Rev. Fitotec. Mex. 18: 188-203. 1995

EL ANALISIS BIPLOT EN CLASIFICACION

BIPLOT ANALYSIS IN CLASSIFICATION

José de Jesús Sánchez González[1]

RESUMEN

El método de análisis Biplot se presenta como una alternativa gráfica en clasificación. Se enfatiza el uso del método para la inspección de matrices de datos, sobre todo cuando se estudian grandes números de variables y la inspección visual es poco práctica. A través de este análisis, es posible examinar en forma simultánea la estructura de los datos; es decir, varianzas y correlaciones aproximadas, así como las interrelaciones, patrones o grupos entre las unidades taxonómicas. Se presentan los aspectos teóricos relevantes en los análisis de componentes principales y de la estructura de una matriz de datos para finalmente proponer un programa escrito en lenguaje matricial del Sistema de Análisis Estadístico (IML/SAS) ilustrando el método Biplot con un ejemplo.

PALABRAS CLAVE ADICIONALES

Análisis multivariado, clasificación, taxonomía numérica, biplot, componentes principales.

SUMMARY

The Biplot technique is presented as a graphical multivariate tool useful in classification of taxonomic units. The method is particularly useful in the display and analysis of large data matrices where visual examination is impractical. The Biplot obtains a two-dimensional approximation to a matrix and allows inspection of variances, approximate correlations, relationships, patterns and clusters existing in the data. The basic theory related to Principal Component Analysis and the stucture of a data matrix is presented. A program was written in the matrix languaje of the Statistical Analysis System (IML/SAS) and the method of Biplot analysis is illustrated with an example.

ADDITIONAL INDEX WORDS

Multivariate analysis, classification, numerical taxonomy, biplot, principal component analysis.

INTRODUCCION

En las últimas décadas ha habido grandes cambios en las metodologías utilizadas en las áreas de sistemática y estudio de las poblaciones biológicas; muchas de ellas son de tipo cuantitativo y emplean una serie de disciplinas como son análisis multivariado, cómputo, taxomía numérica, etc. La clasificación de organismos generalmente implica el análisis conjunto de un gran número de caracteres con base en técnicas estadísticas multivariadas. La mayoría de los usuarios de estas técnicas inician el proceso de clasificación con una matriz que contiene información acerca de las características o propiedades de un número de unidades taxonómicas (razas, especies, variedades, etc.); a partir de la matriz de datos se calculan diferentes medidas de similitud entre todos los pares de unidades taxonómicas y la información se resume en términos de conjuntos de objetos similares a través de análisis de agrupamiento. Paralelamente, se puede elegir un método de ordenación. Los métodos de ordenación tratan de reducir la dimensionalidad del problema, es decir, de todo el conjunto de características,

[1] Recursos Genéticos del Campo Experimental Centro de Jalisco del INIFAP. Apartado Postal No. 10, Tlajomulco de Zuñiga, Jalisco.

Fecha de recepción: 7 de abril de 1995

simplificar el problema en un menor número de variables, para finalmente lograr un examen visual de los datos multivariados en una gráfica simple, de preferencia con no más de dos dimensiones.

La técnica multivariada conocida como Biplot (Gabriel's Biplot) es una opción poco usada en aspectos de clasificación en México. Esta presenta de manera gráfica al mismo tiempo las similitudes entre las unidades taxonómicas, las relaciones entre las variables que caracterizan las unidades taxonómicas y los valores relativos de las observaciones para cada variable. El método Biplot es general y se ha aplicado en diferentes campos de la ciencia, como la Meteorología (Gabriel, 1972), la Toxicología (Shy-Modjeska *et al.*, 1984) y en estudios de la interacción genotipo-ambiente (Murillo, 1988; Zobel *et al.*, 1988; Shafii *et al.*, 1992).

En este trabajo se presentan los detalles teóricos más relevantes del Análisis de Componentes Principales, su relación con el análisis Biplot y estructura de la matriz de datos. Se propone un programa escrito en lenguaje matricial del Sistema de Análisis Estadístico (Interactive Matrix Languaje, IML-SAS) para el análisis Biplot y se incluye un ejemplo para ilustrar el método.

TEORIA Y DEFINICIONES

En esta sección se pretende la descripción de algunos conceptos relacionados con el tema y de los aspectos teóricos relevantes del Análisis de Componentes Principales y Estructura de la Matriz de Datos. Los conceptos se encuentran de manera un tanto dispersa en la literatura, y se conjuntan con el fin de dar mayor claridad al método Biplot.

