CONSTRUCCIÓN DE DENDROGRAMAS DE TAXONOMÍA NUMÉRICA MEDIANTE EL COEFICIENTE DE DISTANCIA χ^2: UNA REVISIÓN

C. A. Núñez-Colín¶; J. E. Rodríguez-Pérez; R. Nieto-Ángel; A. F. Barrientos-Priego

Programa de Postgrado en Horticultura, Departamento de Fitotecnia, Universidad Autónoma Chapingo, Chapingo, Estado de México, C. P. 56230. MÉXICO. Correo-e: lit007@hotmail.com (¶Autor responsable).

RESUMEN

En taxonomía numérica, los coeficientes de disimilaridad más utilizados son los que corresponden a las distintas distancias taxonómicas, en la presente investigación se realizó una revisión acerca de la distancia χ^2. Esta tiene ventajas sobre la distancia taxonómica media o la euclidiana, ya que permite realizar una prueba de hipótesis no exacta para determinar la igualdad estadística existente entre pares de unidades taxonómicas operativas obtenidas a partir de la distribución probabilística de χ^2 con un grado de libertad y usa el nivel de significancia (α) como medida de similitud. La altura de corte de un dendrograma obtenido en un análisis de agrupación basado en esta matriz se puede determinar por lo que el investigador, al elegir el nivel de α deseado, para la conformación de grupos, no necesita del cálculo de otras pruebas de partición de dendrogramas como el criterio de agrupación cúbica o la pseudo estadística de t^2 de Hotelling. Para ilustrar las ventajas de la distancia χ^2, en comparación con la distancia euclidiana y Manhatan, se presenta un ejemplo con genotipos de tejocote (*Crataegus* spp.) mediante un análisis de agrupación con la comprobación de un análisis discriminante canónico. Encontrando que la distancia χ^2 fue la mejor opción.

PALABRAS CLAVE ADICIONALES: distancias taxonómicas, análisis de agrupación, caracterización.

NUMERICAL TAXONOMY DENDROGRAM CONSTRUCTION USING THE DISTANCE COEFFICIENT χ^2: A REVIEW

ABSTRACT

The most used dissimilarity coefficients in numerical taxonomy are those corresponding to different taxonomical distances. In the present paper we performed a review about the distance χ^2. This distance has advantages over the Average or Euclidian taxonomical distance, because it allows to perform non-exact hypothesis testing to determine existing statistical equality between pairs of operative taxonomical units obtained from the probabilistic distribution of χ^2 with one degree of freedom, and it uses the significance level (α) as a similitude measure. The cutting height for a dendrogram, obtained in a cluster analysis based on this matrix can be determined; thus, the researcher, when choosing the desired α level for placing groups together, does not need to calculate other dendrogram partitioning tests such as the cubic clustering criterion or Hottelling's pseudo statistic t^2. To illustrate the advantages of the χ^2 distance, compared to the Euclidian and Manhattan distances, we show an example with hawthorn genotypes (*Crataegus* spp.) using cluster analysis validated with canonical discriminant analysis. We found that the χ^2 distance was the best option.

ADDITIONAL KEY WORDS: taxonomical distances, cluster analysis, characterization.

INTRODUCCIÓN

El estudio de los recursos fitogenéticos se ha convertido en una prioridad científica sobre todo aquellos con poco estudio y potencial comercial, lo que hace importante el estudio de esta diversidad mediante métodos cuantitativos que ayuden a agrupar poblaciones de un mismo género o especie, mediante similitudes de caracteres homólogos. Sokal y Sneath (1963) propusieron el tratado de taxonomía numérica, el cual mediante variables binomiales, multinomiales y cuantitativas analizadas con técnicas estadísticas multivariadas, agrupa poblaciones de acuerdo a similitudes de dichos caracteres, con la finalidad de facilitar su interpretación, desde el punto de vista taxonómico o agronómico de acuerdo con el interés del investigador. Para ello se hace referencia al concepto de unidad taxonómica operativa (OTU, "operational taxonomic unit", por sus siglas en Inglés), que es definida como aquellos individuos o poblaciones que son el objeto del

Recibido: 4 de julio, 2003
Aceptado: 18 de diciembre, 2003

estudio y pueden ser representados en un dendrograma (González–Andrés, 2001).

La matriz de datos obtenidos de un estudio de caracterización utiliza generalmente, las medias de la población, pero si en ésta no se presenta una gran variación, es recomendable corroborar la homogeneidad de datos dentro de la población para ver si la media es representativa.

La semejanza o desemejanza entre OTU diferentes se basa en la obtención inicial de distancias o índices multivariados, generados mediante el tratamiento simultáneo de todos los caracteres. Dependiendo del tipo de investigación que se realice pueden ser distancias taxonómicas, para estudios de caracterización morfológica; o distancias genéticas, para el estudio de caracterización mediante marcadores genético moleculares.

Las distancias están basadas en distribuciones multivariadas que a su vez se basan en planos euclidianos *p*–dimensionales (Lindgren, 1968), por lo que es importante conocer qué es un plano euclidiano, concepto básico para entender de forma correcta la medición de las distancias multivariadas. El espacio euclidiano puede ser *p*–dimensional el cual está definido por el álgebra vectorial y la geometría analítica (Haaser *et al.*, 2001)

El cálculo estadístico de la distancia es necesario para realizar un análisis por agrupación, ya que en principio, se debe medir la semejanza o desemejanza entre dos observaciones separadas y, a continuación, la semejanza o desemejanza entre dos grupos de observaciones.

Se ha generado gran número de propuestas de distancias multivariadas para estudios de caracterización mediante taxonomía numérica; sin embargo, la elección de aquella que sea más adecuada, con base en su manejo desde el punto de vista teórico – estadístico para su manipulación, presenta complicaciones; por lo que el investigador requiere de información adecuada acerca de los supuestos de dichas distancias. Así, el objetivo de este ensayo es revisar teórica y empíricamente las ventajas e inconvenientes de las distancias más utilizadas, la euclidiana, la taxonómica media, la de Manhatan y la opción de la distancia χ^2 que es recomendada por el International Plant Genetic Resourses Institute (IPGRI) para datos en escalas similares, aunque es poco utilizado en estudios de taxonomía numérica (Hidalgo, 2003).

