Aplicación de técnicas estadísticas multivariadas para el agrupamiento de materiales genéticos de cacao (*Theobroma cacao* L.)

Autores:
Ana Julia Righetto[1]
Luiz Ricardo Nakamura[2]
Ezequiel López Bautista[3]
Carlos Tadeu dos Santos Dias[3]

Recibido en el CEA el 14 de noviembre de 2013. Aprobado el 22 de abril de 2014.

[1/]Doctorando, Programa de Estadística y Experimentación Agronómica, ESALQ/USP, Piracicaba SP.
[2/]Docente Investigador, Facultad de Agronomía, Universidad de San Carlos de Guatemala.
[3/]Profesor Titular, Departamento de Ciencias Exactas, PPGEEA, ESALQ/USP, Piracicaba SP.

Resumen

El objetivo de este artículo es presentar una aplicación de técnicas estadísticas multivariadas (análisis factorial y de agrupamiento) para clasificar 20 materiales genéticos de cacao (*Theobroma cacao* L.) provenientes de diversos países, localizados en el Centro de Agricultura Tropical "Bulbuxyá" de la Facultad de Agronomía de la Universidad de San Carlos de Guatemala, en el municipio de San Miguel Panán, departamento de Suchitepéquez, y estudiadas 12 características cuantitativas. Debido al número de características estudiadas, se optó por la utilización, inicialmente, del análisis factorial (AF), que busca explicar un determinado conjunto de datos *p*-dimensional en un número menor de dimensiones (*m*-dimensional). El AF realiza la rotación de los factores para que estos sean interpretables desde el punto de vista del problema en cuestión. Para esta rotación, se utilizó el criterio Varimax. Una vez obtenidos las puntuaciones (escores) de cada una de las observaciones (árboles), se utilizó el análisis de agrupamiento, con la finalidad de formar grupos diferentes, en cada uno de ellos contenía árboles con características similares. El método utilizado para el agrupamiento fue el jerárquico de Ward. Con el auxilio de estas técnicas estadísticas multivariadas, los materiales genéticos de cacao fueron divididos en cinco grupos con características distintas.

Palabras clave: Análisis factorial, análisis de agrupamiento, cacao.

Abstract

The aim of this paper is to present an application of multivariate statistical techniques (factor analysis and clustering) to classify 20 genetic materials of cocoa (*Theobroma cacao* l.) from different countries, located in the Centro de Agricultura Tropical "Bulbuxyá" of the Facultad de Agronomía of the Universidad de San Carlos de Guatemala, in San Miguel Panán, Suchitepéquez, and studied 12 quantitative traits. Due to the number of traits studied, it is opted for use, initially, factor analysis (AF), which seeks to explain a given set of *p*-dimensional data in one smaller number of dimensions (*m*-dimensional). AF performs the rotation of factors so that they may be interpreted from the point of view of the problem in question. In this case, the approach of Varimax rotation was used. Once obtained the scores of each observations (trees), we used the cluster analysis and the Ward's hierarchical clustering method was applied, to form different groups, each of them containing trees with similar characteristics. With the support of these multivariate statistical techniques, genetic materials of cocoa were divided into five groups with different characteristics.

Keywords: Factor analysis, cluster analysis, cocoa.

Introducción

El cacao (*Theobroma cacao* L.) es una planta originaria de la floresta amazónica y de la región mesoamericana, crece en el trópico entre las latitudes 20º norte y 20º sur del Ecuador, con clima cálido (temperatura media de 25 a 28ºC), con precipitación variando entre 1200 a 2500 mm y altitud de 10 hasta 1000 metros sobre el nivel del mar (Avendaño et al., 2011). El cacao es un cultivo de gran importancia económica, principalmente para la utilización de las semillas de sus frutos, que son la materia prima de la industria del chocolate (ICCO, 2012).

En Guatemala ocupa 3,990 ha y la producción anual es de 10,927 tm. El volumen de producción se concentra en tres departamentos: Alta Verapaz (31%), Suchitepéquez (31%) y San Marcos (25%). De acuerdo con Ávalos et al. (2012) el valor de este cultivo para Guatemala está en la producción de cacao de alta calidad (tipo gourmet), más que en el aumento en el volumen de producción. Siendo por eso necesario realizar una caracterización agronómica de los materiales genéticos presentes en colecciones de árboles de cacao para poder seleccionar a los mejores individuos y tratar de obtener una población élite.