Sneath y Sokal (1973) definen los términos siguientes:

"**Clasificación**, como el ordenamiento de organismos en grupos con base en sus interrelaciones",

"**Taxonomía**, como el estudio teórico de la clasificación incluyendo sus bases, principios, procedimientos y reglas",

"**Taxonomía Numérica**, como el agrupamiento por medio de métodos numéricos de unidades taxonómicas (taxa) con base en el estado de sus caracteres".

Gabriel (1981) indica que el prefijo "bi" del término BIPLOT no se refiere a una gráfica tradicional de dos dimensiones sino a una gráfica conjunta de hileras y columnas de una matriz de datos. Este autor enfatiza el uso del método Biplot para la inspección de matrices de datos, sobre todo cuando se estudian muchas variables y la inspección visual es poco práctica; en tales casos el Biplot permite examinar la estructura de los datos y no es requisito llevar a cabo análisis estadísticos y pruebas de significancia, aunque de ser necesario es posible calcular valores que permiten la construcción de elipses para ciertos puntos. Estas juegan el papel de la desviación estándar en datos univariados.

Análisis de componentes principales (ACP)

El **ACP** consiste en transformar la serie de variables originales en un nuevo conjunto de variables no correlacionadas, llamadas componentes principales. Esas nuevas variables son combinaciones lineales de las variables originales y se derivan en orden decreciente de importancia (varianza), de tal manera que el primer componente principal es responsable de la mayor proporción posible de la variación con respecto a los datos originales.

Es común que el objetivo del **ACP** sea ver si los primeros componentes pueden

explicar la mayor parte de la variación en los datos originales. Cuando éste es el caso, se supone que la dimensionalidad efectiva del problema es menor que el número total de variables en estudio. En problemas aplicados, un método común es graficar los dos primeros componentes principales para cada taxa, tratando de identificar grupos en los datos. El ACP se define como una técnica matemática que no requiere que el investigador especifique un modelo estadístico para explicar la estructura del error experimental, sin embargo se podrá tener un mejor significado en los componentes en el caso en que las observaciones se ajusten a una distribución normal multivariada (Chatfield y Collins, 1980).

Sea **X** es una matriz de orden n x p, de np observaciones correspondientes a los valores de p variables de cada una de n unidades de estudio. Tal y como se mencionó anteriormente, el ACP consiste en transformar un conjunto de variables $X_1, X_2, \ldots X_p$ a un nuevo conjunto $Y_1, Y_2, \ldots Y_p$. Estas nuevas variables deben tener las propiedades siguientes:

1. Cada Y es una combinación lineal de las X's. Por ejemplo, para el primer componente,

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p = \mathbf{a'_1 x}$$

donde $\mathbf{x} = [\, X_1 \;\; X_2 \;\; \ldots \;\; X_p \,]$ es el vector de valores muestrales de la variables originales, y a_{ij} es el valor del j-ésimo elemento del vector característico $\mathbf{a_1}$ asociado al valor característico más grande λ_1 . En forma matricial para todos los componentes, $\mathbf{Y=X\,A}$, en donde **Y** es la matriz de orden n x p de componentes principales; **A** es una matriz de orden p x p de vectores característicos y **X** es la matriz de orden n x p de observaciones.

2. La suma de cuadrados de los coeficientes a_{ij} para cada i (j= 1,....,p) es la unidad.

3. De todas las posibles combinaciones, Y_1 tiene la máxima varianza:

$$\mathrm{Var}(Y_1) > \mathrm{Var}(Y_2) > \ldots > \mathrm{Var}(Y_p)$$

4. Las Y no están correlacionadas.

Si llamamos **S** a la matriz de varianzas y covarianzas; el primer componente principal Y_1, se obtiene por medio de la elección del vector $\mathbf{a_1}$ de tal manera que Y_1 tenga la varianza máxima posible. Dado que

$$Y_1 = \mathbf{a'_1 X};\;\; \mathrm{Var}(Y_1) = \mathbf{a'_1 S\, a_1}\ , \qquad \ldots\ldots (1) \qquad (1)$$

entonces, el problema es determinar un vector $\mathbf{a_1}$ que maximice $\mathbf{a'_1 S\, a_1}$ sujeto a la condición $\mathbf{a'_1\, a_1 = 1}$. Esta condición se aplica dado que si esto no se hace, entonces $\mathrm{Var}(Y_1)$ puede incrementarse simple y sencillamente incrementando cualquiera de los valores de a_{ij} .

