

ANÁLISIS DE AGRUPAMIENTO PARA LA CARACTERIZACIÓN DEL GERMOPLASMA DE ÑAME (*Dioscorea* spp.) MEDIANTE CONGLOMERADOS PARTICIONALES

Osmany Molina Concepción*, Yuniel Rodríguez García, Yaselis Guillen López, Carmen C. Pons Pérez

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

* Autor para correspondencia: taxonumeric@inivit.cu, <https://orcid.org/0000-0003-4199-6652>

Recibido: 9 de abril de 2024; Aceptado: 12 de junio de 2024

RESUMEN

En el presente trabajo se evalúan diferentes métodos de agrupamiento particionales implementados en la función *kmeans*, unido a los métodos *Pam* y *Clara*. El objetivo es identificar patrones naturales dentro del germoplasma de ñame conservado en el Inivit. La matriz de datos está conformada por variables mixtas donde se integran las variables cualitativas (nominales y ordinales) y las cuantitativas según sus características morfológicas. A partir de la matriz de distancia se combinan las soluciones para obtener una estructura consenso, en aras de lograr un mejor rendimiento y calidad de los resultados alcanzados por los algoritmos individuales. Para los análisis se utilizaron funciones implementadas sobre la base del lenguaje de programación R. En esta investigación se demostró que los algoritmos particionales estudiados, así como la combinación de agrupamientos para lograr una partición consenso son técnicas efectivas en la conformación de estructuras consistentes del germoplasma de ñame. Este enfoque puede ser extrapolado a otros estudios de clasificación en germoplasma vegetal, destacando su utilidad y aplicabilidad en investigaciones similares.

Palabras clave: algoritmo K-medias, conglomerado particional, consenso, germoplasma

CLUSTERING ANALYSIS FOR THE CHARACTERIZATION OF YAM GERMPLASM (*Dioscorea* spp.) THROUGH PARTITIONING CLUSTERS

ABSTRACT

In this work, different partitioning clustering methods implemented in the *kmeans* function are evaluated, along with the *Pam* and *Clara* methods. The objective is to identify natural patterns within the yam germplasm preserved at Inivit. The data matrix consists of mixed variables where qualitative variables (nominal and ordinal) and quantitative variables are integrated according to their morphological characteristics. From the distance matrix, solutions are combined to obtain a consensus structure, aiming to achieve better performance and quality of the results obtained by the individual algorithms. Functions implemented in the R programming language were used for the analyses. This research demonstrated that the studied partitioning algorithms, as well as the combination of clusters to achieve a consensus partition, are effective techniques in forming consistent structures of yam germplasm. This approach can be

extrapolated to other classification studies in plant germplasm, highlighting its utility and applicability in similar research.

Keywords: K-means algorithm, partitioning clustering, consensus, germplasm

INTRODUCCIÓN

Los bancos de germoplasma son esenciales para los programas de mejoramiento genético, si se desea desarrollar este último sobre bases sólidas y obtener un avance consistente en todos los lineamientos de mejoramiento que se persigan (Castillo, 1991). A lo largo de la historia, los recursos fitogenéticos han contribuido a la estabilidad de los agroecosistemas y han proporcionado la materia prima fundamental para el surgimiento del fitomejoramiento científico moderno.

La caracterización morfo-agronómica de un banco de germoplasma es normalmente la forma más accesible de conocer y cuantificar la diversidad genética de los cultivos, incluida la batata (Ritschel y Huamán, 2002; Amoatey *et al.*, 2015; Ahiakpa *et al.*, 2013). Ésta consiste en establecer una identidad para cada entrada (accesión) a través del conocimiento de una serie de informaciones que permita estudiar la variabilidad genética de cada muestra (Daros *et al.*, 2002).

Algunos investigadores han utilizado caracteres morfológicos y agronómicos para identificar accesiones duplicadas (Elameen *et al.*, 2011; Tairo *et al.*, 2008). Los caracteres evaluados arrojan datos tanto cuantitativos como cualitativos. La caracterización morfológica de los datos cuantitativos, aunque más bien debería llamarse caracterización morfométrica, es la más usada en la actualidad (Núñez y Escobedo, 2011). En el caso de la evaluación de las accesiones, se trata de un tipo de caracterización que registra la expresión de la mayoría de los caracteres poligénicos que solo se expresan como reacción a estímulos del medio ambiente, como rendimiento y productividad agronómica (Franco e Hidalgo, 2003; Amin y Single, 2010). Sin embargo, la mayoría de las colecciones son mal aprovechadas por los mejoradores, debido a la falta de información útil sobre las accesiones (Lebot, 2010), por lo que se debe hacer una exhaustiva selección en base a los resultados de caracterización y/o evaluación sobre las accesiones, para identificar caracteres deseables.

Una vez que se reúne la información se procede a su análisis estadístico mediante métodos multivariados de descripción, ordenación o clasificación adecuados.

La clasificación puede considerarse una de las tareas fundamentales de la ciencia, es una acción propia del hombre dado que los objetos no están ordenados por sí mismos; toda clasificación se considera artificial ya que el hombre, como resultado de su capacidad de razonar e interpretar el mundo que le rodea, es el que asigna el orden a las entidades según sus ideas, necesidades y criterios (Kragh, 1989).

Los algoritmos de agrupación varían entre sí por el mayor o menor grado de reglas heurísticas que utilizan e, inversamente, por el nivel de procedimientos formales involucrados. 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.

En el Instituto de Investigaciones en Viandas Tropicales (Inivit) se mantiene desde 1967, una colección de germoplasma de ñame del género *Dioscorea* procedente de colectas, de introducciones foráneas y del programa de mejoramiento genético del cultivo. La experiencia adquirida en Cuba sobre el cultivo, la riqueza genética conservada en la Isla, constituyen motivaciones importantes para profundizar en el

conocimiento de tan importante género de plantas hacia el objetivo de garantizar el manejo sostenible del patrimonio genético existente, lo que incluye un mayor acercamiento a la adecuada ubicación taxonómica de todos los cultivares y a un mejor conocimiento de sus características y potencial genético, que permita lograr su empleo efectivo en los programas de mejoramiento; así como el perfeccionamiento de las estrategias de conservación de los recursos disponibles.

Además, dadas las circunstancias actuales de escasez de recursos financieros y materiales se hace necesario conservar y estudiar solo aquellos materiales que sean únicos, es decir, que no se encuentren duplicados y, de esta forma facilitar que se profundice en los estudios morfológicos, bioquímicos y moleculares.