En este artículo se presenta una aplicación de técnicas estadísticas multivariadas para clasificar 20 materiales genéticos de cacao de la colección localizada en el Centro de Agricultura Tropical “Bulbuxyá” de la Facultad de Agronomía de la Universidad de San Carlos de Guatemala.

Materiales y métodos

Área de estudio: Los datos utilizados en este estudio fueron tomados de la tesis de Ing. Agr. de Gatica (1994), realizada en el Centro de Agricultura Tropical "Bulbuxyá" (CATBUL) de la Facultad de Agronomía de la Universidad de San Carlos de Guatemala, localizado en el municipio de San Miguel Panán, Suchitepéquez, en las coordenadas geográficas 14º39' latitud norte y 91º 22' de longitud oeste y altitud de 325 metros.

Material experimental: Los materiales genéticos incluidos en el estudio fueron: **1.** Pound-12 × Catongo, **2**. EET-400 × SCA-12, **3.** UF-613 × Pound-7, **4.** IMC-67 × SCA-12, **5.** EET-62 × SCA-12, **6.** UF-667 × SCA-12, **7.** UF-613 × Pound-12, **8.** IMC 67 × UF 613, **9.**EET 162 × SCA 12, **10.** UF 668 ×Pound 12, **11. SCA** 6 × EET 95, **12.** SCA 6 × EET 62, **13.** EET 95 × SCA 12, **14.** 75-R, **15**. SGU-50, **16**.SGU-71, **17.** SGU-69, **18**. SGU-72, **19.** SGU-88 y **20**. SGU-54, conforme descrito por Gatica (1994).

Variables medidas: Fueron considerados las siguientes variables, relacionadas con:

a) Semilla: fueron seleccionadas aleatoriamente 15 semillas por material genético y medido: X_1: peso húmedo da semilla (gr), X_2: peso seco de la semilla (gr), X_3: número de semillas por fruto, X_4: largo de la semilla (mm), X_5: ancho de la semilla (mm), X_6: grosor de la semilla (mm).

b) Fruto: fue tomada una muestra de 30 frutos maduros por material genético y medido: X7: largo del fruto (mm), X_8: ancho del fruto (mm), X_9: peso del fruto (gr) e X_{10}: grosor de las paredes del fruto en el surco secundario (mm).

c) X_{11}: Índice de fruto (IF): es definido como la cantidad de frutos necesarios para producir un kg de semilla seca y fermentada de cacao; calculado de acuerdo con la ecuación: IF = [1000/ Peso semilla (g)] × 0,4.

d) X_{12}: Índice de semilla: es el peso promedio de semilla seca expresado en gramos, medido de 15 semillas de cada una de las 10 frutas maduras seleccionadas por material genético.

Análisis estadístico de los datos: Inicialmente fue construida la matriz de correlación entre las variables, calculando los coeficientes de correlación de Pearson y verificando su significancia estadística. Dada la cantidad de variables consideradas en el estudio, se optó por usar el Análisis Factorial (AF), que busca explicar un determinado conjunto de datos *p*-dimensional en un número menor de dimensiones (*m*-dimensional) o factores (Hairet al., 2005). El AF realiza la rotación de los factores para que estos sean interpretables desde el punto de vista del problema. Para ello se utilizó el criterio Varimax. Matemáticamente, el modelo factorial ortogonal es dado por (Mardia, Kent &Bibby, 1992):

$$\mathbf{X} - \boldsymbol{\mu} = \boldsymbol{\Lambda} \boldsymbol{F} + \boldsymbol{\varepsilon}, \qquad (1)$$

en que $\mathbf{X}$ es un vector p-dimensional de variables aleatorias, con media $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}_{p \times p}$; es la matriz de cargas factoriales de dimensión $p \times m$; $\boldsymbol{F}$ es el vector m-dimensional de variables latentes o factores; y $\mathbf{e}$ es el vector p-dimensional de errores aleatorios, con media 0 e varianza diagonal $(\boldsymbol{\Psi})$.