El método tradicional de resolver el problema de maximización de una función de varias variables sujeto a condiciones, es el de Multiplicadores de Lagrange. Para el modelo $Y_1 = a_{11}X_1 + a_{12}X_2 + \ldots\ldots\ldots + a_{1p}X_p = \mathbf{a'_1\, X}$, deseamos escoger $\mathbf{a_1}$ de tal manera que se maximice la varianza de Y_1 sujeta a la condición $\mathbf{a'_1\, a_1 = 1}$. Con base en lo anterior, la función es representada por:

$$F(\mathbf{a_1}) = \mathbf{a'_1 S\, a_1} + \lambda(\mathbf{a'_1\, a_1} - 1\,).$$

Chatfield y Collins (1980) presentan los detalles del problema de maximización de la función $F(\mathbf{a_1})$. Después de obtener la primera derivada e igualandola con el vector nulo se obtiene:

$$(\mathbf{S} - \lambda_1 \mathbf{I})\mathbf{a_1} = 0 \ldots\ldots \qquad (2)$$

Para que sea posible encontrar solución para $\mathbf{a}_1$, además de la solución trivial $\mathbf{a}_1 = 0$, la matriz $(\mathbf{S} - \lambda_1\mathbf{I})$ debe ser una matriz singular; es decir, el valor λ_1 debe ser escogido de tal manera que el determinante $|\mathbf{S} - \lambda_1\mathbf{I}| = 0$. En otras palabras, la solución existe sólo si λ_1 es un valor característico de S y $\mathbf{a}_1$ es el vector característico correspondiente a λ_1. Estas mismas condiciones se aplican al resto de componentes principales Y_i .

Si **L** es la matriz diagonal de valores característicos y **A** la matriz de vectores característicos, con base en las ecuaciones (1) y (2):

$$\text{Var}(Y_i) = \mathbf{a}'_i\mathbf{S}\,\mathbf{a}_i = \lambda_i$$

$$\Sigma\,\text{Var}(Y_i) = \Sigma\,\lambda_i = \text{Traza (L)}, \text{ y}$$

$\mathbf{S = A\,L\,A}$, o $\mathbf{L = A'\,S\,A}$. Así mismo,

$$\begin{aligned}\text{Traza (L)} &= \text{Traza (A' S A)}\\ &= \text{Traza (S A A')}\\ &= \text{Traza (S)}\\ &= \Sigma\,\text{Var}(X_i)\end{aligned}$$

En otras palabras, la suma de las varianzas de las variables originales es la misma que la de sus componentes principales y, consecuentemente, la proporción de la varianza total explicada por el i-ésimo componente principal es $\lambda_i / \Sigma\,\lambda_i$.

Estructura de la matriz de datos

La estructura de correlaciones de una matriz de datos y las dependencias que involucran cualquier número de variables es revelada con base en la descomposición de dicha matriz y obtención de sus valores singulares (**SVD**) o con base en el análisis de los valores característicos de la matriz de sumas de cuadrados y productos (Rawlings, 1988). Con base en análisis **SVD** de la matriz de datos será posible usar el análisis **BIPLOT** (Gabriel, 1972, 1981; Rawlings, 1988), el cual permite el examen visual tanto de la estructura de los datos (varianzas y correlaciones), como de los patrones que existen en los datos (similitud entre unidades taxonómicas). Es importante señalar que **SVD** es el primer paso en el Análisis de Componentes Principales.

Por lo general, tanto el análisis de valores singulares como de valores característicos se llevan a cabo en la matriz de datos estandarizada; la estandarización se lleva a cabo, en primer término, restando la columna de medias (variables) y posteriormente dividiendo los elementos de cada columna por la raíz cuadrada de la suma de cuadrados de desviaciones. En forma matricial:

$$\mathbf{Z = [X-(JM)](STD^{-1})}$$

en donde **X** representa una matriz de datos n x p (correspondiente, por ejemplo, a n unidades taxonómicas con p variables), **Z** es una matriz de orden n x p de datos estandarizados, **J** es un vector de orden n x 1 conteniendo unos, **M** es el vector de medias de variables de orden 1 x p, y $\mathbf{STD^{-1}}$ es una matriz diagonal p x p que consiste de raíces cuadradas de los elementos diagonales de **[X-(J-M)]'[X-(JM)].**