Para el mejoramiento genético de plantas, en cualquiera de sus formas, es indispensable partir de la mejor clasificación taxonómica posible ya que este conocimiento, optimiza los procesos de cruzamientos y los hace más eficientes.

En el Inivit, se han realizados diversos estudios de técnicas estadísticas no supervisada de clasificación jerárquica y de reducción de dimensionalidad sobre el germoplasma de ñame (*Dioscorea* spp.) (Milián, 2008; Molina, 2013a; 2013b; 2015; 2017; 2018; Milián, 2018), a partir de los cuales se pretende desarrollar el presente trabajo con el objetivo de contribuir a una mejor clasificación específica de este germoplasma. Se pretende ampliar el estudio con la aplicación de técnicas estadísticas no supervisada de clasificación con métodos particionales, que integren las diferentes variables que componen las accesiones, lo cual es indispensable para obtener una mayor comprensión sobre la distribución de las accesiones dentro del germoplasma del género *Dioscorea* que se conservan en Cuba.

MATERIALES Y MÉTODOS

Para realizar la investigación, se usaron datos procedentes de un estudio de los diferentes genotipos de una colección de ñame (*Dioscorea* spp.) (85 accesiones donde se evaluaron 43 variables cualitativas y nueve variables cuantitativas incluidas en el Sistema de Descriptores Mínimos del banco de germoplasma, que se conserva en el Instituto de Investigaciones de Viandas Tropicales (Inivit).

Las 85 accesiones están distribuidas en 72 accesiones de la especie *D. alata* (ALA), tres de la especie *D. bulbifera* (BUL), cinco de la especie *D. cayenensis* (CAY), tres de la especie *D. rotundata* (RTA) y dos de la especie *D. esculenta* (ESC) (Milián, 2008; 2018).

En un primer paso se integraron las variables cualitativas transformadas a través del análisis de correspondencia múltiple (Tenenhaus, 1985) con la función *MCA()* descrita en el paquete *FactoMineR*, con las cuantitativas normalizadas para lo cual se usó la función *data.Normalization(n1)* del paquete *clusterSim* (Walesiak, 2017).

Como medidas para evaluar las diferencias y similitudes entre accesiones se usó la distancia Euclidiana implementada en la función *dist()* descrita en el paquete *stats*.

A pesar que la investigación está centrada en una clasificación no supervisada, para estimar el número de agrupaciones se usó la identificación ya establecida de algunas especies basado en investigaciones de los curadores del germoplasma en estudio de malanga.

Se usaron los métodos de aglomeración particionales *k-means* descrito en el paquete *stats* en la función *kmeans()* con los cuatro algoritmos: Forgy (Forgy, 1965), Hartigan-Wong (Hartigan, 1979), Lloyd (Lloyd, 1982) y MacQueen (MacQueen, 1967); que forma

parte de la librería básica de R que se instala por defecto. Y los métodos *Pam* y *Clara* implementados en el paquete *cluster* (Maechler *et al.*, 2019) en las funciones *pam()* y *clara()*.

Una vez obtenidos los resultados e interpretados es conveniente establecer un criterio de validación de la coherencia dentro de los conglomerados mediante los coeficientes devueltos con el Ancho de la Silueta (Rousseeuw, 1987; Kaufman y Rousseeuw, 1990) a través de la función *silhouette()* implementada en el paquete *cluster*. El índice proporciona una representación gráfica que da una medida de la similitud de una accesión con su propio conglomerado (cohesión) en comparación con otros conglomerados (separación). El valor de la silueta varía de -1 a 1, donde un valor alto muestra que el objeto está bien adaptado a su propio conglomerado y un valor bajo indica que está mal adaptado a los conglomerados vecinos.

Con el objetivo de lograr una mejor calidad de los resultados alcanzados por los algoritmos individuales y compensar posibles errores cometidos por un algoritmo individual se combinaron los resultados de los diferentes algoritmos de aglomeración particionales con la función *cl_ensemble()* y finalmente se obtiene las particiones consenso, las cual es capaz de combinar toda la información existente de los diferentes resultados individuales, se obtuvo con la función *cl_consensus()* ambas descritas en el paquete *clue* utilizando siete de los métodos disponibles en esta función ("GV1", "SM", "SE", "GV3", "soft/symdiff" y "DWH", "HE") para calcular las particiones.

Para medir el desempeño de la clasificación de cada uno de los métodos particionales y el consensos se usó las matrices de confusión (Kohavi y Provost, 1998), tratadas posteriormente por (Provost y Fawcett, 2001; Fawcett, 2003; Weiss y Provost, 2003; Maimon y Rokach, 2005) a través de la función *confusion.matrix()* implementadas en el paquete *clv* y el área bajo la curva ROC (AUC) (Swets, 1988), tratados más profundamente *a posteriori* por el propio Swets y otros autores (Swets *et al.*, 2000a; 2000b) con la *multiclass.roc()* descrita en el paquete *pROC*.

Las matrices de confusión contienen información acerca de los valores reales y las clasificaciones predichas hechas. El desempeño de un tal sistema es usualmente evaluado usando los datos en dicha matriz. Las curvas ROC (en inglés *Receiver Operating Characteristics*) conjuntamente con las matrices de confusión, constituyen otra manera de examinar el desempeño de una clasificación. El área bajo la curva ROC (AUC) se puede usar como una medida de la exactitud en muchas aplicaciones. Entre mayor área bajo la curva mejor clasificación.

El software utilizado para realizar los experimentos presentes en esta investigación es R versión 3.6.1 (R Development Core Team, 2019).

RESULTADOS Y DISCUSIÓN

Atendiendo a la naturaleza mixta de los datos las variables cualitativas (nominales y ordinales) se transformaron a través del análisis de correspondencia múltiple, se toman 25 dimensiones con un porcentaje de varianza mayor que uno, que representan el 87.076 % de la variabilidad total. Las 9 variables cuantitativas se estandarizaron y, posteriormente, se unieron ambos resultados en una sola matriz de datos continuos a la cual se le aplicó la distancia Euclidiana. Sorensen (1948) y Gatna (2004) han demostrado que al menos el 70 % de la inercia total explicada es el valor de corte más adecuado.