ESPACIO EUCLIDIANO

Para definir el espacio euclidiano se utilizará como ejemplo del espacio euclidiano tridimensional, puesto que se visualiza de una manera más sencilla por tener noción del espacio tridimensional x, y, z. El cual, a su vez, puede ser generalizado en espacios euclidianos *p*–dimensionales (Haaser *et al.*, 2001).

El espacio euclidiano tridimensional se denota R^3. Los puntos de R^3 son ternas ordenadas (x, y, z) del espacio vectorial V_3. Los cuales son las coordenadas rectangulares del punto $\mathbf{p} = (x, y, z)$.

Una recta en R^3 está determinada por un punto $\mathbf{p}_0$ y una dirección **a** (**a** es un vector no nulo). Los puntos **p** sobre la recta **L** que pasa por $\mathbf{p}_0$ en la dirección **a**, son de la forma $\mathbf{p} = \mathbf{p}_0 + t\mathbf{a}$, donde *t* es un número real y representa la magnitud de la línea **L**.

Un plano en R^3 está determinado por dos rectas no paralelas $\mathbf{L}_1$ y $\mathbf{L}_2$ de direcciones respectivas **a** y **b** que se cortan en un punto $\mathbf{p}_0$. Los puntos **p** sobre el plano **P** determinado por $\mathbf{L}_1$ y $\mathbf{L}_2$ son los puntos de la forma $\mathbf{p} = \mathbf{p}_0 + u\mathbf{a} + v\mathbf{b}$, donde *u* y *v* son números reales y representan la magnitud de las líneas $\mathbf{L}_1$ y $\mathbf{L}_2$, respectivamente.

La distancia en el espacio euclidiano tridimensional R^3, desde el punto $\mathbf{p}_1$ a un punto $\mathbf{p}_2$, se define como la longitud del vector $\mathbf{p}_2 - \mathbf{p}_1$, que va de $\mathbf{p}_1$ a $\mathbf{p}_2$, es decir, $d(\mathbf{p}_2,\mathbf{p}_1) = |\mathbf{p}_2 - \mathbf{p}_1| = [(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2]^{½}$ (Figura 1).

Pero se generaliza mediante la fórmula:

$$E_{ij} = [\Sigma_k \ (X_{ki} - X_{kj})^2]^{½}$$

Donde:

E_{ij} = la distancia taxonómica media entre la OTU_i y la OTU_j

X_{ki} = el carácter *k* de la OTU_i donde $k = 1, 2, 3, \ldots, n$

X_{kj} = el carácter *k* de la OTU_j donde $k = 1, 2, 3, \ldots, n$

n = número de caracteres evaluados

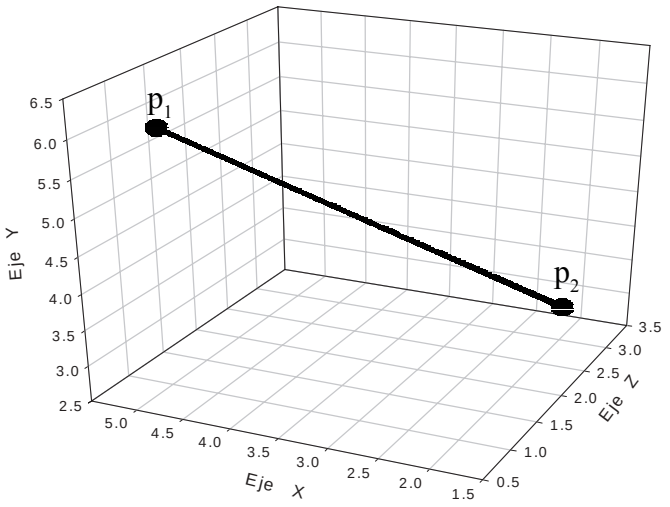

FIGURA 1. Distancia entre dos puntos en un espacio euclidiano tridimensional.

Y la distancia taxonómica media (Sokal y Sneath, 1963) es definida como la media aritmética de las distancias euclidianas entre el número de caracteres evaluados, esto fue hecho para poder comparar valores de distancia sin importar el número de caracteres evaluados; la distancia taxonómica media presenta la siguiente fórmula:

$$E_{ij} = [\Sigma_k 1/n\ (X_{ki} - X_{kj})^2]^{½}$$

Donde:

E_{ij} = la distancia taxonómica media entre la OTU_i y la OTU_j

X_{ki} = el carácter *k* de la OTU_i donde *k* = 1, 2, 3,…, *n*

X_{kj} = el carácter *k* de la OTU_j donde *k* = 1, 2, 3,…, *n*

n = número de caracteres evaluados

Y la distancia City Block o de Manhatan (Sokal y Sneath, 1963) es definida como la media aritmética de la suma del valor absoluto de las diferencias de los caracteres evaluadas; la cual presenta la siguiente fórmula:

$$M_{ij} = 1/n\ [\Sigma_k\ |\ X_{ki} - X_{kj}|]$$

Donde:

M_{ij} = la distancia de Manhatan entre la OTU_i y la OTU_j

X_{ki} = el carácter *k* de la OTU_i donde *k* = 1, 2, 3,…, *n*

X_{kj} = el carácter *k* de la OTU_j donde *k* = 1, 2, 3,…, *n*

n = número de caracteres evaluados

La distancia χ^2 está definida por la fórmula:

$$d_{ii} = \left(\sum_k \frac{1}{X_{k\bullet}} \left(\frac{X_{ki}}{X_{\cdot i}} - \frac{X_{kj}}{X_{\cdot j}} \right)^2 \right)^{½}$$ Donde:

d_{ij} = la distancia entre la OTU_i y la OTU_j

X_{ki} = el carácter *k* de la OTU_i

$X_{k\bullet} = \Sigma_k\ (X_{ki} + X_k)$; donde *k* = 1, 2, 3,…, *n*

$X_{\bullet i} = \Sigma_i\ (X_{ki})$; donde *k* = 1, 2, 3,…, *n*

$X_{\cdot j} = \Sigma_j\ (X_{kj})$; donde *k* = 1, 2, 3,…, *n*

n = número de caracteres evaluados

Mediante esta distancia se comparan a la vez solamente dos OTU por lo tanto, esta distancia se basa en la distribución χ^2 con un grado de libertad. Derivando de ello, se presenta una gráfica de probabilidad de la distribución χ^2 mediante valores de α de tablas de probabilidad (Figura 2)

Cuando en el cálculo de la distancia se obtiene un valor que tiende a 0, las OTU estudiadas presentan un mayor parecido y entre mayor sea el nivel de α ($0 \leq \alpha \leq 1$), el valor en tablas de probabilidad de la distribución χ^2 con un grado de libertad tiende a 0, por lo que si dos OTU son iguales, el valor de la distancia es 0 y el de α=1 (Figura 2), lo que indica que tiene un nivel de similaridad del 100 %. Por el contrario, entre más se aleja de 0 (χ^2 no acepta valores negativos, por ser una distribución cuadrática) mayor es el grado de disimilaridad, por lo que α corresponde a los valores de probabilidad de la distancia χ^2.

