Rev. Agricultura Tropical Vol. 4 No. 2:58-64, 2018
ISSN on line: 2517-9292

ANÁLISIS DE VARIABLES CUANTITATIVAS EN BANCOS DE GERMOPLASMA VEGETAL

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
Uno de los problemas en el análisis de conglomerados es la decisión de codificar o no las variables cuantitativas antes de aplicar alguna medida de distancia, por lo cual, el presente estudio examina este inconveniente a partir de dos métodos de agrupamiento jerárquico (promedio y Ward), para determinar la mejor clasificación del germoplasma de malanga (*Xanthosoma* spp.), ñame (*Dioscorea* spp.) y plátanos (*Musa* spp.), con dos estructuras de datos, una primera matriz cuantitativa, y otra con estos datos codificados y tratados como ordinales. En la investigación se utilizó la métrica de Gower para obtener la matriz de distancia de las dos variantes. Las estructuras de los conglomerados se evaluaron con los coeficientes aglomerativo y de correlación cofenético. Para los análisis se utilizaron funciones implementadas sobre la base del lenguaje de programación R. En este estudio se demuestra que las dos variantes de tratamiento de los datos cuantitativos sobre estos germoplasmas varían su respuesta en los agrupamientos de acuerdo al método de aglomeración que se use: el método de aglomeración promedio da una mejor agrupación de los genotipos con la variante uno, y el método de Ward con la variante dos; aunque con la variante dos se obtienen de forma general mejores resultados de los coeficientes aglomerativo y de correlación cofenético. La concepción de este análisis puede ser aplicada a otros estudios de clasificación en bancos de germoplasma vegetal.
Palabras clave: clasificación, lenguaje de programación R, métrica de Gower.

QUANTITATIVE VARIABLES ANALYSIS IN VEGETABLE GERMPLASM BANKS

ABSTRACT
One of the problems in the cluster analysis is the decision of coding or not the quantitative variables before applying some distance measure. The present research examines this inconvenience using two hierarchical cluster methods (average and Ward) to determine the best classification of taro germplasm (*Xanthosoma* spp.), yam (*Dioscorea* spp.) and plantain (*Musa* spp.), with two data structures, first, one quantitative data matrix and another one with these coded data and treated as ordinals. In the research, Gower´s metric was used to obtain the two variants distance matrix. The cluster structures were evaluated with the agglomerative and cophenetic correlation coefficients. Implemented functions were used for the analysis based on R language program. Through this work it is shown that the two variants of quantitative data treatment about these germplasms vary their response in the clusters according to the agglomeration method used: the average agglomeration method offers a better genotype cluster with the first variant, and the Ward's method with the second one; although using the second variant, best results are generally obtained in relation to the

agglomerative and cophenetic correlation coefficients. This analysis conception can be applied to other classification studies in vegetable germplasm banks.
Keywords: classification**,** R programing language, Gower's metric.

INTRODUCCIÓN

En los bancos de germoplasma cada base de datos está compuesta por un conjunto de accesiones o casos representativos de un problema en cuestión y a su vez, cada accesión, se describe mediante un conjunto de descriptores, también conocidos como variables de entrada o características, los cuales pueden ser cualitativos (categóricos) o cuantitativos (numéricos). Las variables categóricas pueden definirse como binarias, nominal u ordinal (Núñez *et al*., 2003; Crossa and Franco, 2004; Podani and Schmera, 2006).

La medida con escala nominal asigna números que se usan para etiquetar o identificar sujetos u objetos. Las escalas nominales, también conocidas como categóricas, proporcionan el número de ocurrencias en cada clase o categoría de la variable que se está usando y donde las categorías no tienen un orden inherente. En las bases de datos de germoplasma los números o símbolos asignados a las accesiones no tienen más significado cuantitativo que indicar el valor del descriptor morfoagronómico acorde a las listas de descriptores (por ejemplo: color de la masa, hábito foliar). La escala ordinal son datos categóricos donde las categorías tienen un orden significativo, pero sin una distancia medible entre las categorías. Por ejemplo, hay un orden para los valores alto, medio y bajo, pero no se puede calcular la "distancia" entre los valores. Los datos numéricos continuos son medidos en una escala de intervalo o de razón, donde los valores de los datos indican el orden y la distancia entre ellos, estos proporcionan el nivel más alto de medida de precisión y permiten realizar casi todas las operaciones matemáticas. Estas dos escalas tienen unidades constantes de medida, de tal forma que las diferencias entre dos puntos adyacentes de cualquier parte de la escala son iguales. La única diferencia real entre las escalas de intervalo y de razón es que la primera tiene un punto cero arbitrario, mientras que la segunda tiene un punto de cero absolutos.