A partir de la ejecución del método de k-medias con los cuatro algoritmos que emplea para realizar las agrupaciones y los métodos *Pam* y *Clara* con la distancia Euclidiana, con cinco grupos que corresponden con el número de especies presente en la base de datos, en un análisis global se puede observar en la tabla 1 la distribución de las accesiones en los cinco conglomerados. Las tres accesiones de las especies de *D. bulbifera* (BUL) siempre comparten el mismo conglomerado, excepto en los algoritmos *Lloyd* y *Clara*, igualmente cuatro accesiones (1278, 1279, 1280 y 1281) de la especie *D. cayenensis* (CAY) comparten el mismo clúster junto a la accesión 1318 de la especie *D. rotundata*, menos en el algoritmo *Clara*. Las accesiones 1275 y 1491 de la especie *D. esculenta* siempre permanecen en el mismo conglomerado independientemente del algoritmo de agrupamiento, lo cual muestra una medida de similitud entre ellas. Las accesiones 1282 y 1311 de la especie *D. rotundata*, y la accesión 1310 de la especie *D. cayenensis*, nunca se unen en un cluster a las semejantes de su especie, independientemente del método de aglomeración, por lo cual se puede afirmar que no tienen una afinidad morfológica demostrada a través de estos métodos de clasificación.

Tabla 1. Salida en consola de accesiones por conglomerados de los algoritmos de aglomeración particionales.

Algoritmos	Accesiones por conglomerados
<i>Hartigan</i>	1: {903-ALA, 1270-ALA, 1285-ALA, 1287-ALA, 1292-ALA, 1294-ALA, 1295-ALA, 1297-ALA, 1300-ALA, 1301-ALA, 1302-ALA, 1307-ALA, 1308-ALA, 1426-ALA, 1444-ALA, 1456-ALA, 1469-ALA, 1493-ALA, 1617-ALA, 1618-ALA, 1829-ALA, 1832-ALA} 2: {1283-ALA, 1284-ALA, <u>1286-BUL</u>, 1434-ALA, 1447-ALA, 1473-ALA, 1487-ALA, 1492-ALA, 1498-ALA, <u>1627-BUL</u>, <u>1744-BUL</u>} 3: {<u>1275-ESC</u>, <u>1282-RTA</u>, 1303-ALA, 1305-ALA, <u>1310-CAY</u>, <u>1311-RTA</u>, 1461-ALA, 1467-ALA, 1488-ALA, <u>1491-ESC</u>} 4: {1272-ALA, 1274-ALA, 1276-ALA, 1277-ALA, 1291-ALA, 1293-ALA, 1296-ALA, 1298-ALA, 1304-ALA, 1309-ALA, 1312-ALA, 1313-ALA, 1317-ALA, 1320-ALA, 1423-ALA, 1432-ALA, 1452-ALA, 1472-ALA, 1499-ALA, 1819-ALA, 1822-ALA, 1828-ALA, 1833-ALA} 5: {1271-ALA, <u>1278-CAY</u>, <u>1279-CAY</u>, <u>1280-CAY</u>, <u>1281-CAY</u>, 1288-ALA, 1289-ALA, 1290-ALA, 1306-ALA, 1315-ALA, 1316-ALA, <u>1318-RTA</u>, 1319-ALA, 1442-ALA, 1449-ALA, 1453-ALA, 1463-ALA, 1497-ALA, 1613-ALA}
<i>Lloyd</i>	1: {1272-ALA, 1274-ALA, 1277-ALA, 1291-ALA, 1293-ALA, 1296-ALA, 1304-ALA, 1312-ALA, 1320-ALA, 1432-ALA, 1452-ALA, 1472-ALA, 1819-ALA, 1822-ALA, 1828-ALA} 2: {1271-ALA, <u>1278-CAY</u>, <u>1279-CAY</u>, <u>1280-CAY</u>, <u>1281-CAY</u>, 1288-ALA, 1289-ALA, 1290-ALA, 1305-ALA, 1306-ALA, 1315-ALA, 1316-ALA, <u>1318-RTA</u>, 1319-ALA, 1442-ALA, 1449-ALA, 1453-ALA, 1461-ALA, 1463-ALA, 1497-ALA, 1613-ALA} 3: {1276-ALA, 1283-ALA, 1284-ALA, 1298-ALA, 1309-ALA, 1313-ALA, 1317-ALA, 1423-ALA, 1434-ALA, 1473-ALA, 1499-ALA, 1833-ALA} 4: {<u>1275-ESC</u>, <u>1282-RTA</u>, <u>1286-BUL</u>, 1303-ALA, <u>1310-CAY</u>, <u>1311-RTA</u>, 1447-ALA, 1467-ALA, 1487-ALA, <u>1491-ESC</u>, 1492-ALA, 1498-ALA, <u>1627-BUL</u>, <u>1744-BUL</u>} 5: {903-ALA, 1270-ALA, 1285-ALA, 1287-ALA, 1292-ALA, 1294-ALA, 1295-ALA, 1297-ALA, 1300-ALA, 1301-ALA, 1302-ALA, 1307-ALA, 1308-ALA,