De ahí que α, para este caso, no es la probabilidad de cometer error tipo 1, sino que corresponde al de similitud, por lo que se usará ζ (z del alfabeto griego) para evitar confusiones. Los valores de tablas de similitud (Cuadro 1) son obtenidos de la función inversa gamma de MS Excel.

El valor obtenido al utilizar la fórmula de la distancia χ^2, es el valor de χ^2_c de las OTU evaluadas. Se tiene entonces que la hipótesis nula es H_0: $OTU_i = OTU_j$; es decir las dos OTU son iguales a un nivel de similitud fijado por el investigador, que corresponde a la gráfica de probabilidad. Con lo que se tiene la regla de decisión de aceptar H_0 si $\chi^2_c \geq \chi^2_\zeta$, es decir, las dos OTU evaluadas son similares a la ζ fijada. Y se rechaza la hipótesis nula si $\chi^2_c < \chi^2_\zeta$, es decir, las OTU evaluadas no tienen la similitud de la ζ fijada.

COMPARACIÓN DE LA DISTANCIA χ^2 CON LA DISTANCIA EUCLIDIANA, LA DISTANCIA TAXONÓMICA MEDIA Y LA DISTANCIA DE MANHATAN

Para poder ejemplificar cómo obtener los grupos en un análisis de caracterización mediante taxonomía numérica, se utilizará como ejemplo 41 genotipos de tejocote (*Crataegus* spp.) del Banco de Germoplasma de la Universidad Autónoma Chapingo (Cuadro 2), con 27 variables morfológicas (Cuadro 3).

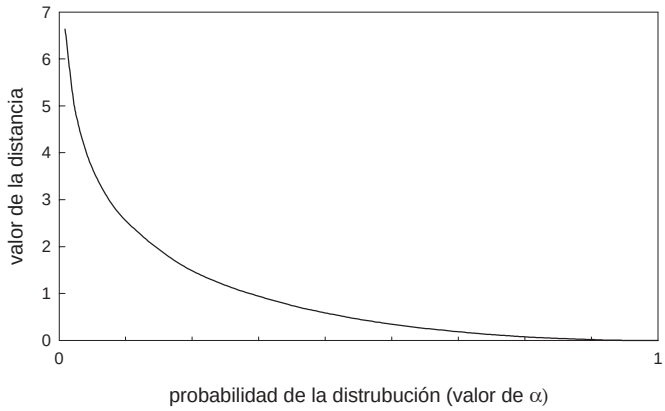

FIGURA 2. Valores de probabilidad de la distribución χ^2 obtenido mediante valores de tablas de probabilidad.

CUADRO 1. Valores de χ^2 con un grado de libertad para distintos valores de α.

Nivel de significancia (ζ)	Valor de χ^2 en tablas para 1 grado de libertad
0.990	0.000200
0.975	0.000982
0.950	0.003930
0.900	0.015800
0.800	0.064160
0.750	0.102000
0.700	0.148450
0.600	0.275000
0.500	0.455000
0.400	0.708290
0.300	1.074130
0.200	1.642500
0.100	2.706000
0.050	3.841500
0.025	5.023900
0.010	6.634900

Este género está constituido por 150 especies, 10 se encuentran en México (Nieto-Ángel y Borys, 1993), y presentan una gran diversidad de tipos, tanto silvestres como cultivados.

Para evitar problemas por utilizar unidades de medida diferentes, los datos de dichas expresiones se transformaron con la función logarítmica con base 10 (Little y Hills, 2001); lo que homogeneiza las varianzas de las variables evaluadas

Mediante las distancias antes mencionadas se obtuvieron dendrogramas con la técnica de agrupación de la media aritmética no ponderada (UPGMA) para poder observar las diferencias existentes entre la distancia χ^2 y las otras distancias. El primer dendrograma que se expone es el de la distancia taxonómica media en la cual se puede observar que tiene una bifurcación inicial en dos grupos (Figura 3); pero se recomienda tener por lo menos tres agrupaciones de acuerdo a las pruebas de partición de grupos como el criterio de agrupación cúbica, y la prueba pseudo estadística t^2 de Hotelling (Johnson, 1998).

Al no contar con esta prueba de partición para obtener una altura de corte más estricta, no se tiene certidumbre de dónde cortar, por lo que el criterio del investigador es el que lo fija arbitrariamente, por lo que para este ejemplo se fijó o en cinco grupos y con un criterio más estricto en ocho grupos. No se tiene la certeza de que estos grupos estén correctamente agrupados, por lo que se procede a elaborar un análisis discrimínate canónico y evaluar las diferencias entre grupos con la distancia de Mahalanobis (Johnson, 1998).

CUADRO 2. Accesiones analizadas de tejocote.