Es posible llevar a cabo la obtención de valores singulares a partir de **Z** o los valores característicos a partir de **Z'Z** (matriz de correlaciones). La matriz **Z** puede descomponerse en su estructura básica (Green y Carroll, 1978; Rawlings, 1988) como:

$$\mathbf{Z}_{n\,x\,p} = \mathbf{U}_{n\,x\,p}\,\mathbf{L}^{1/2}{}_{p\,x\,p}\,\mathbf{V}'_{p\,x\,p} \quad \ldots\ldots\ldots (3)$$

donde **U** y **V** son matrices ortonormales y satisfacen la relación $\mathbf{U'U = I}$, $\mathbf{V'V = I}$,

U es conocida como la matriz izquierda de vectores singulares

V es conocida como la matriz derecha de vectores singulares

L es la matriz diagonal de valores característicos.

I es la matriz identidad.

V puede obtenerse del análisis de valores característicos de **Z'Z;** con base en (3),

$$\mathbf{Z'Z} = (\mathbf{U}\,\mathbf{L}^{1/2}\,\mathbf{V'})'\,(\mathbf{U}\,\mathbf{L}^{1/2}\,\mathbf{V'}) = \mathbf{V}\,\mathbf{L}^{1/2}\,\mathbf{U'U}\,\mathbf{L}^{1/2}\,\mathbf{V'} = \mathbf{V}\,\mathbf{L}\,\mathbf{V'},$$

Después de resolver para **L** y **V** del análisis de **Z'Z, U** puede obtenerse como:

$$\mathbf{U} = \mathbf{Z}\,\mathbf{V}\,(\mathbf{L}^{1/2})^{-1}$$

De esta manera, la estructura de correlación de las variables y los patrones existentes en los datos pueden mostrarse de manera gráfica en un Biplot tal como lo describen Gabriel (1972, 1981) y Rawlings (1988).

Las coordenadas para los vectores que representan las variables (column markers) se obtienen de:

$$\mathbf{T} = \mathbf{L}^{1/2}\,\mathbf{V'}$$

Las coordenadas para las unidades taxonómicas (row markers) son los elementos de **U** provenientes de la descomposición de la estructura básica de la matriz **Z**, es decir:

$$\mathbf{U} = \mathbf{Z}\,\mathbf{V}\,(\mathbf{L}^{1/2})^{-1}$$

Uno de los problemas que poco se discuten en el método Biplot, es acerca de las escalas que requieren las matrices **T** y **U**, es decir, las coordenadas para variables y unidades taxonómicas que servirán para obtener el Biplot. Dado que las coordenadas provienen de las matrices **T** y **U** respectivamente, las cuales se ponderan de manera diferente por los valores característicos, λ, algunos autores como Rohlf (1993) sugieren multiplicar **U** por $\mathbf{L}^{1/2}$. Así, las coordenadas para las unidades taxonómicas serán los elementos de:

$$\mathbf{W} = \mathbf{Z}\,\mathbf{V}\,(\mathbf{L}^{1/2})^{-1}\,\mathbf{L}^{1/2} = \mathbf{Z}\,\mathbf{V},$$

en donde **W** es la matriz de componentes principales.

La longitud del i-ésimo vector-variable en la gráfica de dos dimensiones, relativo a su longitud en el espacio original de p dimensiones puede calcularse por medio de la correlación entre las variables originales y los componentes principales o por el valor de la correlación al cuadrado. Para el primer componente principal y la primera variable, la correlación al cuadrado es:

$$\rho^2_{1i} = \frac{\lambda_1 v^2_{1i}}{SS_i}$$

donde λ_1 es el primer valor característico, v_{1i} es el i-ésimo elemento del primer vector característico y SS_i es la suma de cuadrados de la i-ésima variable estandarizada.

Interpretación del Biplot

Rawlings (1988) da los elementos clave para la interpretación del Biplot de la manera siguiente:

1. La longitud de cada vector-variable en un Biplot, relativo a su longitud en el espacio original de p dimensiones indica qué tan bien se representa dicho vector en la gráfica de dos dimensiones. Los vectores que no caen cerca del plano definido por los dos primeros componentes principales, proyectarán en el Biplot vectores mucho más cortos en relación al espacio original de p dimensiones; para tales variables, dicho Biplot en particular tendrá una representación pobre de las relaciones entre variables y las interpretaciones relacionadas con ellas deberán evitarse.