	1426-ALA, 1444-ALA, 1456-ALA, 1469-ALA, 1488-ALA, 1493-ALA, 1617-ALA, 1618-ALA, 1829-ALA, 1832-ALA}
Forgy	1: {<u>1275-ESC, 1426-ALA, 1491-ESC, 1493-ALA</u>} 2: {1271-ALA, <u>1282-RTA</u>, 1290-ALA, 1303-ALA, 1305-ALA, 1306-ALA, <u>1310-CAY, 1311-RTA, 1318-RTA</u>, 1442-ALA, 1449-ALA, 1461-ALA, 1463-ALA, 1467-ALA, 1488-ALA} 3: {1272-ALA, 1274-ALA, 1277-ALA, <u>1278-CAY, 1280-CAY, 1281-CAY</u>, 1293-ALA, 1296-ALA, 1304-ALA, 1312-ALA, 1320-ALA, 1423-ALA, 1452-ALA, 1472-ALA, 1499-ALA, 1819-ALA, 1822-ALA, 1828-ALA, 1833-ALA} 4: {1276-ALA, <u>1279-CAY</u>, 1283-ALA, 1284-ALA, <u>1286-BUL</u>, 1288-ALA, 1289-ALA, 1291-ALA, 1298-ALA, 1309-ALA, 1315-ALA, 1316-ALA, 1319-ALA, 1434-ALA, 1447-ALA, 1453-ALA, 1473-ALA, 1487-ALA, 1492-ALA, 1497-ALA, 1498-ALA, 1613-ALA, <u>1627-BUL, 1744-BUL</u>} 5: {903-ALA, 1270-ALA, 1285-ALA, 1287-ALA, 1292-ALA, 1294-ALA, 1295-ALA, 1297-ALA, 1300-ALA, 1301-ALA, 1302-ALA, 1307-ALA, 1308-ALA, 1313-ALA, 1317-ALA, 1432-ALA, 1444-ALA, 1456-ALA, 1469-ALA, 1617-ALA, 1618-ALA, 1829-ALA, 1832-ALA}
MacQueen	1: {903-ALA, 1270-ALA, 1285-ALA, 1287-ALA, 1294-ALA, 1295-ALA, 1297-ALA, 1300-ALA, 1301-ALA, 1302-ALA, 1307-ALA, 1444-ALA, 1456-ALA, 1469-ALA, 1493-ALA, 1617-ALA, 1618-ALA, 1832-ALA} 2: {1272-ALA, 1274-ALA, 1276-ALA, 1277-ALA, 1283-ALA, 1291-ALA, 1293-ALA, 1296-ALA, 1298-ALA, 1304-ALA, 1309-ALA, 1312-ALA, 1313-ALA, 1317-ALA, 1320-ALA, 1423-ALA, 1432-ALA, 1452-ALA, 1472-ALA, 1473-ALA, 1499-ALA, 1819-ALA, 1822-ALA, 1828-ALA, 1833-ALA} 3: {1271-ALA, <u>1278-CAY, 1279-CAY, 1280-CAY, 1281-CAY</u>, 1288-ALA, 1289-ALA, 1290-ALA, 1306-ALA, 1315-ALA, 1316-ALA, <u>1318-RTA</u>, 1319-ALA, 1442-ALA, 1449-ALA, 1453-ALA, 1463-ALA, 1497-ALA, 1613-ALA} 4: {<u>1275-ESC</u>, 1292-ALA, 1305-ALA, 1308-ALA, <u>1310-CAY</u>, 1426-ALA, 1447-ALA, 1461-ALA, 1488-ALA, <u>1491-ESC</u>, 1829-ALA} 5: {<u>1282-RTA</u>, 1284-ALA, <u>1286-BUL</u>, 1303-ALA, <u>1311-RTA</u>, 1434-ALA, 1467-ALA, 1487-ALA, 1492-ALA, 1498-ALA, <u>1627-BUL, 1744-BUL</u>}
Pam	1: {903-ALA, 1270-ALA, 1285-ALA, 1287-ALA, 1292-ALA, 1294-ALA, 1295-ALA, 1297-ALA, 1300-ALA, 1301-ALA, 1302-ALA, 1307-ALA, 1308-ALA, 1313-ALA, 1317-ALA, 1426-ALA, 1432-ALA, 1444-ALA, 1456-ALA, 1469-ALA, 1493-ALA, 1617-ALA, 1618-ALA, 1829-ALA, 1832-ALA} 2: {1271-ALA, <u>1278-CAY, 1279-CAY, 1280-CAY, 1281-CAY</u>, 1288-ALA, 1289-ALA, 1290-ALA, 1306-ALA, 1315-ALA, 1316-ALA, <u>1318-RTA</u>, 1319-ALA, 1442-ALA, 1449-ALA, 1453-ALA, 1463-ALA, 1492-ALA, 1497-ALA} 3: {1272-ALA, 1274-ALA, 1276-ALA, 1277-ALA, 1284-ALA, 1291-ALA, 1293-ALA, 1296-ALA, 1298-ALA, 1304-ALA, 1309-ALA, 1312-ALA, 1320-ALA, 1423-ALA, 1434-ALA, 1452-ALA, 1472-ALA, 1499-ALA, 1613-ALA, 1819-ALA, 1822-ALA, 1828-ALA, 1833-ALA} 4: {<u>1275-ESC</u>, 1283-ALA, <u>1286-BUL</u>, 1447-ALA, 1473-ALA, 1487-ALA, <u>1491-ESC</u>, 1498-ALA, <u>1627-BUL, 1744-BUL</u>} 5: {<u>1282-RTA</u>, 1303-ALA, 1305-ALA, <u>1310-CAY, 1311-RTA</u>, 1461-ALA, 1467-ALA, 1488-ALA}
Clara	1: {903-ALA, 1270-ALA, 1283-ALA, 1285-ALA, 1287-ALA, 1292-ALA, 1294-ALA, 1295-ALA, 1297-ALA, 1300-ALA, 1301-ALA, 1302-ALA, 1307-ALA, 1313-ALA, 1317-ALA, 1432-ALA, 1444-ALA, 1456-ALA, 1617-ALA, 1832-ALA} 2: {1271-ALA, <u>1278-CAY, 1279-CAY, 1280-CAY, 1281-CAY, 1282-RTA</u>,

- 1288-ALA, 1289-ALA, 1290-ALA, 1305-ALA, 1306-ALA, **1311-RTA**, 1315-ALA, 1316-ALA, **1318-RTA**, 1442-ALA, 1449-ALA, 1453-ALA, 1463-ALA, 1467-ALA, 1487-ALA, 1492-ALA, 1497-ALA, 1498-ALA, **1627-BUL**, **1744-BUL** }
- 3: {1272-ALA, 1274-ALA, 1277-ALA, 1296-ALA, 1303-ALA, 1319-ALA, 1461-ALA, 1472-ALA, 1499-ALA}
- 4: {**1275-ESC**, **1286-BUL**, 1308-ALA, **1310-CAY**, 1426-ALA, 1447-ALA, 1469-ALA, 1488-ALA, **1491-ESC**, 1493-ALA, 1618-ALA, 1829-ALA}
- 5: {1276-ALA, 1284-ALA, 1291-ALA, 1293-ALA, 1298-ALA, 1304-ALA, 1309-ALA, 1312-ALA, 1320-ALA, 1423-ALA, 1434-ALA, 1452-ALA, 1473-ALA, 1613-ALA, 1819-ALA, 1822-ALA, 1828-ALA, 1833-ALA}

El objetivo general perseguido por las técnicas de clustering consiste en identificar grupos o clusters compactos, pero también se hace necesario la evaluación para medir la eficiencia de las técnicas de clustering aplicado. Para el caso de evaluación de clusters en nuestro trabajo utilizaremos el concepto de silueta de un clúster y la entrega de sus coeficientes de medición. Así pues, se tiene que mientras más cercana este la silueta del individuo a 1 (eje Y), quiere decir que el miembro del grupo tiene mayor afinidad al clúster asignado. Por el contrario, si la silueta del miembro del clúster está cercana a 0 (eje Y), quiere decir que tiene afinidad morfológica con otro grupo. Por otra parte, si la silueta del individuo es menor a 0 (eje Y), la probabilidad de que el miembro del clúster pertenezca a otro grupo es muy alta.

Teniendo en cuenta los resultados obtenidos por los métodos de agrupamiento, es conveniente aglutinarlos en aras de lograr una mejor calidad de los resultados alcanzados por los algoritmos individuales, hacer un conjunto de particiones diversas es recomendable, ya que mientras más variado es el conjunto de particiones, mayor es la información disponible en la etapa de combinación.