Número de accesión	Clave	Localidad	Origen	Colector
1	RNU02	Rancho Nuevo	Chiapas	Borys, 1982
2	RNU06	Rancho Nuevo	Chiapas	Borys, 1982
3	SCC02	San Cristóbal de las Casas	Chiapas	Borys, 1982
6	RNU01	Rancho Nuevo	Chiapas	Borys, 1982
8	SSC05	San Cristóbal de las Casas	Chiapas	Borys, 1982
9	RNU04	Rancho Nuevo	Chiapas	Borys, 1982
12	RRO01	Rancho Robelo	Chiapas	Borys, 1982
13	RNU03	Rancho Nuevo	Chiapas	Borys, 1982
14	RRO06	Rancho Robelo	Chiapas	Borys, 1982
15	MIT02	Mitzitán	Chiapas	Borys y Nieto, 1985
16	MIT01	Mitzitán	Chiapas	Borys y Nieto, 1985
17	RNU07	Rancho Nuevo	Chiapas	Borys, 1982
18	RRO03	Rancho Robelo	Chiapas	Borys, 1982
19	MIT04	Mitzitán	Chiapas	Borys y Nieto, 1985
20	MIT05	Mitzitán	Chiapas	Borys y Nieto, 1985
21	RRO04	Rancho Robelo	Chiapas	Borys, 1982
22	MIT03	Mitzitán	Chiapas	Borys y Nieto, 1985
23	RNU08	Rancho Nuevo	Chiapas	Borys, 1982
27	RRO05	Rancho Robelo	Chiapas	Borys, 1982
35	MIT07	Mitzitán	Chiapas	Borys y Nieto, 1985
72	SJY02	San José Yashitinín	Chiapas	Nieto y Barrientos, 1989
83	SCC09	San Cristóbal de las Casas	Chiapas	Nieto y Barrientos, 1989
24	CALP5	Calpan	Puebla	Borys y Nieto, 1983
25	CALP6	Calpan	Puebla	Borys y Nieto, 1983
26	CALP3	Calpan	Puebla	Borys y Nieto, 1983
31	CALP1	Calpan	Puebla	Borys y Nieto, 1983
33	HUE01	Huejotzingo	Puebla	Borys y Nieto, 1983
45	HUE02	Huejotzingo	Puebla	Borys y Nieto, 1983
46	CALP2	Calpan	Puebla	Borys y Nieto, 1983
48	HUE04	Huejotzingo	Puebla	Borys y Nieto, 1983
54	HUE05	Huejotzingo	Puebla	Borys y Nieto, 1983
55	HUE06	Huejotzingo	Puebla	Borys y Nieto, 1983
56	HUE07	Huejotzingo	Puebla	Borys y Nieto, 1983
86	HUE10	Huejotzingo	Puebla	Nieto y Barrientos, 1988
93	XAM04	Xamimilulco	Puebla	Nieto y Barrientos, 1988
28	SPI01	San Pablo Ixayoc	México	Borys y Nieto, 1982
32	BAT02	Batán	México	Nieto, 1983
37	SPI04	San Pablo Ixayoc	México	Borys y Nieto, 1982
62	TEQ01	Tequexquinauac	México	Borys y Nieto, 1982
63	TEQ02	Tequexquinauac	México	Borys y Nieto, 1982
65	TEQ05	Tequexquinauac	México	Borys y Nieto, 1982

CUADRO 3. Caracteres medidos de las accesiones de tejocote.

Tipo de variable	Variable	Unidades
Vigor del genotipo	Relación diámetro del injerto/diámetro del portainjerto	
	Diámetro de unión del injerto	cm
Hábitos de	Altura del árbol	m
crecimiento	Área máxima de la copa	m^2
de genotipos	Tipo de Copa	Nivel [z]
	Diámetro máximo de la copa	m
	Diámetro de la sombra zenital proyectada	m
	Densidad de población potencial	Árboles·ha^{-1}
	Número de hojas por brote vegetativo	
	Área foliar por brote vegetativo	cm^2
	Longitud por brote vegetativo	cm
	Número de espinas por brote vegetativo	
	Longitud de espinas por brote vegetativo	cm
	Número de hojas por brote reproductivo	
	Área foliar por brote reproductivo	cm^2
	Longitud por brote reproductivo	cm
	Número de frutos por brote reproductivo	
	Relación diámetro longitud del fruto	
	Peso de fruto	g
	Número de semillas	
	Peso de semillas	g
Características	Acidez titulable (Ácido málico)	%
de fruto	pH	Escala pH
	Sólidos solubles totales	°Bx
	Escala de color de amarillo a azul	a Hunter
	Escala de color de negro a blanco	L Hunter
	Escala de color de rojo a verde	b Hunter

[z]1 = Esferoide, 2 = Semiesferoide, 3 = Semielipsoide, 4 =Elipsoide, 5 = Ovalada, 6 = Cuadrangular, 7 = Cónico invertido

En el caso del ejemplo con cinco grupos, se calculó el nivel de significancia de la distancia de Mahalanobis (Cuadro 4) y se encontró que no existen diferencias significativas entre el grupo 5 con respecto al grupo 4, por lo que probablemente estén mal agrupados, por lo tanto, dicho coeficiente establece la probabilidad de que un grupo y otro similar sean un mismo grupo.

Lo anterior se puede observar de una mejor manera en la gráfica de dispersión entre los componentes canónicos uno y dos (Can 1 y Can 2, respectivamente), obtenidos del análisis discriminante canónico (Figura 4), donde las agrupaciones 4 y 5 se presentan casi como un mismo grupo y no hay separación de grupos. Por lo antes referido, se diría que no hay un agrupamiento adecuado de las OTU del grupo 4 y 5, lo que hace al análisis menos confiable.

Aunque las agrupaciones formadas son bastante aceptables, esa confusión existente entre los grupos 4 y 5, que en la gráfica de dispersión de los componentes canónicos se ve evidente, por lo que el investigador podría tener problemas para justificar la razón de porque separar esos grupos.

CUADRO 4. Nivel de significancia de la distancia de Mahalanobis con 90 % de confiabilidad de las agrupaciones de genotipos de tejocote obtenidas a partir de la distancia taxonómica media.