Para la estimación de las matrices $\boldsymbol{\Lambda}$ y $\boldsymbol{\Psi}$ fue utilizado el método de los componentes principales por medio de la matriz de correlación y para la selección del número m de factores fueron utilizados los siguientes criterios (Johnson &Wichern, 2007): **i)** raíz latente: m corresponde al número de autovalores extraídos de la matriz de correlación, mayores que uno; y **ii)** análisis de representatividad con relación a la varianza original, los m factores deben representar, en conjunto, un porcentaje $\gamma \times 100\%$ de la varianza original de los datos. Una vez obtenidos las puntuaciones de cada una de las observaciones, se utilizó el análisis de agrupamiento (AA), con el objetivo de formar grupos diferentes, cada uno de ellos con materiales genéticos con características similares. El método utilizado fue el jerárquico de Ward.

Resultados y discusión

Inicialmente fue calculada la matriz de correlación entre los datos, presentando un número substancial de correlaciones superiores a 0.30 (en valor absoluto), además de presentar un valor de medida de la adecuación muestral (Coeficiente KMO) de 0.62, lo que según Hair et al. (2005) valida la aplicación de este método multivariado. En el Cuadro 1 son presentados los autovalores y varianza explicada (%) referente a cada una de las variables estudiadas. Tal como fue expuesto en la sección de materiales y métodos, esa información será utilizada para la selección del número de *m* factores.

Cuadro 1 – Autovalores, varianza explicada y varianza explicada acumulada de cada uno de los factores.

Table 1– Eigenvalues, explained variance and accumulated explained variance for each factors.

Factor	Autovalor (λ)	Varianza explicada (%)	Varianza acumulada (%)
1	6.10	50.81	50.81
2	2.57	21.40	72.21
3	1.43	11.93	84.14
4	0.60	5.02	89.16
5	0.33	2.78	91.94
6	0.32	2.68	94.62
7	0.25	2.05	96.67
8	0.19	1.57	98.24
9	0.13	1.05	99.29
10	0.06	0.46	99.75
11	0.02	0.16	99.91
12	0.01	0.09	100.00

Por el criterio de la raíz latente, tres factores deben ser retenidos en el sistema de autovalores, ll $= 6.10$, $\lambda_1 = 2.57$ e $\lambda_1 = 1.43$, totalizando 84.14% de la varianza acumulada explicada de las variables originales, que según el criterio preestablecido por Mardia, Kent y Bibby (1992) es un porcentaje aceptable. De esta manera, se continuó el análisis utilizán-

dose los tres factores, con el cálculo de la matriz de cargas factoriales por medio del método de los componentes principales y la rotación usando el criterio Varimax (Cuadro 2).

En el Cuadro 2 se muestran las cargas factoriales que más influyeron en cada uno de los factores, destacadas en negrito, y que auxiliaron para etiquetarlos, de la siguiente manera:

a) Factor 1: las variables con cargas factoriales más elevadas son: largo del fruto (mm), ancho del fruto (mm), peso del fruto (g) y grosor de las paredes del fruto en un surco secundario (mm). Este factor fue etiquetado como "características del fruto".

b) Factor 2: las variables con cargas factoriales más elevadas son: peso húmedo de semilla (g), peso seco de semilla (g), índice de fruto e índice de semilla. Este factor fue etiquetado como "características de la semilla".

c) Factor 3: las variables con cargas factoriales más elevadas son: número de semillas por fruto, largo de la semilla (mm), ancho de la semilla (mm) y grosor de la semilla. Este factor fue etiquetado como: "componente del rendimiento".

Cuadro 2 – Matriz factorial rotada por el criterio Varimax.

Table 2 – Rotated factor matrix using Varimax criterion.

Variables	Factor		
	1	2	3
X_1	0.402	**0.838**	-0.037
X_2	0.027	**0.931**	0.237
X_3	0.217	-0.179	**-0.874**
X_4	0.555	0.442	**0.567**
X_5	0.639	0.227	**0.655**
X_6	0.531	0.272	**0.637**
X_7	**0.790**	0.255	-0.034
X_8	**0.873**	-0.007	0.003
X_9	**0.956**	0.174	0.039
X_{10}	**0.723**	-0.068	0.456
X_{11}	-0.019	**-0.954**	-0.112
X_{12}	0129	**0.852**	0.440

El etiquetado de los factores fue realizado de forma simple y objetiva, ya que la aplicación del AF proporcionó la separación de las variables en estudio de forma coherente. Con los valores de las cargas factoriales, así como del significado de cada uno de los factores, fue realizado el agrupamiento por el método jerárquico de Ward de los árboles de cacao en estudio utilizando las puntuaciones (scores) obtenidos con el AF. En la Figura 1 se presenta el dendrograma resultante.