Medidas de distancia que analizan diferentes tipos de variables simultáneamente no son utilizadas con frecuencia para cuantificar disimilitud genética (Vieira *et al.*, 2007). Entre las alternativas metodológicas para abordar este problema está la métrica de Gower (1971), que permite el análisis de variables mixtas.

Los algoritmos de agrupación (o *clustering*), solo requieren de la definición previa del vector de características. Una vez establecido dicho vector, los procedimientos de agrupación reciben como datos de entrada los objetos convertidos lógicamente en vectores numéricos a clasificar, de modo que, a partir de estos datos de entrada, el algoritmo sin supervisión de ningún tipo, agrupa esos vectores en clases disjuntas. Todos ellos se basan en el empleo sistemático de las distancias entre los vectores (objetos a agrupar), así como entre los grupos que se van haciendo y deshaciendo a lo largo del proceso correspondiente.

El análisis de los datos implica la separación, identificación y medida de variación en un conjunto de variables. Un análisis cuidadoso de los datos conduce a una mejor predicción y a una evaluación más precisa de los resultados.

Uno de los problemas en el análisis de conglomerados es la decisión de codificar o no las variables cuantitativas antes de aplicar alguna medida de distancia. No es conocido

el resultado de este paso en la clasificación de los bancos de germoplasma vegetal que se conservan en el Instituto de Investigaciones de Viandas Tropicales (INIVIT).
Esta investigación tiene como objetivo analizar la respuesta de dos métodos de aglomeración jerárquicos sobre una matriz de datos cuantitativos y una con los datos cuantitativos codificados y tratados como ordinales correspondientes a datos cuantitativos de accesiones que se conservan en el germoplasma de malanga (*Xanthosoma* spp.), ñame (*Dioscorea* spp.) y plátanos (*Musa* spp.) del INIVIT.

MATERIALES Y MÉTODOS

Se usaron datos procedentes de un estudio de accesiones de malanga (*Xanthosoma* spp.), ñame (*Dioscorea* spp.) y plátanos (*Musa* spp.) del Banco de Germoplasma, que se conserva en el Instituto de Investigaciones de Viandas Tropicales (INIVIT).
La colección analizada de malanga con 71 accesiones, se evaluaron 16 variables cuantitativas (Milián, 2008). Dentro de la colección cubana de malanga se incluyen 57 accesiones de la especie *Xanthosoma sagittifolium* (L.) Schott (SAG) y 10 accesiones que conforman un grupo que no tienen una especie definida, o cuya clasificación especifica no está definida de *Xanthosoma* spp. (SP). Además, está presente la especie *Xanthosoma brasilense* Engl. (BRA), una de la especie *Xanthosoma atrovirens* Koch & Bouché (ATR) y una de la especie *X. nigrum* (Vell.) Mansf. (NIG) (Milián, 2008).
El germoplasma de ñame contiene 86 accesiones donde se evaluaron nueve variables cuantitativas incluidas en el Sistema de Descriptores Mínimos (Sánchez *et al.* 1995), y el de plátano y bananos incluye 131 accesiones, con siete variables cuantitativas incluidas en el Sistema de Descriptores Mínimos (IPGRI-INIBAP/CIRAD, 1996).
Para el análisis se conforman dos variantes de estructura de datos, una primera está formada por las matrices cuantitativas de las accesiones de los tres germoplasmas en estudio, a las cuales se le aplica la métrica de Gower para este tipo de variable, y una segunda en la cual las variables cuantitativas de las tres bases de datos fueron codificadas como cualitativas, según el sistema de descriptores establecido para cada germoplasma, su tratamiento en los análisis es como variables ordinales como está descrito por Podani (1999).
Como medida de distancia se utilizó el coeficiente general de Gower (Gower, 1971). Esta métrica está implementada en la función *gowdis*(), descrita en el paquete "FD" (Laliberté and legendre, 2010; Laliberté and Shipley, 2011).
En esta investigación se usaron los métodos de aglomeración jerárquicos de Ward (Ward, 1963; Murtagh and Legendre, 2014) en su opción *ward.D* y promedio (UPGMA) implementados en la función *hclust*() en el paquete "*stats"* (R Core Team, 2017) que forma parte de la librería básica de R que se instala por defecto.
En el análisis taxonómico, una vez obtenido el resultado del método de aglomeración y su correspondiente dendrograma, 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 desempeño del método de aglomeración se verificó a través del coeficiente de correlación cofenético (CCC), es decir, la fiabilidad de los métodos de la agrupación depende en gran medida de esta magnitud (Sokal and Rohlf, 1962; Mantel, 1967), que está implementada en la función *cophenetic*() del paquete "*stats"*.
Para procesar la información se utilizó el lenguaje de programación orientado a objetos, denominado R (R Core Team, 2017) el cual es un conjunto de programas integrados

para análisis estadísticos y gráficos. R es un software libre y se basa en comandos con los que se pueden acceder a todas sus bibliotecas y ejecutar múltiples funciones.

