.J. 1986. A visual display for hierarchical classification*. Data Anal. Inform.,* 4:743–748.

SÁNCHEZ, I.; M. MILIÁN; A. RAYAS y S. RODRÍGUEZ. 1995. Lista de descriptores y caracterización de la colección cubana de ñame (*Discorea* spp).

SNEATH, P.H.A. and R.R. SOKAL. 1973. Numerical taxonomy. The principles and practice of numerical classification. W.H. Freeman and Co, San Francisco, California, USA.

SOKAL, R.R. and F.J. ROHLF. 1962. The comparisons of dendrograms by objective methods*. Taxon*, 11: 3–40.

SORENSEN, T.A. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of vegetation on Danish commons*. Biol. Skr.,* 5:1–34.

WALESIAK, M. and A. DUDEK. 2015. ClusterSim: Searching for optimal clustering procedure for a data set. Consultado: 12 de octubre de 2015. Disponible en: http://cran.fhcrc.org/web/packages/clusterSim/clusterSim.pdf.

WARD, J.H.J. 1963. Hierarchical grouping to optimize an objective function*. J. Am. Stat. Assoc.,* 58:236–244.