La combinación de agrupamientos para lograr una partición consenso se realizó con el algoritmo “soft/symdiff” de los siete disponibles para calcular los conglomerados consenso en los algoritmos particionales, ya que fue el de mayor área bajo la curva (*Multi-class area under the curve*: 0.8694). En la tabla 2 se puede observar la distribución de las accesiones por clústeres, lográndose una mejor clasificación por especies, aunque no hay cluster definidos, únicos para cada especie.

Tabla 2. Salida en consola de las accesiones por conglomerados, método consenso con el algoritmo “soft/symdiff”.

Algoritmo consenso	Accesiones por conglomerados
SV3	1: { <u>1286-BUL</u> , 1447-ALA, 1487-ALA, 1492-ALA, 1498-ALA, <u>1627-BUL</u> , <u>1744-BUL</u> }
	2: {903-ALA, 1270-ALA, 1285-ALA, 1287-ALA, 1292-ALA, 1294-ALA, 1295-ALA, 1297-ALA, 1300-ALA, 1301-ALA, 1302-ALA, 1307-ALA, 1308-ALA, 1313-ALA, 1317-ALA, 1426-ALA, 1432-ALA, 1444-ALA, 1456-ALA, 1469-ALA, 1493-ALA, 1617-ALA, 1618-ALA, 1829-ALA, 1832-ALA}
	3: {1271-ALA, <u>1278-CAY</u> , <u>1279-CAY</u> , <u>1280-CAY</u> , <u>1281-CAY</u> , 1288-ALA, 1289-ALA, 1290-ALA, 1305-ALA, 1306-ALA, 1315-ALA, 1316-ALA, <u>1318-RTA</u> , 1319-ALA, 1442-ALA, 1449-ALA, 1453-ALA, 1461-ALA, 1463-ALA, 1497-ALA, 1613-ALA}

4: {1272-ALA, 1274-ALA, 1276-ALA, 1277-ALA, 1291-ALA, 1293-ALA, 1296-ALA, 1298-ALA, 1304-ALA, 1309-ALA, 1312-ALA, 1320-ALA, 1423-ALA, 1452-ALA, 1472-ALA, 1473-ALA, 1499-ALA, 1819-ALA, 1822-ALA, 1828-ALA, 1833-ALA}
5: {**1275-ESC**, **1282-RTA**, 1283-ALA, 1284-ALA, 1303-ALA, **1310-CAY**, **1311-RTA**, 1434-ALA, 1467-ALA, 1488-ALA, **1491-ESC**}

En la figura 1 se puede observar que el ancho de silueta para el consenso es de 0.18, y en la figura 2 se muestran los miembros que conforman cada uno de los cinco conglomerados según similitudes morfológicas.

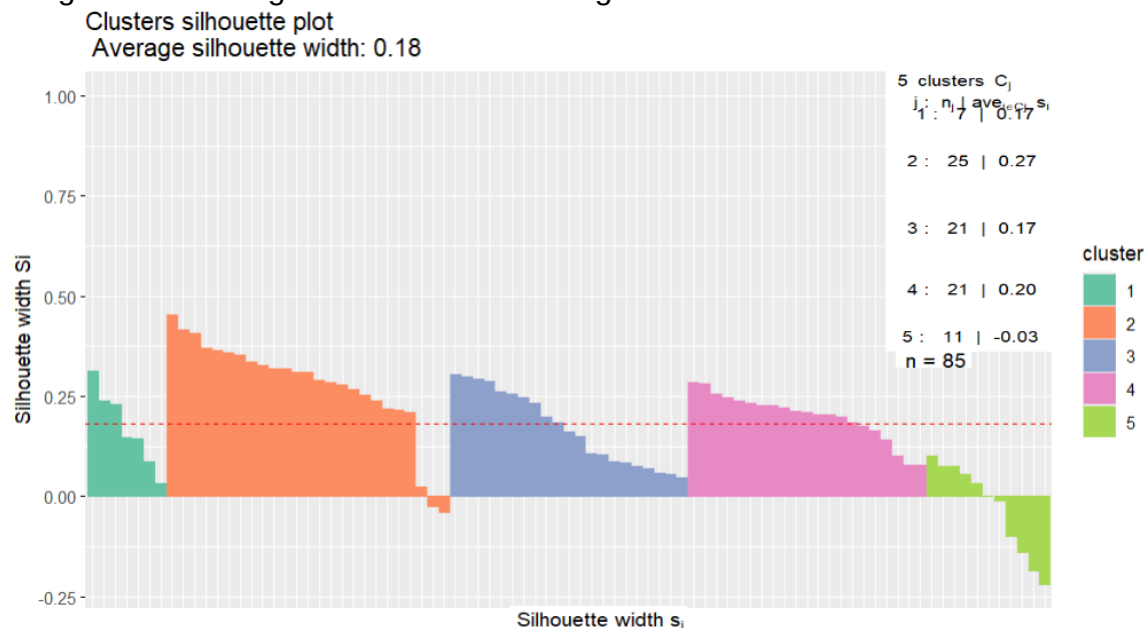


Figura 1. Ancho de silueta. Método consenso con el algoritmo “soft/symdiff”
Cluster plot