RESULTADOS Y DISCUSIÓN

Como ya se enunció se usaron datos procedentes de un estudio de accesiones de malanga (*Xanthosoma* spp.), ñame (*Dioscorea* spp.) y plátanos y bananos (*Musa* spp.).
Se obtuvieron dos matrices de distancia para cada base de datos, una como resultado de aplicar el coeficiente general de Gower a los datos cuantitativos, y la otra como resultado de aplicar esta métrica a los datos cuantitativos codificados según el sistema de descriptores establecido para dicho germoplasma y tratados como variables ordinales con la variante de "podani" para este tipo de variables implementada en la función *gowdis*().
Los mayores valores del coeficiente aglomerativo, se obtuvieron con la segunda estructura de datos para todos los métodos de análisis de conglomerados no jerárquicos en las tres bases de datos, por lo cual se puede enunciar que hay una fuerte estructura presente en los datos en la aplicación de esta variante; siendo el método de ward.D el de mayor índice para los tres germoplasmas en estudio y las dos variantes de tratamiento de los datos, como se puede observar en los resultados que se reflejan en la tabla 1.
En este estudio, los dos métodos de aglomeración son comparados con el coeficiente de correlación cofenético. UPGMA reveló mayor coeficiente en comparación con el método de ward.D para las dos variantes (Tabla 1), por lo cual es el que mejor se ajusta a la matriz de distancia original (Podani and Schmera, 2006).

Tabla 1. Resumen de los coeficientes de aglomeración y de correlación cofenético para las dos estructuras de datos con el germoplasma de malanga *Xanthosoma* spp. Métodos de aglomeración: ward.D y UPGMA.

Base de datos	Métodos de aglomeración	Variante 1		Variante 2	
		CA	CCC	CA	CCC
Xanthosoma spp.	ward.D	0,928	0,422	0,995	0,638
	UPGMA	0,564	0,671	0,930	0,712
Dioscorea spp.	ward.D	0,956	0,438	0,998	0,635
	UPGMA	0,671	0,685	0,986	0,787
Musa spp.	ward.D	0,995	0,766	0,999	0,603
	UPGMA	0,877	0,824	0,992	0,685

CA: coeficiente de aglomeración. CCC: coeficiente de correlación cofenético.

Los dendrogramas obtenidos para la base de datos de malanga (Figura 1) tienen una relativa diferencia de estructura entre ellos; iguales resultados se obtuvieron para las bases de datos restantes germoplasmas en estudio. En el primer dendrograma derivado de la métrica de Gower variante "numeric" se define claramente a la altura de 0,29, dos grupos bien separados: el grupo uno está formado por una mezcla de nueve accesiones de SP, cuatro de la especie SAG y una de la especie BRA, el grupo dos está formado por 55 accesiones de la especie SAG, una de la ATR, una de la SP y una de la especie NIG.

En el segundo dendrograma obtenido con la variante “podani” de la métrica de Gower estableciendo un corte a la altura 0,17 se forman dos grupos: el grupo uno tiene ocho accesiones de la especie SP, siete de la especie SAG y una de la BRA, el grupo dos mezclas 52 accesiones de la especie SAG con dos de la especie SP, una de la especie NIG y una de la ATR. Por lo cual, el dendrograma derivado de la métrica de Gower variante “numeric” con el método promedio se obtiene una leve mejor explicación taxonómica de la base de datos.