2. El ángulo entre dos vectores-variable reflejará su correlación en una proyección de dos dimensiones. La correlación es el

coseno del ángulo; un ángulo de 90^0 indicará cero correlación; un ángulo de 0^0 ó de 180^0 indica una correlación de 1.0 ó -1.0 respectivamente. Cabe aclarar que los ángulos entre vectores se traducen en correlaciones sólamente debido a que las variables fueron centradas (a cada valor se le restó la media) antes de que se llevara a cabo el análisis.

3. La proximidad espacial de observaciones individuales entre las unidades taxonómicas refleja sus similitudes; puntos cercanos tienen valores similares y viceversa.

4. Los valores relativos de las observaciones para una variable en particular pueden verse proyectando los puntos de las observaciones sobre el vector que representa las variables, sea esta proyección en sentido positivo o negativo. El vector señala en la dirección de los valores mayores para la variable.

DESCRIPCION DEL PROGRAMA

El programa para el análisis Biplot (Cuadro 1) fue desarrollado para computadoras personales que usan el sistema operativo DOS; está escrito en el lenguaje matricial del Sistema de Análisis Estadístico (SAS/IML). El IML (Interactive Matrix Languaje) es un lenguaje bastante flexible y sencillo que se ajusta a las necesidades de programación del usuario; sin embargo, es recomendable consultar el manual de referencia (SAS/IML Guide for personal computers, 1985), sobre todo para hacer modificaciones y mejoras al programa aquí propuesto. Con el fin de verificar la confiabilidad del programa, se usaron los datos y resultados presentados por Gabriel (1981).

El programa requiere un cuadro de doble entrada, en donde las hileras corresponden a las unidades taxonómicas (razas, especies, etc.) y las columnas a las variables; cada uno de los valores del cuadro deberá corresponder al promedio sobre repeticiones y ambientes. Debido a limitaciones del SAS, el IML sólo permite trabajar con matrices de 4096 elementos, es decir, matrices de 100 x 40, 200 x 20, o dimensiones equivalentes.

Para usar el programa sólo se requieren modificar las instrucciones **data** e **input** en función del conjunto particular de datos. Con el fin de facilitar el uso y/o modificación del programa, se incluyen comentarios en las instrucciones de programación que se consideraron relevantes; dichos comentarios aparecen encerrados entre los caracteres /* y */, los cuales pueden ser escritos en cualquier lugar y no tienen efecto sobre las instrucciones de programación.

Las matrices de coordenadas para caracteres y unidades taxonómicas pueden usarse para construir el Biplot dentro de **PROC PLOT** de SAS o pueden ser importados a otros programas como Lotus 1-2-3, Quattro Pro, Excell, etc. dado que IML genera archivos ASCII.

Ejemplo

La matriz de datos **X**, proviene de la caracterización morfológica de 43 accesiones de teocintle del Banco de germoplasma de Maíz del INIFAP y del Colegio de Posgraduados, Montecillo, Méx. (Cuadro 2). Las accesiones fueron sembradas durante el ciclo primavera-verano de 1991 en Tlajomulco, Jal. en condiciones de temporal. Cada accesión fue sembrada en un surco de 10 m con 25 plantas cada uno; se utilizó un diseño experimental de bloques al azar con dos repeticiones. El espaciamiento de los surcos fue de 0.80 m y entre plantas de 0.40 m. Se fertilizó y se mantuvo libre de malezas como si se tratara del cultivo de maíz.

Caracteres medidos

Los caracteres medidos comprendieron cuatro categorías: (1) caracteres vegetativos

