Análisis de diferentes métodos multivariados para la clasificación de las colecciones cubanas de germoplasma de plátanos (*Musa* spp.) y malanga (*Xanthosoma* spp.)

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

1. Instituto de Investigaciones de Viandas Tropicales (INIVIT), Cuba.
2. Universidad Central "Marta Abreu" de Las Villas (UCLV), Cuba.

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 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, se ha difundido recientemente el empleo de formas no lineales del análisis de componentes principales, o de análisis de componentes principales incluyendo variables categóricas.

El Análisis de Componentes Principales (PCA) con escalamiento óptimo (Gifi, 1990; Meulman *et al*., 2004), también conocido como CatPCA (SPSS, 2003) es el equivalente no lineal del PCA, el cual puede manipular variables categóricas. Además, reduce la dimensionalidad de un conjunto de datos y convierte las variables categóricas en variables cuantitativas usando escalamiento óptimo. El Análisis de Componentes Principales No Lineal (NLPCA) puede llevarse a cabo con el paquete Homals en R (R Development Core Team2009).

El objetivo del presente trabajo es clasificar un grupo de genotipos de plátanos (*Musa* spp.) y malanga (*Xanthosoma* spp.) a partir de la evaluación de la variables cualitativas y cuantitativas, de los rasgos que los caracterizan pertenecientes al genofondo cubano a partir de la estrategia de integrar variables categóricas mediante la aplicación del Análisis NLPCA implementado en el paquete *Homals* sobre la base del lenguaje de programación R y un análisis de conglomerado jerárquico que permita la integración de las variables cuantitativas para establecer 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 genotipos de de plátanos y malanga (*Xanthosoma* spp.) de la colección cubana de germoplasma, que se conserva en el Instituto de Investigaciones de Viandas Tropicales (INIVIT).

La colección de análisis de plátanos contiene 131 accesiones, con 36 variables cualitativas (nominales y ordinales) y 6 variables cuantitativas; y la de malanga *Xanthosoma* 71 accesiones donde se evaluaron 9 variables cualitativas (nominales y ordinales) y 9 variables cuantitativas incluidas en el Sistema de Descriptores Mínimo. De esta forma quedó conformada dos matrices de datos por colección, una con las variables cualitativas y otra con las cuantitativas, a la

primera se le realiza un Análisis de Componentes Principales No Lineal (NLPCA) con la función *homals*. Esta técnica permite la combinación de variables categóricas nominal y ordinal reduciendo a un menor número el banco de datos, perdiendo la menor cantidad de información posible.
Las nuevas variables se unen a la matriz de variables cuantitativas y se les realiza un análisis de cluster jerárquico usando el método de aglomeración de mejor respuesta a ambas colecciones y determinar el mejor coeficiente aglomerativo entre Ward y UPGMA, y como medida de disimilitud, la distancia de mejor coeficiente divisivo entre euclidiana y manhattan.
Para procesar la información se utilizó un lenguaje de programación, orientado a objetos denominado R; el cual es un conjunto de programas integrados para análisis estadísticos y gráficos. R es un *software* libre, por lo cual la implementación de las técnicas de clasificación en este lenguaje le dará mayor potencialidad e independencia.
Del paquete cluster de R se aplicó la función *agnes* para determinar el mejor coeficiente aglomerativo y la función *diana* determinar el coeficiente divisivo.

Resultados y Discusión
Al aplicar el Análisis de Componentes Principales No Lineal (NLPCA) con la función *homals* del paquete Homals (Leeuw, 2009) de R a las base de datos de variables cualitativas obtenidas a partir de la caracterización morfoagronómica de la colección de plátanos y malanga (*Xanthosoma* spp.) se tomaron las cuatro primeras dimensiones que acumulan más del 80% de la variabilidad. La elección de los factores se realizó de tal forma que el primero recoja la mayor proporción posible de la variabilidad original; el segundo factor debe recoger la máxima variabilidad posible no recogida por el primero, y así sucesivamente.
Con las cuatro primeras dimensiones principales obtenidas del análisis NLPCA, más las variables cuantitativas estandarizadas se aplicó la función *diana* para determinar el mayor coeficiente divisivo entre la distancia euclidiana y manhattan, donde se puede observar en la tabla 1 que la mejor medida de disimilitud es la distancia manhattan para ambas bases de datos.

Tabla 1. Coeficiente divisivo de la función *diana*.

Medidas disimilitud *Xanthosoma*		**Medidas disimilitud plátanos**	
Euclidean	Manhattan	Euclidean	Manhattan
0.7895129	0.8258541	0.8787804	0.9149492

Al aplicar la función *agnes* para determinar el mejor coeficiente aglomerativo entre los métodos de aglomeración de Ward y Average se determino un mejor comportamiento para el método de Ward combinado con la distancia de manhattan (Tabla 2 y 3).

Tabla 2. Coeficiente aglomerativo función *agne* para los genotipos de Xanthosoma.

Ward		**Average**	
Euclidean	Manhattan	Euclidean	Manhattan
0.8694898	0.8982666	0.7641314	0.7931706

Tabla 3. Coeficiente aglomerativo función *agne* para los genotipos de plátanos

Ward		**Average**	
Euclidean	Manhattan	Euclidean	Manhattan
0.9626573	0.9740902	0.7950644	0.8468477

Teniendo en cuenta los resultados obtenidos podemos afirmar que el empleo de la estrategia de combinar la función *homals* con los datos cualitativos y posteriormente con las dimensiones más la variables cuantitativas estandarizadas hacer un análisis de conglomerados con el empleo del método Ward con la distancia manhattan como medida de disimilitud constituye una herramienta fundamental para el análisis morfológico y la correcta identificación de los materiales utilizados en este estudio. 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)

Esta estrategia es aplicada por primera vez para la clasificación taxonómica de las colecciones cubanas de germoplasma de plátanos (*Musa* spp.) y malanga (*Xanthosoma* spp.) y se confirmó que esta propuesta permite determinar semejanzas y diferencias entre individuos a clasificar perteneciente a estos genofondos.

Bibliografía

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

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

Gifi, A. *Nonlinear multivariate analysis*. Wiley, Chichester. 1990.

SPSS. CATPCA algorithm, Retrieved May 20, 2004. http://support.spss.com/. 2003.

R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. 2009.

Leeuw, J. de and P. Mair. Homogeneity Analysis in R: the Package homals. Journal of Statistical Software, (in press), 2009.