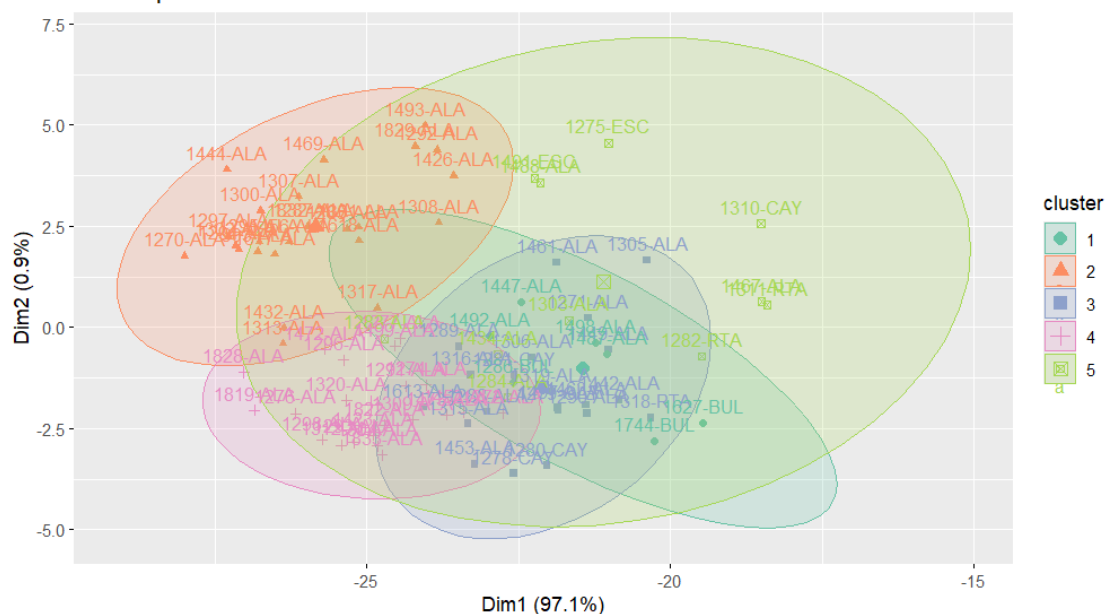


Figura 2. Visualización para cinco clústeres identificando accesiones. Método consenso con el algoritmo “soft/symdiff”

En la tabla 3 se puede hacer una evaluación de la información sobre los coeficientes de silueta individuales, el cluster asignado y el vecino.

Tabla 3. Ancho de silueta de miembros de cada clúster. Método consenso con el algoritmo “soft/symdiff”.

ID.cultivares	cluster	neighbor	sil_width
903-ALA	2	4	0.358
1270-ALA	2	4	0.215
1271-ALA	3	4	0.292
1272-ALA	4	3	0.077
1274-ALA	4	3	0.208
1275-ESC	5	1	-0.011
1276-ALA	4	1	0.202
1277-ALA	4	2	0.174
1278-CAY	3	4	0.076
1279-CAY	3	4	0.200
1280-CAY	3	4	0.160
1281-CAY	3	4	0.057
1282-RTA	5	3	0.032
1283-ALA	5	4	-0.218
1284-ALA	5	4	-0.183
1285-ALA	2	4	0.320
1286-BUL	1	4	0.145
1287-ALA	2	4	0.364
1288-ALA	3	4	0.087
1289-ALA	3	4	0.150
1290-ALA	3	4	0.246
1291-ALA	4	3	0.102
1292-ALA	2	5	0.239
1293-ALA	4	3	0.221
1294-ALA	2	4	0.251
1295-ALA	2	4	0.334
1296-ALA	4	2	0.197
1297-ALA	2	4	0.285
1298-ALA	4	1	0.212
1300-ALA	2	4	0.354
1301-ALA	2	4	0.290
1302-ALA	2	4	0.327
1303-ALA	5	4	0.076
1304-ALA	4	3	0.226
1305-ALA	3	5	0.083
1306-ALA	3	4	0.303
1307-ALA	2	4	0.407
1308-ALA	2	3	0.211
1309-ALA	4	3	0.204
1310-CAY	5	3	0.101
1311-RTA	5	3	0.076
1312-ALA	4	3	0.285
1313-ALA	2	4	-0.038
1315-ALA	3	4	0.103

1316-ALA	3	4	0.183
1317-ALA	2	4	0.024
1318-RTA	3	5	0.231
1319-ALA	3	4	0.070
1320-ALA	4	3	0.233
1423-ALA	4	3	0.254
1426-ALA	2	1	0.310
1432-ALA	2	4	-0.023
1434-ALA	5	4	-0.099
1442-ALA	3	1	0.287
1444-ALA	2	4	0.453
1447-ALA	1	3	0.088
1449-ALA	3	4	0.262
1452-ALA	4	3	0.226
1453-ALA	3	4	0.107
1456-ALA	2	4	0.369
1461-ALA	3	5	0.046
1463-ALA	3	1	0.298
1467-ALA	5	3	0.054
1469-ALA	2	4	0.416
1472-ALA	4	2	0.164
1473-ALA	4	3	0.078
1487-ALA	1	3	0.239
1488-ALA	5	2	-0.137
1491-ESC	5	2	0.000
1492-ALA	1	3	0.032
1493-ALA	2	5	0.217
1497-ALA	3	4	0.255
1498-ALA	1	5	0.146
1499-ALA	4	2	0.184
1613-ALA	3	4	0.055
1617-ALA	2	4	0.279
1618-ALA	2	4	0.267
1627-BUL	1	3	0.229
1744-BUL	1	3	0.313
1819-ALA	4	2	0.247
1822-ALA	4	3	0.280
1828-ALA	4	2	0.141
1829-ALA	2	3	0.308
1832-ALA	2	4	0.319
1833-ALA	4	1	0.238

La tabla cuatro representa la probabilidad de pertenencia de las accesiones a cada uno de los cinco conglomerados. Como se puede observar las accesiones 1286, 1627 y 1744 de la especie *D. bulbifera*; 1275 y 1491 de la especie *D. esculenta*; 1282 y 1311 de la especie *D. rotundata*; 1310 de la especie *D. cayenensis*; 1283, 1284, 1303, 1305, 1434, 1447, 1461, 1467, 1473, 1487, 1488, 1492 y 1498 de la especie *D. alata*, tienen probabilidades compartidas de pertenecer a más de un clusters a la vez.

Tabla 4. Probabilidad de pertenencia de miembros de cada clúster. Método consenso con el algoritmo “soft/symdiff”.