Cuadro 1. Programa en IML para el Método Biplot.

```
/* ANALISIS BIPLOT PROPUESTO POR GABRIEL (1972).
  NOTA: EN LOS CASOS EN QUE SE TRABAJA CON VARIABLES MEDIDAS EN LA MISMA
        ESCALA, SERA NECESARIO SUPRIMIR LAS INSTRUCCIONES 4, 5 Y 6  */

data a;infile 'b:\tla91.x14' lrecl=158;
input (col ncol alt fm nh ah hij ram lpr lrp aes les agr lgr roy pse)
  (5. $16. 13*10.5 6.1);drop col agr lgr;run;
PROC IML WORKSIZE=200;RESET LINESIZE=80 PAGESIZE=55 NOPRINT NOLOG;
USE A;
READ ALL  VAR _NUM_ INTO X[COLNAME=NAMEX ];
USE A;READ ALL VAR _CHAR_ INTO RAZA;
START;
N=NROW(X);                  /*DIMENSION DE X*/
MEAN=X[+, ]/N;              /*MEDIAS POR COLUMNA*/
X=X-J(N,1)*MEAN;            /*CENTRAR X A MEDIA CERO*/
SS=X[##, ];                 /* SUMAS DE CUADRADOS  (4)  */
STD=SQRT(SS');              /*  MATRIZ DE RAIZ CUAD. DE SS  (5)  */
Z=X*DIAG(1/STD);            /*MATRIZ ESTANDARD    (6)  */
ZPZ=Z'*Z;                   /*MATRIZ DE CORRELACION CUANDO SE ESTANDARIZA*/
VAR=VECDIAG(ZPZ);
CALL EIGEN( E,V,ZPZ);       /*EIGENANALYSIS DE LA MATRIZ ZPZ, */
L=SQRT(ABS(E));             /* E=VALORES CARACTERISTICOS (EIGENVALUES) DE Z'Z  */
L12=DIAG(L);                /* MATRIZ DIAGONAL DE VALORES SINGULARES*/
U=Z*V;                      /* V=RIGHT EIGENVECTORS (DE Z'Z) , U = W=  COMPONENTES
                                                                    PRINCIPALES*/
T=L12*V';                   /* COLUMN MARKERS-COORDENADAS PARA LOS CARACTERES-*/
PX=T[1:3,];  CORDX=PX';  CMARK=(CORDX)##2;
CMARK1=CMARK[,1];PCMARK1=CMARK1[,+]/VAR;
CMARK2=CMARK[,1:2];PCMARK2=CMARK2[,+]/VAR;
PCMARK3=CMARK[,+]/VAR;
FREE T PX V ;
TOTAL=SUM(E);
PROPOR=E/TOTAL;             /* PROPORCION PARA CADA EIGENVALUE */
PROP1=SUM(E[1,])/TOTAL;     /* PROPORCIONES ACUMULATIVAS*/
PROP2=SUM(E[1:2,])/TOTAL;
PROP3=SUM(E[1:3,])/TOTAL;
PROP4=SUM(E[1:4,])/TOTAL;
PROP5=SUM(E[1:5,])/TOTAL;
PROP6=SUM(E[1:6,])/TOTAL;
```

Continúa.....

Cuadro 1. Continuación...

```
CORDU=U[,1:4];              /*ROW MARKERS-COORDENADAS PARA UNIDADES
                            TAXONOMICAS-MATRIZ- U */
FREE U;
NAME3={U1 U2 U3 U4};
CP13=PCMARK1 || PCMARK2 || PCMARK3; /* MATRIZ DE VALORES CP*/
NAME4={CM1 CM2 CM3 };NAME5={CP1 CP2 CP3};
PRINT , "VALORES CARACTERISTICOS Y PROPORCIONES",,
 E[FORMAT=8.4] PROPOR[FORMAT=12.5],,
"-COORDENADAS PARA CARACTERES (L1/2*V')",,
 CORDX[R=NAMEX C=NAME4 FORMAT=8.4],
,"-COORDENADAS PARA UNIDADES TAXONOMICAS (MATRIZ U=W)",,
 CORDU[ R=RAZA C=NAME3 FORMAT=8.4];
PRINT ,,"PROPORCIONES ACUMULATIVAS",,
 "PC1" PROP1,"PC2" PROP2,"PC3" PROP3,"PC4" PROP4,"PC5" PROP5,"PC6" PROP6 ;
PRINT "VALORES -CP- (CORRELACIONES ENTRE LOS COMPONENTES",
 " PRINCIPALES Y LAS VARIABLES ORIGINALES)",,
CP13[R=NAMEX C=NAME5 FORMAT=10.4] ;
CREATE PRX FROM CORDX;     APPEND FROM CORDX;CLOSE PRX;
CREATE PRU FROM CORDU;     APPEND FROM CORDU;CLOSE PRU;
FINISH;RUN;
PROC PLOT DATA=PRX;PLOT COL2*COL1='*' / HREF=0 VREF=0;RUN;
PROC PLOT DATA=PRU;PLOT COL2*COL1='*' / HREF=0 VREF=0;RUN;
```

de la planta, (2) caracteres de "mazorca" y grano, (3) caracteres de espiga, y (4) caracteres de espiguilla para un total de 27 variables. En cada unidad experimental, aproximadamente tres semanas después de la siembra, se escogieron al azar 10 plantas dentro de cada parcela, las que fueron marcadas con pintura; posteriormente, se etiquetaron cinco plantas que constituyeron la muestra para la obtención de los datos. Las espigas fueron obtenidas del tallo principal una vez que la floración fue completa; se etiquetaron, secaron y prensaron para las mediciones subsecuentes.