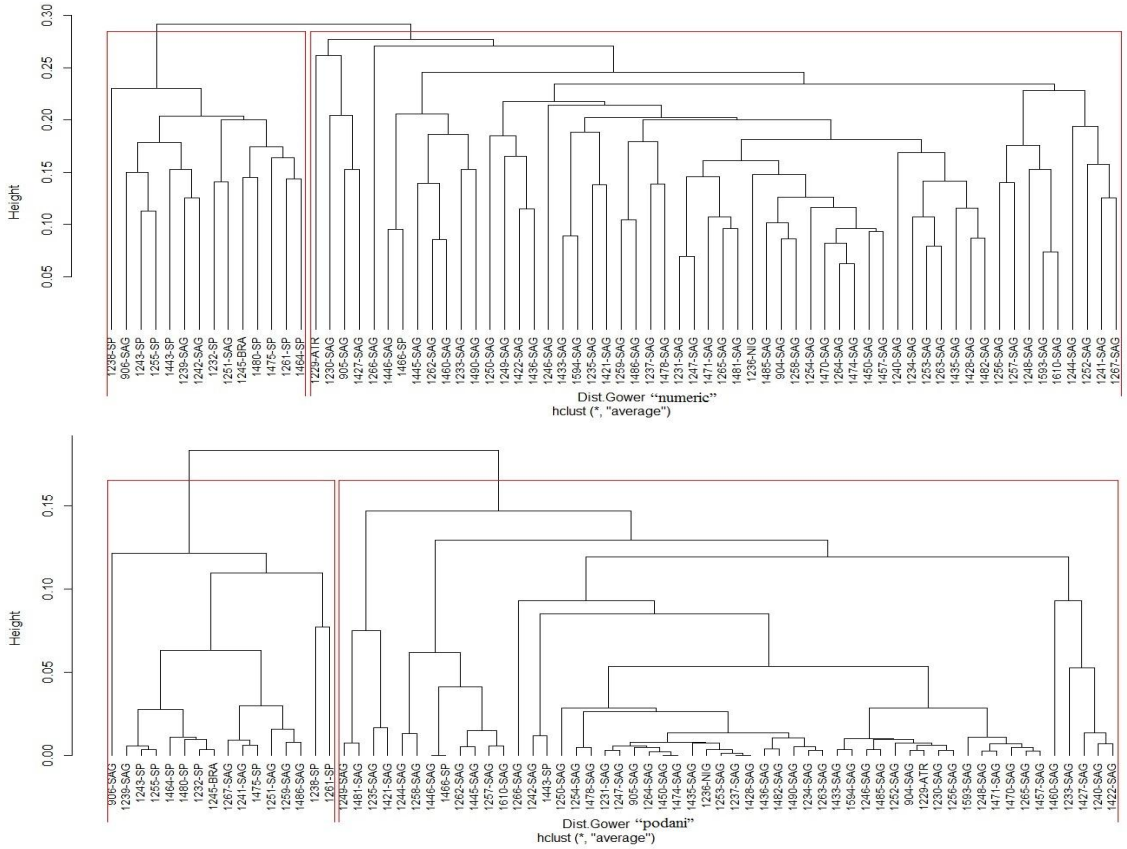


Figura 1. Dendrogramas obtenidos al aplicar el método UPGMA con la métrica de Gower, variantes “numeric” y “podani”, sobre la matriz de datos *Xanthosoma* spp.

En el primer dendrograma de la figura 2 con la variante “numeric” definida en la métrica de Gower se definen tres grupos a la altura de uno, con mayor dispersión de las especies con respecto al dendrograma obtenido con la variante “podani”. En el segundo dendrograma un corte a la altura de dos se forman dos grupos: el grupo uno tiene ocho accesiones de la especie SP, siete de la especie SAG y una de la BRA, el grupo dos mezclas 52 accesiones de la especie SAG con dos de la especie SP, una de la especie NIG y una de la ATR. Por lo cual, el dendrograma derivado de la métrica de Gower variante “podani” con el método Ward proporciona una mejor descripción taxonómica de la base de datos.