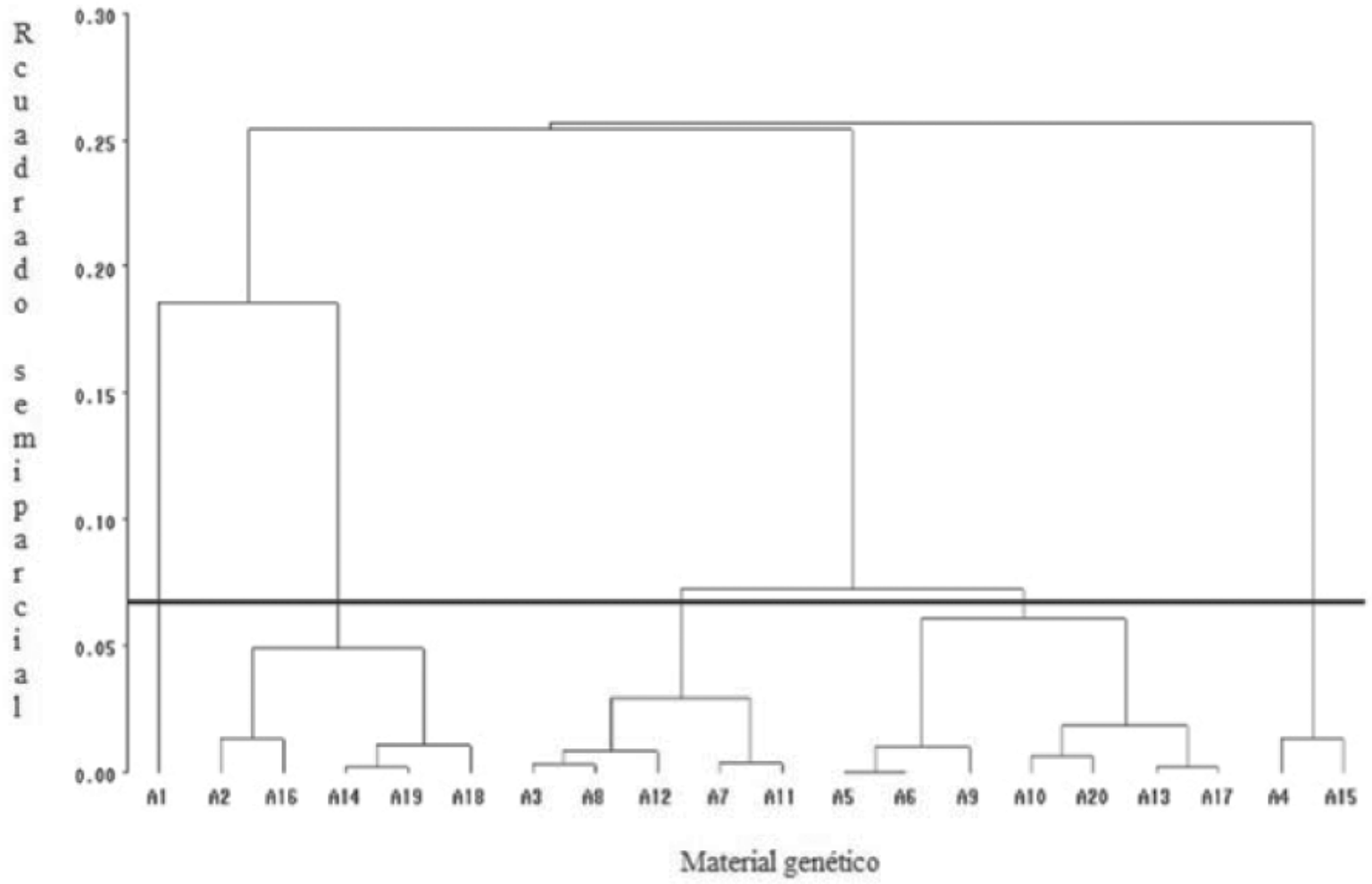

Figura 1.–Dendrograma resultante del agrupamiento jerárquico por el método de Ward.