Grupo	1	2	3	4	5
1	1.0000				
2	0.0009	1.0000			
3	0.0323	0.0218	1.0000		
4	<0.0001	0.0037	0.0018	1.0000	
5	0.0131	0.0526	0.0760	**0.2184**	1.0000

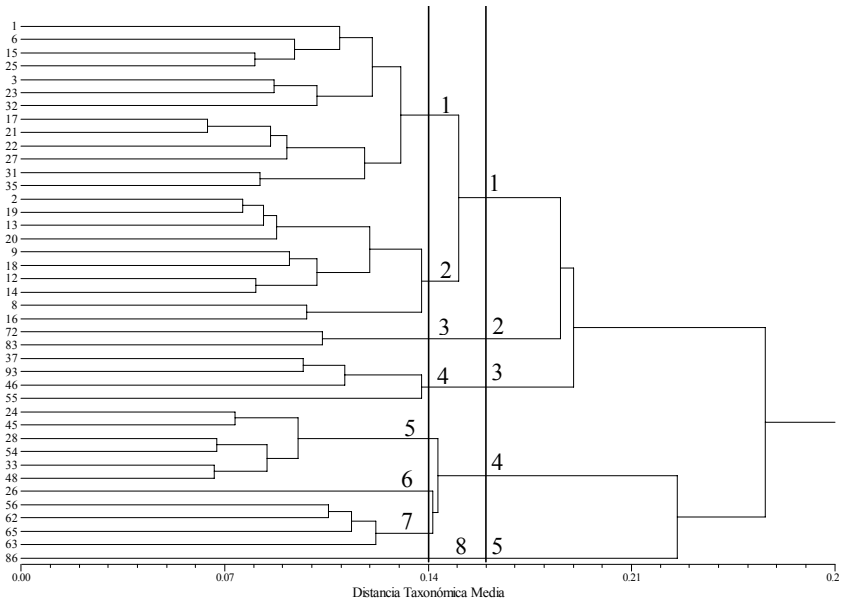

FIGURA 3. Dendrograma de disimilitud de 41 genotipos de tejocote (*Crataegus* spp.) y 27 caracteres morfológicos, mediante la distancia taxonómica media.

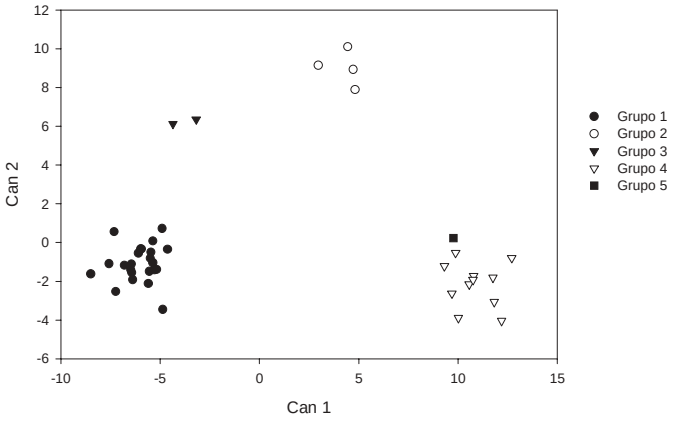

FIGURA 4. Dispersión obtenida de los valores canónicos para cinco grupos obtenidos de la distancia taxonómica media de genotipos de tejocote.

CUADRO 5. Nivel de significancia de la distancia de Mahalanobis con 90 % de confiabilidad de las agrupaciones de genotipos de tejocote obtenidos a partir de la distancia taxonómica media.

Grupo	1	2	3	4	5	6	7	8
1	1.0000							
2	0.0015	1.0000						
3	0.0164	0.0012	1.0000					
4	0.0964	0.0130	0.0513	1.0000				
5	0.0014	0.0004	0.0266	0.0093	1.0000			
6	**0.1695**	0.0178	**0.4140**	**0.2034**	**0.5653**	1.0000		
7	0.0013	0.0003	0.0109	0.0131	0.0944	**0.6769**	1.0000	
8	0.0168	0.0016	0.0485	0.0352	0.0334	**0.3705**	0.0399	1.0000

Pasando al ejemplo cuando tenemos ocho grupos que tiene que ser una prueba más confiable presenta muchos problemas puesto que existe un mayor número de grupos sin diferencias significativas (Cuadro 5).

En la gráfica que se obtuvo del análisis discriminante canónico (Figura 5) los grupos 4 y 1 están sobrepuestos por lo que el agrupamiento no es correcto, lo que hace que pierda confiabilidad.

Para el caso de la distancia euclidiana se tiene un dendrograma (Figura 6) muy parecido al de la distancia taxonómica media pero con una escala diferente.

Las accesiones se comportan y se agrupan de la misma manera que con la distancia taxonómica media, por

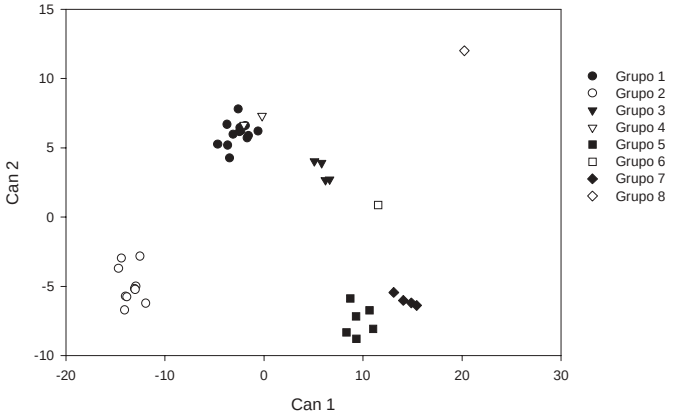

FIGURA 5. Dispersión obtenida de los valores canónicos para ocho grupos obtenidos de la distancia taxonómica media de genotipos de tejocote.

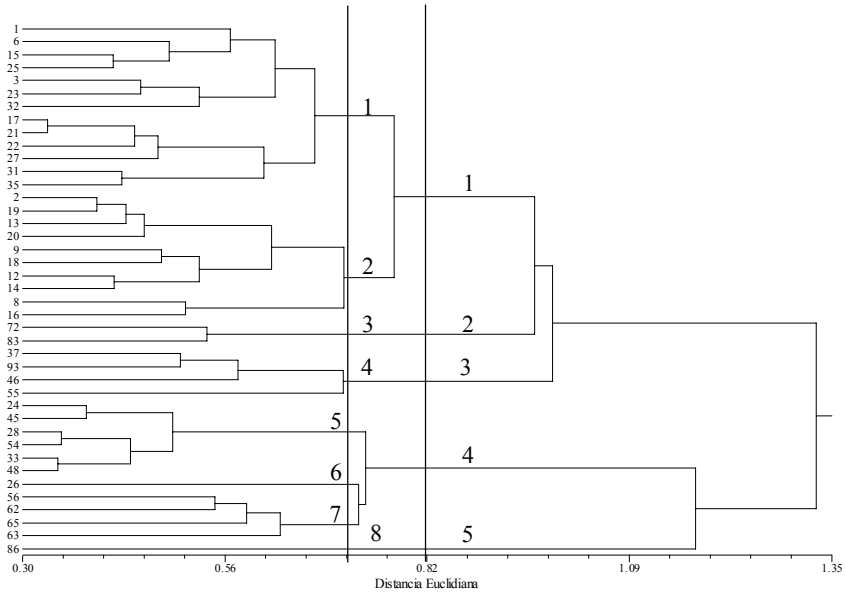

FIGURA 6. Dendrograma de disimilitud de 41 genotipos de tejocote (*Crataegus* spp.) y 27 caracteres morfológicos, mediante la distancia euclidiana.