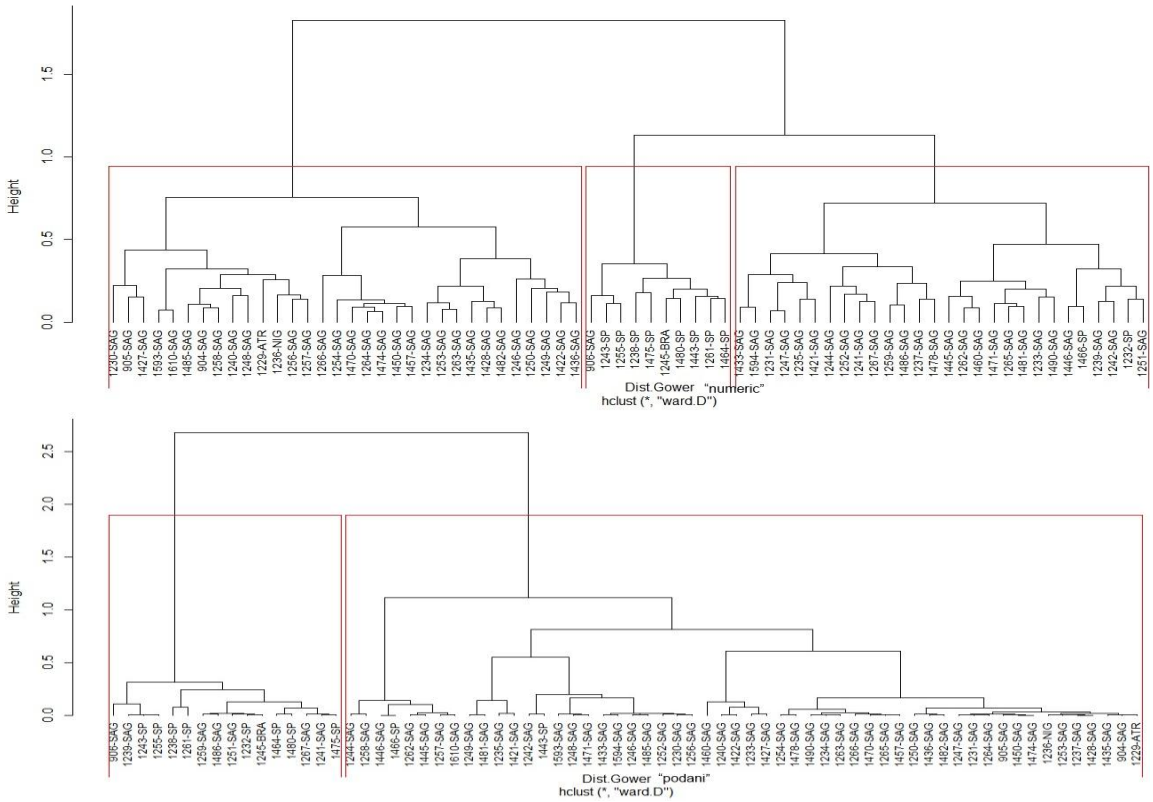


Figura 2. Dendrogramas obtenidos al aplicar el método de Ward con la métrica de Gower, variante “numeric” y “podani” sobre la matriz de datos de *Xanthosoma* spp.

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 existían antecedentes sobre estas colecciones, sobre un mejor resultado en el análisis de variables cuantitativas codificadas o no, antes de aplicar alguna medida de distancia.

CONCLUSIONES

En este estudio se demuestra que las dos variantes de tratamiento de los datos cuantitativos sobre estos germoplasmas varían su respuesta en los agrupamientos de acuerdo al método de aglomeración que se use, aunque con la variante dos se obtienen de forma general mejores resultados de los coeficientes aglomerativo y de correlación cofenético. Por lo cual, se recomienda valorar ambas variantes de tratamiento a variables cuantitativas para cada base de datos que se vaya a analizar, para así encontrar la solución más sólida de acuerdo a los resultados que se esperan de los datos analizados.

BIBLIOGRAFÍA

CROSSA, J. and J. FRANCO. 2004. Statistical methods for classifying genotypes. *Euphytica*, 137:19-37.

GOWER, J.C. 1971. A general coefficient of similarity and some of its properties. *Biometrics*, 27:857-871.

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.

LALIBERTÉ, E. and P. LEGENDRE. 2010. A distance-based framework for measuring functional diversity from multiple traits. *Ecology*, 91:299-305.

LALIBERTÉ, E. and B. SHIPLEY. 2011. FD: measuring functional diversity from multiple traits, and other tools for functional ecology. R package version 1.0-11.

MANTEL, N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2):209-220.

MILIÁN, M.J. 2008. Caracterización de la variabilidad de los cultivares de la colección cubana de germoplasma del género *Xanthosoma* (Araceae). Tesis presentada en opción al grado científico de Doctor en Ciencias Biológicas, Universidad de La Habana, Ciudad de La Habana,123 p.

MURTAGH, F. and P. LEGENDRE. 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification*, 31:274-295.

NÚÑEZ, M.; A. VILLARROYA and J.M. OLLER. 2003. Minimum distance probability discriminant analysis for mixed variables. *Biometrics,* 59:248-253.

PODANI, J. 1999. Extending Gower's general coefficient of similarity to ordinal characters. *Taxon,* 48:331-340.

R CORE TEAM. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Disponible en: https://www.R-project.org/. Consultado el 14/9/2017.

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 F.J. ROHLF. 1962. The comparisons of dendrograms by objective methods. *Taxon*,11:33-40.

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