ID.cultivares	[,1]	[,2]	[,3]	[,4]	[,5]
903-ALA	0.000	1.000	0.000	0.000	0.000
1270-ALA	0.000	1.000	0.000	0.000	0.000
1271-ALA	0.000	0.000	1.000	0.000	0.000
1272-ALA	0.000	0.000	0.000	1.000	0.000
1274-ALA	0.000	0.000	0.000	1.000	0.000
1275-ESC	0.443	0.000	0.000	0.000	0.557
1276-ALA	0.000	0.000	0.000	1.000	0.000
1277-ALA	0.000	0.000	0.000	1.000	0.000
1278-CAY	0.000	0.000	1.000	0.000	0.000
1279-CAY	0.000	0.000	1.000	0.000	0.000
1280-CAY	0.000	0.000	1.000	0.000	0.000
1281-CAY	0.000	0.000	1.000	0.000	0.000
1282-RTA	0.405	0.000	0.001	0.000	0.595
1283-ALA	0.420	0.000	0.000	0.000	0.580
1284-ALA	0.382	0.000	0.000	0.002	0.617
1285-ALA	0.000	1.000	0.000	0.000	0.000
1286-BUL	0.586	0.000	0.000	0.000	0.414
1287-ALA	0.000	1.000	0.000	0.000	0.000
1288-ALA	0.000	0.000	1.000	0.000	0.000
1289-ALA	0.000	0.000	1.000	0.000	0.000
1290-ALA	0.000	0.000	1.000	0.000	0.000
1291-ALA	0.000	0.000	0.000	1.000	0.000
1292-ALA	0.000	1.000	0.000	0.000	0.000
1293-ALA	0.000	0.000	0.000	1.000	0.000
1294-ALA	0.000	1.000	0.000	0.000	0.000
1295-ALA	0.000	1.000	0.000	0.000	0.000
1296-ALA	0.000	0.000	0.000	1.000	0.000
1297-ALA	0.000	1.000	0.000	0.000	0.000
1298-ALA	0.000	0.000	0.000	1.000	0.000
1300-ALA	0.000	1.000	0.000	0.000	0.000
1301-ALA	0.000	1.000	0.000	0.000	0.000
1302-ALA	0.000	1.000	0.000	0.000	0.000
1303-ALA	0.119	0.000	0.000	0.000	0.881
1304-ALA	0.000	0.000	0.000	1.000	0.000
1305-ALA	0.063	0.000	0.928	0.000	0.010
1306-ALA	0.000	0.000	1.000	0.000	0.000
1307-ALA	0.000	1.000	0.000	0.000	0.000
1308-ALA	0.000	1.000	0.000	0.000	0.000
1309-ALA	0.000	0.000	0.000	1.000	0.000
1310-CAY	0.321	0.000	0.000	0.000	0.679
1311-RTA	0.405	0.000	0.001	0.000	0.595
1312-ALA	0.000	0.000	0.000	1.000	0.000
1313-ALA	0.000	1.000	0.000	0.000	0.000
1315-ALA	0.000	0.000	1.000	0.000	0.000
1316-ALA	0.000	0.000	1.000	0.000	0.000
1317-ALA	0.000	1.000	0.000	0.000	0.000
1318-RTA	0.000	0.000	1.000	0.000	0.000
1319-ALA	0.000	0.000	1.000	0.000	0.000

1320-ALA	0.000	0.000	0.000	1.000	0.000
1423-ALA	0.000	0.000	0.000	1.000	0.000
1426-ALA	0.000	1.000	0.000	0.000	0.000
1432-ALA	0.000	1.000	0.000	0.000	0.000
1434-ALA	0.382	0.000	0.000	0.002	0.617
1442-ALA	0.000	0.000	1.000	0.000	0.000
1444-ALA	0.000	1.000	0.000	0.000	0.000
1447-ALA	0.503	0.000	0.000	0.000	0.497
1449-ALA	0.000	0.000	1.000	0.000	0.000
1452-ALA	0.000	0.000	0.000	1.000	0.000
1453-ALA	0.000	0.000	1.000	0.000	0.000
1456-ALA	0.000	1.000	0.000	0.000	0.000
1461-ALA	0.048	0.000	0.632	0.000	0.321
1463-ALA	0.000	0.000	1.000	0.000	0.000
1467-ALA	0.421	0.000	0.000	0.000	0.579
1469-ALA	0.000	1.000	0.000	0.000	0.000
1472-ALA	0.000	0.000	0.000	1.000	0.000
1473-ALA	0.026	0.000	0.000	0.960	0.014
1487-ALA	0.627	0.000	0.000	0.000	0.373
1488-ALA	0.091	0.001	0.000	0.000	0.908
1491-ESC	0.443	0.000	0.000	0.000	0.557
1492-ALA	0.701	0.000	0.139	0.000	0.160
1493-ALA	0.000	1.000	0.000	0.000	0.000
1497-ALA	0.000	0.000	1.000	0.000	0.000
1498-ALA	0.627	0.000	0.000	0.000	0.373
1499-ALA	0.000	0.000	0.000	1.000	0.000
1613-ALA	0.000	0.000	1.000	0.000	0.000
1617-ALA	0.000	1.000	0.000	0.000	0.000
1618-ALA	0.000	1.000	0.000	0.000	0.000
1627-BUL	0.627	0.000	0.000	0.000	0.373
1744-BUL	0.627	0.000	0.000	0.000	0.373
1819-ALA	0.000	0.000	0.000	1.000	0.000
1822-ALA	0.000	0.000	0.000	1.000	0.000
1828-ALA	0.000	0.000	0.000	1.000	0.000
1829-ALA	0.000	1.000	0.000	0.000	0.000
1832-ALA	0.000	1.000	0.000	0.000	0.000
1833-ALA	0.000	0.000	0.000	1.000	0.000

Se puede afirmar que el método consenso con el algoritmo “soft/symdiff” representa mejor la estructura de la base de datos de ñame (*Dioscorea* spp.).

CONCLUSIONES

La combinación de agrupamientos para lograr una partición consenso es una técnica efectiva en aras de lograr una mejor calidad de los resultados alcanzados por los algoritmos individuales sobre la clasificación del germoplasma de ñame (*Dioscorea* spp.) y se debe de tener en cuenta como una posible solución para mejorar la clasificación de especies vegetales.

El uso del ancho de silueta para evaluar la clasificación, tanto individual como por conglomerados, es efectiva en la identificación de la afinidad morfológica entre los miembros de un mismo conglomerado y entre éstos.

Se amplía el conocimiento sobre la combinación técnicas de agrupamiento particionales en la clasificación de bancos de germoplasma vegetal.