lo que en el dendrograma se establecerán las mismas líneas de corte; pero las OTU que componen esos grupos son los mismos que dicha distancia, por lo que sería repetitivo presentar los cuadros y figuras del análisis discriminante canónico y las probabilidades de la distancia de Mahalanobis, por lo que se omitieron

En el dendrograma que se obtuvo con la distancia Manhatan se tienen dos alturas de corte, una con cinco y otra con ocho grupos (Figura 7) las cuales se evaluaron mediante un análisis discriminante canónico.

En el caso del ejemplo con cinco grupos, se calculó el nivel de significancia de la distancia de Mahalanobis (Cuadro 6) y se encontró que no existen diferencias significativas del grupo 5 con respecto al grupo 4, por lo que probablemente esté mal agrupado, esto se puede distinguir de una mejor manera en la gráfica de dispersión canónica (Figura 8).

CUADRO 6. Nivel de significancia de la distancia de Mahalanobis con 90 % de confiabilidad de las agrupaciones obtenidas a partir de la distancia de Manhatan de genotipos de tejocote.

Grupo	1	2	3	4	5
1	1.0000				
2	0.0323	1.0000			
3	<0.0001	0.0018	1.0000		
4	0.0009	0.0218	0.0037	1.0000	
5	0.0131	0.0760	**0.2184**	0.0526	1.0000

Aunque son aceptables las agrupaciones formadas por este análisis, esa confusión, existente entre los grupos

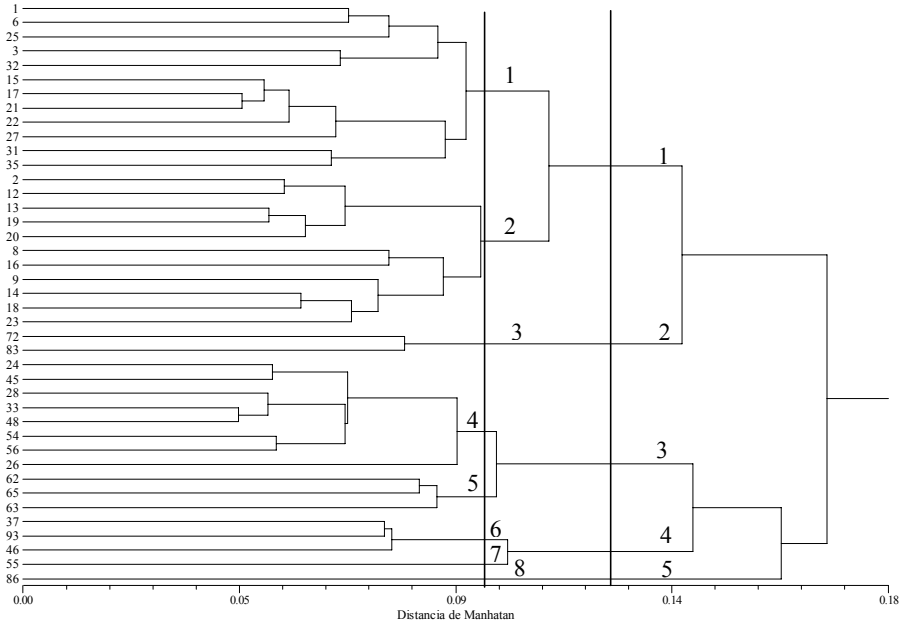

FIGURA 7. Dendrograma de disimilitud de 41 genotipos de tejocote (*Crataegus* spp.) y 27 caracteres morfológicos, mediante la distancia de Manhatan.

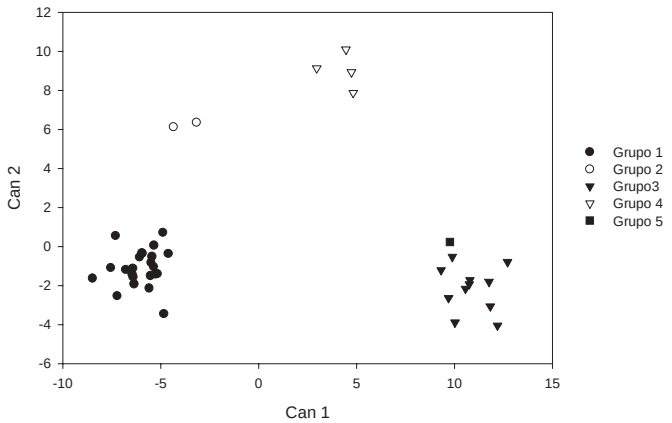

FIGURA 8. Dispersión obtenida de los valores canónicos para 5 grupos obtenidos de la distancia de Manhatan de genotipos de tejocote.

3 y 5 que en la gráfica de dispersión de los componentes canónicos es evidente (Figura 8), lo cual puede ocasionar problemas al investigador para justificar el por qué separar esos grupos.

Recurriendo al ejemplo, cuando se tienen ocho grupos se requiere de una prueba más confiable, en ésta se presentan muchos problemas, puesto que existe un mayor número de grupos sin diferencias significativas (Cuadro 7), al igual que en el caso de la distancia taxonómica media.

En la gráfica que se obtuvo en el análisis discriminante canónico (Figura 9) se observó que los grupos 4 y 8 están prácticamente aglutinados en cierta área, además de que los grupos 2 y 3 están muy cercanos, al igual que los grupos 6 y 7, lo que hace que se pierda confiabilidad.