Para ilustrar el análisis Biplot se eligieron las 12 variables siguientes:

Altitud del sitio de recolección (**ALT**),

Días al 50% de floración masculina (**FM**),

Número total de hojas (**NH**),

Ancho de hoja (**AH**),

Numero de hijos (**HIJ**),

Número de ramas de la espiga (**RAM**),

Longitud de la parte ramificada de la espiga(**LPR**),

Longitud de la rama principal de la espiga(**LRP**),

Ancho de espiguilla (**AES**),

Longitud de espiguilla (**LES**),

Reacción a roya (*Puccinia sorghi*)(**ROY**),

Peso de 100 semillas (**PSE**).

Cuadro 2. Accesiones de teocintle usadas en el estudio.

NUM.	CLAVE	SITIO DE RECOLECCION	RAZA
1	JSG Y LOS-95	SAN PEDRO NEXAPA, MEX.	CHALCO
2	JSG Y LOS-99	JUCHITEPEC, MEX.	CHALCO
3	JSG Y LOS-103	CHALCO, MEX.	CHALCO
4	JSG Y LOS-135	BOYEROS, MEX.	CHALCO
5	JSG Y LOS-48	CHURINTZIO, MICH.	MESA CENTRAL
6	JSG Y LOS-53	CHUCANDIRO, MICH.	MESA CENTRAL
7	JSG Y LOS-55	MANUEL DOBLADO, GTO.	MESA CENTRAL
8	C-19-78	URIANGATO, GTO.	MESA CENTRAL
9	JSG Y LOS-86	PUENTE GAVILANES, DGO.	DURANGO
10	JSG Y LOS-88	FRANCISCO VILLA, DGO.	DURANGO
11	JSG Y LOS-83	NABOGAME, CHIH.	NOBOGAME
12	JSG Y LOS-106	MAZATLAN, GRO.	BALSAS
13	JSG Y LOS-109	PALO BLANCO, GRO.	BALSAS
14	JSG Y LOS-111	IXCATEOPAN, GRO.	BALSAS
15	JSG Y LOS-117	ALCHOLOA, GRO.	BALSAS
16	JSG Y LOS-124	BENITO JUAREZ, MICH.	BALSAS
17	JSG Y LOS-130	TZITZIO, MICH.	BALSAS
18	JSG Y LOS-172	COLORINES, MEX.	BALSAS
19	JSG Y LOS-165	PALMAR CHICO, MEX.	BALSAS
20	C-17-78	ERENDIRA, MICH.	BALSAS
21	JSG Y LOS-15	PIEDRA ANCHA, JAL.	*Zea perennis*
22	JSG Y LOS-17	LA MESA, JAL.	*Zea perennis*
23	JSG Y LOS-68	MANANTLAN, JAL.	*Zea diploperennis*
24	MAS-13	LA VENTANA, JAL.	*Zea diploperennis*
25	U DE G	LAS JOYAS, JAL.	*Zea diploperennis*
26	JSG Y LOS-128	CIUDAD HIDALGO, MICH.	CHALCO ?
27	JSG-194	OPOPEO, MICH.	CHALCO ?
28	JSG Y LOS-134	EL SALITRE, MICH.	MESA CENTRAL ?
29	JSG Y LOS-75	COJUMATLAN, MICH.	MESA CENTRAL ?
30	JSG Y LOS-5	SAN JERONIMO, JAL.	MESA CENTRAL ?
31	JSG Y LOS-142	EL SAUCITO, JAL.	BALSAS
32	JSG Y LOS-159	MALINALCO, MEX.	BALSAS
33	JSG-183	AMATLAN, MOR.	BALSAS
34	JSG-205	EL COYOTOMATE, JAL.	BALSAS
35	JSG Y LOS-40	ZACATONGO, JAL.	BALSAS
36	JSG Y LOS-43	EL TABLILLO, JAL.	BALSAS
37	JSG Y ALA-197	SAN CRISTOBAL HONDURAS, OAX.	BALSAS
38	JSG-196	TARETAN, MICH.	BALSAS
39	JSG-202	JIROSTO, JAL.	BALSAS

Continúa.....