Figure 1.–Dendrogram computed from the Ward's hierarchical clustering method.

De acuerdo con el dendrograma (Figura 1), los 20 materiales genéticos fueron clasificados en cinco grupos, los cuales son descritos a continuación.

Grupo 1: formado por el material genético Pound 12 x Catongo, en las condiciones del lugar donde fue realizado el estudio, presentó los rendimientos más bajos, entre los materiales genéticos evaluados, siendo de 311.58 kg de semilla seca de cacao por hectárea (rendimiento promedio durante los 5 períodos evaluados); bajo peso (húmedo y seco) de semillas, y valor alto de índice de fruta (30 frutas para producir un kg de cacao seco) y menor índice de semilla (13.2).

Grupo 2: formado por los materiales genéticos: 2, 14, 16, 18 y 19. Este grupo posee como característica principal, semillas grandes (mayor longitud, ancho y grosor) y número bajo de semillas por fruto (33) y rendimiento de 334.5 kg de semilla seca de cacao por hectárea (media del rendimiento durante los 5 períodos evaluados).

Grupo 3: formado por los materiales genéticos: 3,7,8,11 y 12. Los individuos de este grupo produjeron los mayores rendimientos, 421.81 kg de semilla seca de cacao por hectárea, e valores medios de las características medidas.

Grupo 4: en este grupo están incluidos los materiales genéticos: 5,6,9,10,13,17 y 20, siendo el más numeroso. Como característica principal, los individuos de este grupo poseen semillas pequeñas, como consecuencia número alto de semillas por fruto (38.7) y rendimiento de 378.7 kg de semilla seca de cacao por hectárea, el segundo mayor dentro de los observados.

Grupo 5: formado por los materiales genéticos: 4 y 15, los individuos de este grupo se caracterizan por poseerlos frutos con menores dimensiones (longitud, ancho y grosor de las paredes del fruto) y como consecuencia pesos bajos de frutos. Rendimiento de 369.27 kg de semilla seca de cacao por hectárea (media del rendimiento durante los 5 períodos evaluados).

Conclusiones

Con el auxilio del análisis factorial, por el criterio de la raíz latente, fueron retenidos tres factores, que explican 84.14% de la varianza acumulada de las variables originales. Posterior al análisis factorial, se aplicó el análisis de agrupamiento, cuyo resultado fue agrupar, los 20 materiales genéticos, en 5 grupos distintos.

Literatura citada

[1] AVALOS, A.; PORRES, M.A.; PÖLL, E.; DARDÓN, E.; ARÉVALO, L.A.; ROSALES, J.A. Caracterización agronómica, botánica y molecular de clones de cacao tipo criollo y mejorado de la zona sur de Guatemala. Revista de la Universidad del Valle de Guatemala, 24:99-104, 2012.

[2] AVENDAÑO, C.H.; VILLAREAL, J.M.; CAMPOS, E.; GALLARDO, R.A. 2011. Diagnóstico del cacao en México. Universidad Autónoma Chapingo, Secretaría de Agricultura, Ganadería, Desarrollo Rural, Pesca y Alimentación, Servicio Nacional de Inspección y Certificación de Semillas, Red de Cacao. 79 p.

[3] GATICA, M.B. 1994. Caracterización agromorfológica de 13 híbridos y 7 clones de cacao (*Theobroma cacao* L.) en el Centro de Agricultura Tropical "Bulbuxyá" de la Facultad de Agronomía de la Universidad de San Carlos de Guatemala. Tesis Ing. Agr. Universidad de San Carlos de Guatemala, Guatemala, Facultad de Agronomía. 100 p.

[4] HAIR, J.F.; TATHAM, R.L.; ANDERSON, R.E.; BLACK, W.C. Análise multivariada de dados. 5 ed. Porto Alegre, Bookman, 2005. 593 p.

[5] INTERNATIONAL COCOA ORGANIZATION (ICCO). 2012. Quarterly Bulletin of Cocoa Statistics, v. 38, n. 4, Cocoa year 2011/12.

[6] JOHNSON, R.A.; WICHERN, D.W. Applied multivariate statistical analysis. 6 ed. New Jersey, Prentice Hall, 2007. 773p.

[7] MARDIA, K.V.; KENT, J.T.; BIBBY, J.M. Multivariate analysis. London, AcademicPress, 1992. 518p.