CUADRO 7. Nivel de significancia de la distancia de Mahalanobis con 90 % de confiabilidad de las agrupaciones obtenidas a partir de la distancia de Manhatan de genotipos de tejocote.

Grupo	1	2	3	4	5	6	7	8
1	1.0000							
2	0.0007	1.0000						
3	0.0151	**0.1177**	1.0000					
4	0.0013	0.0005	0.0082	1.0000				
5	0.0080	0.0006	0.0052	0.0502	1.0000			
6	0.0241	0.0008	0.0093	0.0119	**0.1315**	1.0000		
7	0.0722	0.0086	0.0188	0.0751	**0.2865**	**0.5967**	1.0000	
8	0.0479	0.0216	0.0770	**0.3289**	0.0613	0.0561	0.0718	1.0000

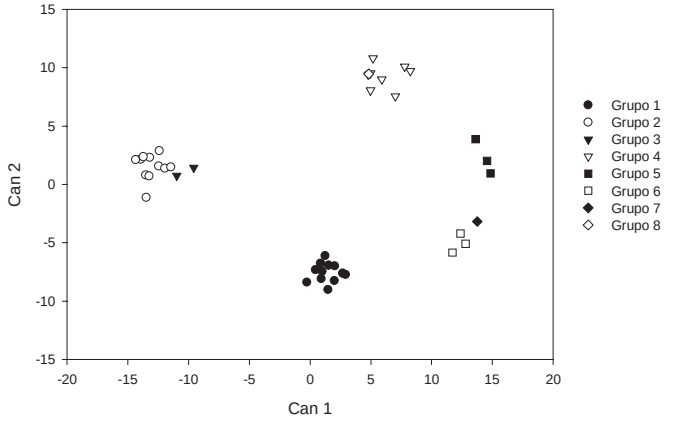

FIGURA 9. Dispersión obtenida de los valores canónicos para 8 grupos obtenidos de la distancia de Manhatan de genotipos de tejocote.

En el dendrograma obtenido mediante la distancia χ^2 (Figura 10) se aprecia diferencia en los agrupamientos obtenidos con respecto a las dos distancias anteriores; pero para decidir la altura de corte se realizó la prueba de hipótesis planteada anteriormente; para lo cual, se busca la distancia de la primera rama de bifurcación. A partir de ese valor, el investigador fija su nivel de similitud (ζ), en nuestro caso la distancia hasta la primera rama es 0.29, por lo que no se puede tener una altura de corte menor a ζ=0.60, que tiene un valor de 0.275 (Cuadro 1) como la primera opción de corte, pero a esta altura se obtienen dos grupos, y generalmente son deseables más de tres agrupaciones; por lo que se fija a un mayor nivel de similitud, fijando ζ=0.70 que presenta un valor de 0.14845≈0.15, con lo que se obtuvieron seis grupos. Al intentar darle un mayor valor a ζ se obtendrían demasiados grupos, por lo que se decidió que tal altura de corte era la correcta y se procedió a la comprobación de las agrupaciones mediante el análisis discriminante canónico y la distancia de Mahalanobis.

Según las probabilidades que se obtuvieron a partir de la distancia de Mahalanobis (Cuadro 8), existen problemas con tres agrupaciones.

CUADRO 8. Nivel de significancia de la distancia de Mahalanobis con 90 % de confiabilidad de las agrupaciones obtenidas a partir de la distancia χ^2 de genotipos de tejocote.

Grupo	1	2	3	4	5	6
1	1.0000					
2	0.0075	1.0000				
3	0.0055	0.0475	1.0000			
4	<0.0001	0.0001	0.0001	1.0000		
5	0.0743	**0.1909**	0.0594	0.0821	1.0000	
6	0.0023	0.0141	0.0028	**0.2042**	**0.1304**	1.0000

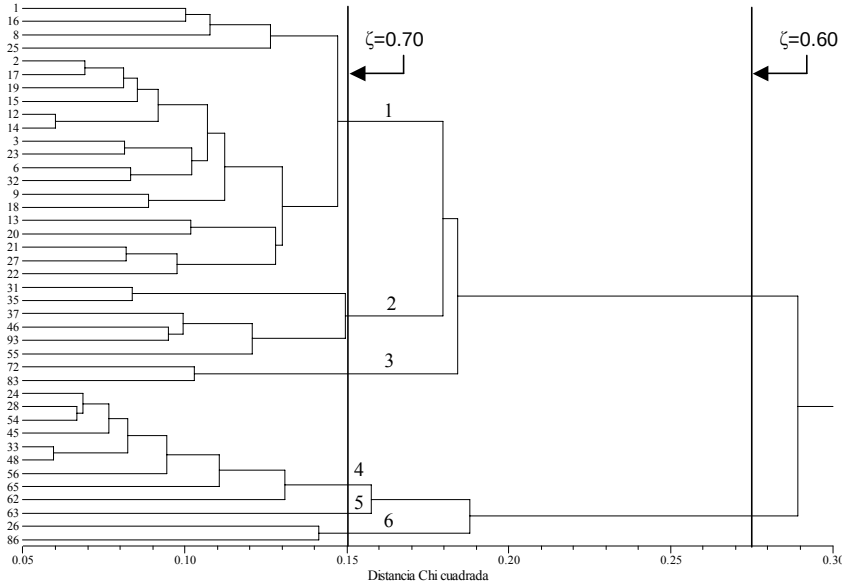

FIGURA 10. Dendrograma de disimilitud de 41 genotipos de tejocote (*Crataegus* spp.) y 27 caracteres morfológicos, mediante la distancia χ^2.

Entre el grupo 2 y el grupo 5 y entre el grupo 6 con los grupos 4 y 5, la distancia de Mahalanobis está basada en la distancia euclidiana; por lo que al obtener las distancias euclidianas en los grupos se presentan similitudes por lo no es un parámetro muy confiable para la distancia χ^2, razón por la cual en la gráfica de dispersión de los valores canónicos (Figura 11) en la cual los grupos presentan similitudes, pero se encuentran separados entre si, por lo que no existe confusión que son diferentes agrupaciones.