BIBLIOGRAFÍA

- AMOATEY, H.M.; G.Y. KLU; E.K. QUARTEY; H.A. DOKU; F.L. SOSSAH; M.M. SEGBEFIA and J.K. AHIKPA. 2015. Genetic Diversity Studies in 29 Accessions of Okra 70 (*Abelmoschus* spp L.) Using 13 Quantitative Traits. *Am. J. Exp. Agric.*, 5(3):217-225.
- AHIKPA, J.K.; P.D. KALEDZI; E.B. ADI; S. PEPRAH and H.K. DAPAAH. 2013. Genetic diversity, correlation and path analyses of okra (*Abelmoschus* spp L. Moench) germplasm collected in Ghana. *Int. J. Dev. Sustain.* 2(2):1396-1415.
- AMIN, A. and J. SINGLE. 2010. Genetic Variability, heritability and genetic advance studies in Carrot (*Daucus carota* var. *sativa* L.). *Electronic Journal of Plant Breeding*. 1(6): 1504-150.
- CASTILLO, R. 1991. Nuevos Departamentos de Recursos Fitogenéticos en Ecuador. *Diversity*, 7(1-2): p.37-39.
- DAROS, M.; J.R. AMARAL; T.S. PEREIRA; N.R. LEAL; S.P. FREITAS y T. SEDIYAMA. (2002). Caracterização morfológica de acessos de batata-doce. *Horticultura Brasileira*, 20: 43-47.
- ELAMEEN, A.; A. LARSEN; S. KLEMSDAL; S. FJELLHEIM; L. SUNDHEIM; S.; MSOLLA, E. MASUMBA and O.A. ROGNLI. 2011. Phenotypic diversity of plant morphological and root descriptor traits within a sweet potato, *Ipomoea batatas* L. Lam., germplasm collection from Tanzania. *Genet. Resour. Crop Evol.* 58: 397-407.
- FRANCO, L. y R. HIDALGO. 2003. Análisis Estadístico de Datos de Caracterización Morfológica de Recursos Fitogenéticos. Boletín técnico no. 8, Instituto Internacional de Recursos Fitogenéticos (IPGRI), Cali, Colombia. 89 pp.
- FORGY, E. 1965. Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*. 21:768-769.
- FAWCETT, T. 2003. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. Hewlett-Packard Company, 28p.
- GATNAR, E. and WALESIAK, M. (eds.) 2004. Metody statystycznej analizy wielowymiarowej w badaniach marketingowych [Multivariate statistical analysis methods in marketing research]: Wroclaw.
- HARTIGAN, J. and WONG, M. 1979. Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics*, 28: p.100-108.
- KRAGH, H. 1989. Introducción a la Historia de la Ciencia. Crítica, Barcelona, 281p.
- KAUFMAN, L. and P. ROUSSEEUW. 1990. Finding groups in data: an introduction to cluster analysis. Wiley, New York.
- KOHAVI, R. and F. PROVOST. 1998. Glossary of Terms. Editorial for the Special Issue on Applications Machine Learning and the Knowledge Discovery Process, 30(2-3).
- Lebot, V. 2010. Root and Tuber Crops Chapter 3 Sweet Potato. Handbook of Plant Breeding. Ed. J.E. Bradshaw. Springer Science Business Media, 97-125.
- LLOYD, S. 1982. Least squares quantization in PCM. Technical Note, Bell Laboratories. In IEEE Transactions on Information Theory, 28:128-137.
- MILIÁN, M. 2008. Caracterización de la variabilidad de los cultivares de la colección cubana de germoplasma del género *Xanthosoma* (Araceae). Tesis para aspirar al grado de Doctor en Ciencias Biológicas, Ciudad de la Habana, 123p.

- 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. *Revista 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.). *Revista Centro Agrícola*, 40(4):7-10.
- MOLINA, O.; R. GARCÍA; M. MILIÁN; L. GONZÁLEZ y C. PONS. 2015. Métodos de estandarización en colecciones de germoplasma vegetal. *Revista Agricultura Tropical*, 1(2):67-73.
- MOLINA, O.; R. GARCÍA; M. MILIÁN; L. GONZÁLEZ y C. PONS. 2017. Taxobanger v.1.0: Aplicación informática en R para el análisis taxonómico en bancos de germoplasma vegetal. *Revista Cuba Ciencias Informáticas*, 11(3):159-173.
- MOLINA, O.; R. GARCÍA; M. MILIÁN; L. GONZÁLEZ y C. PONS. 2018. Evaluación de técnicas para la clasificación del genofondo de raíces, rizomas y tubérculos tropicales, plátanos y bananos en Cuba. Informe del Proyecto PAH P131LH001027. Instituto de Investigaciones de Viandas Tropicales (Inivit), Minag, Cuba. 146p.
- MILIÁN, M.; O. MOLINA and Y. FIGUEROA. Integrated Characterization of Cuban Germplasm of Cocoyam (*Xanthosoma Sagittifolium* (L.) Schott). *Journal of Plant Genetics and Crop Research*. 2018;1(1):1-18.
- MACQUEEN, J. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, editors. LE Cam LM, Neyman J. Berkeley, CA: University of California Press. 281-297.
- MAECHLER, M.; P. ROUSSEEUW; A. STRUYF; M. HUBERT and K. HORNIK. 2019. Cluster: cluster Analysis Basics and Extensions. R package version 2.1.0. Available from: <https://cran.r-project.org/web/packages/cluster/index.html>.
- MAIMON, O. and ROKACH, L. 2005. *Data Mining and Knowledge Discovery Handbook*. Springer US.
- NÚÑEZ, C.A. Y L.D. ESCOBEDO. 2011. Uso correcto del análisis clúster en la caracterización de germoplasma vegetal. *Agronomía mesoamericana*, 22(2): 415-427.
- ROUSSEEUW, P. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math*, 20:53-65.
- PROVOST, F. and T. FAWCETT. 2001. Robust classification for imprecise environments. *Machine Learning*, 42:203-231.
- SORENSEN, T. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analysis of vegetation on Danish commons. *Biologiske Skrifter*, 5, 1-34.
- SWETS, J.; R. DAWES and J. MONAHAN. 2000a. Better decisions through science. 82-87. Consultado: 24 de octubre de 2020. Disponible en: <http://www.psychologicalscience.org/newsresearch/publications/journals/%siam.pdf>.
- SWETS, J.A.; R. DAWES and J. MONAHAN. 2000b. Psychological science can improve diagnostic decisions. *Psy-Chological Science in the Public Interest*, 1(1):1-26.
- RITSCHER, P.S. y Z. HUAMÁN .2002. Variabilidade morfológica da coleção de germoplasma de batata-doce da Embrapa - Centro Nacional de Pesquisa de Hortaliças. *Pesquisa Agropecuária Brasileira*, 37: 485-492.

- R DEVELOPMENT CORE TEAM. 2019. R: A language and environment for statistical computing (version 3.6.1). R Foundation for Statistical Computing. Vienna, Austria. Available from: <http://www.r-project.org/>.
- TAIRO, F.; E. MNENEY and A KULLAYA. 2008. Morphological and agronomical characterization of sweet potato germplasm from Tanzania. *Afr. J. Plant Sci.*, 2(8): 77-85.
- TENENHAUS, M. and F. YOUNG. 1985. An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and Other Methods for Quantifying Categorical Multivariate Data. *Psychometrika*, 50(91).
- WALESIK, M. and A. DUDEK. 2017. ClusterSim: Searching for optimal clustering procedure for a data set. Consultado: 12 de diciembre de 2017. Disponible en: <http://cran.fhcrc.org/web/packages/clusterSim/clusterSim.pdf>.
- WEISS, G.M. and F. PROVOST. 2003. Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. *JAIR*, 19:315-354.