Los grupos en la gráfica de dispersión de los valores canónicos (Figura 11) están perfectamente separados, y aunque la probabilidad de la distancia de Mahalanobis indica un parecido estadístico entre estas agrupaciones, en la dispersión canónica se observó esta separación de los grupos; la cual no existe en las gráficas de dispersión canónica de la distancia taxonómica media ni de la distancia de Manhatan (Figura 4 y 8), por lo que los grupos de esta distancia están más congruentes con el análisis discriminante canónico.

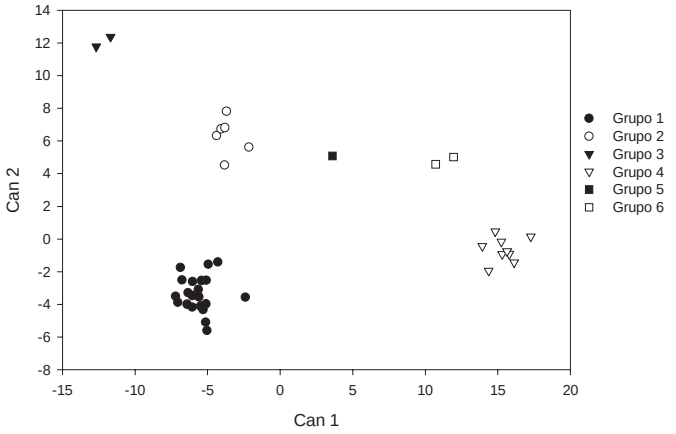

FIGURA 11. Dispersión obtenida de los valores canónicos para seis grupos obtenidos de la distancia χ^2.

Además, para asegurar que no existen errores, se analizan los porcentajes del análisis *a posteriori* (Cuadro 9), lo cual sirve para comprobar que no existen accesiones de un grupo que pertenezca a otro; en este ejemplo se concluye que no existe ninguna accesión que pertenezca a otro grupo, asegurando así que la clasificación sea la correcta.

Aunque la elección de la altura de corte es también arbitraria en esta distancia, se tienen puntos de referencia que facilitan la elección de dónde cortar y cuántas agrupaciones formar, lo que se presenta como la principal ventaja de esta distancia.

CONSIDERACIONES FINALES Y CONCLUSIONES

La distancia χ^2 es recomendable por proporcionar a través de un valor predeterminado la similitud entre grupos de OTU, y poder elegir de una manera sencilla la altura de corte y el número de grupos a formar, lo que representaría evitar el cálculo de prueba de partición de dendrogramas.

En la fórmula de esta distancia se toma en cuenta la variabilidad de los caracteres, lo que da una mayor confiabilidad en el agrupamiento, debido a que esta distancia resalta aquellas variables que tengan una mayor variabilidad para la formación de grupos, por lo que para saber cuales son más importantes para el estudio de caracterización, solamente se obtiene el coeficiente de variación de cada variable de los promedios de las agrupaciones y los que tengan un mayor coeficiente de variación serán los más discriminantes en dicho estudio.

La posible desventaja de esta distancia, puede ser que al agrupar por medio de la variabilidad de los caracteres evaluados, no dé a estos caracteres un peso ecuánime, esto puede dar como resultado problemas por variables con diferencias grandes en la escala en que se midieron lo que podría sesgar la prueba, por esta razón se recomienda tener cuidado con las variables a utilizar en este tipo de coeficientes.

CUADRO 9. Número de observaciones y porcentaje de ellas clasificadas en cada grupo

Grupo	1	2	3	4	5	6	Total
1	21	0	0	0	0	0	21
	100.00	0.00	0.00	0.00	0.00	0.00	100.00
2	0	6	0	0	0	0	6
	0.00	100.00	0.00	0.00	0.00	0.00	100.00
3	0	0	2	0	0	0	2
	0.00	0.00	100.00	0.00	0.00	0.00	100.00
4	0	0	0	9	0	0	9
	0.00	0.00	0.00	100.00	0.00	0.00	100.00
5	0	0	0	0	1	0	1
	0.00	0.00	0.00	0.00	100.00	0.00	100.00
6	0	0	0	0	0	2	2
	0.00	0.00	0.00	0.00	0.00	100.00	100.00
Total	21	6	2	9	1	2	41
	51.22	14.63	4.88	21.95	2.44	4.88	100.00

LITERATURA CITADA

GONZÁLEZ-ANDRÉS, F. 2001. Caracterización morfológica, pp. 199-217. *In*: Conservación y caracterización de recursos filogenéticos. GONZÁLEZ-ANDRÉS, F.; PITA VILLAMIL, J. M. (eds.) Publicaciones Instituto Nacional de Educación Agrícola. Valladolid, España.

HAASER, N. B.; LA SALLE, J. P.; SULLIVAN, J. A. 2001. Análisis Matemático, Curso Intermedio Vol. 2. Traducción F. Velasco Coba. Editorial Trillas, D. F., México. pp. 42-45.

HAIR, J. F. Jr.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. 2001. Análisis Multivariante. 5a edición. Traducido por E. Prense y D. Cano. Prentice Hall Iberia, Madrid, España. pp. 491-545.

HIDALGO, R. 2003. Variabilidad genética y caracterización de especies vegetales. pp. 2-26. *In*: Análisis Estadístico de Datos de Caracterización Morfológica de Recursos Filogenéticos. FRANCO, T. L.; HIDALGO, R. (eds.) Boletín técnico núm. 8 IPGRI, Cali, Colombia.

JOHNSON, D. E. 1998. Métodos Multivariados Aplicados al Análisis de Datos. Traducido por H. Pérez Castellanos. International Thomson Editores, D. F., México.

LINDGREN, B. W. 1968. Statistical Theory. Segunda edición McMillan Company. New York, USA.

LITTLE, T. M.; HILLS, F. J. 2001. Métodos Estadísticos para la Investigación en la Agricultura. Traducido por A. de Paula Crespo. Editorial Trillas, D. F., México.

NIETO-ÁNGEL, R.; BORYS, M. W. 1993. El tejocote (*Crataegus* spp.), un potencial frutícola de zonas templadas. Revista Fruticultura Profesional 54: 64-71

SOKAL, R. R.; SNEATH, P. H. A. 1963. Principles of Numerical Taxonomy. H. Freeman & Company, San Francisco, Cal., USA.