Cuadro 2. Continuación...

40	JSG Y LOS-74	LOS CIMIENTOS, JAL.	BALSAS
41	MAS-11	EL MOLINO, JAL.	BALSAS
42	JSG-200	LA LIMA-EL RODEO, JAL.	BALSAS
43	MAS-15	EL SAUZ, COL.	BALSAS

RESULTADOS

Los resultados proporcionados por el programa del Cuadro 1 se presentan en el Cuadro 3 e incluyen: valores característicos y proporciones acumulativas de cada valor característico, coordenadas para los caracteres ($T = L^{1/2} V'$), coordenadas para unidades taxonómicas ($W = Z V$), y valores de correlación entre los componentes principales y las variables originales (CP).

El análisis Biplot tuvo un ajuste en dos dimensiones de 0.77, es decir, el Biplot de la Figura 1 muestra un 77% de la suma de cuadrados de las variables originales, con un 58% para el primer componente y un 19% para el segundo. Los valores de correlación al cuadrado (CP2) son predominantemente mayores de 0.7, es decir, los vectores-variable en el Biplot tienen una adecuada representación en relación al espacio original de las p dimensiones, por lo que las interpretaciones relacionadas a dichos vectores podrán hacerse con un buen grado de confianza.

Las colecciones de teocintle se dividen en dos grupos (Figura 1), uno de ellos situado en el sector negativo del primer componente principal incluye las poblaciones clasificadas como razas Chalco, Mesa Central, Durango, Nabogame y los teocintles perennes. El segundo grupo incluye las colecciones clasificadas como raza Balsas, las cuales forman un grupo heterogéneo que muestra gran variación y se encuentra en los dos cuadrantes definidos por la escala positiva del componente 1.

Los subgrupos que pueden identificarse en la Figura 1 son:

(1). Chalco, Mesa Central y Durango, (2). Nobogame, (3). *Zea perennis*, (4). *Zea diploperennis*, (5). Cojumatlán, (6). Balsas.

Los resultados mostrados por el análisis Biplot concuerdan en gran medida con los presentados por Doebley (1983) y Doebley *et al.* (1984) con base en datos de morfología de espiga e isoenzimas, respectivamente. Basado en varias fuentes de información, Doebley (1990) dividió a los teocintles en dos secciones. La sección Zea incluye dos subespecies: *Zea mays* ssp. *mexicana* que agrupa a las razas Chalco, Mesa Central, Durango y Nobogame y *Zea mays* ssp. *parviglumis* para la raza Balsas. La sección *Luxuriantes* incluye a los teocintle perennes *Zea perennis* y *Zea diploperennis*. La diferencia más notable de los resultados de la Figura 1 respecto a los trabajos de Doebley es la ubicación de la raza Nobogame, la cual parece no pertenecer a *Zea mays* ssp. *mexicana* como lo propone Doebley o pudiera representar una subespecie adicional.

De acuerdo con la dirección de los vectores variable (Figura 1), el primer grupo está constituido por colecciones provenientes de las localidades de mayor altitud (1880 a 2600 msnm), los mayores valores de longitud de la parte ramificada de la espiga (LPR), tamaño de semilla (PSE) y espiguilla (AES y LES). Por su parte, el segundo grupo se caracteriza por plantas con mayor número de ramas de la espiga (RAM),

Cuadro 3. Resultados del análisis Biplot.

VALORES CARACTERISTICOS Y PROPORCIONES

E	PROPOR
6.9333	0.57778
2.3089	0.19241
1.0274	0.08562
0.4446	0.03705
0.3566	0.02972
0.2980	0.02483
0.2194	0.01828
0.1846	0.01538
0.0948	0.00790
0.0685	0.00571
0.0485	0.00404
0.0153	0.00128

PROPORCIONES ACUMULATIVAS

	PROP1
PC1	0.5777778
	PROP2
PC2	0.7701867
	PROP3
PC3	0.8558031
	PROP4
PC4	0.8928563
	PROP5
PC5	0.9225741
	PROP6
PC6	0.9474059

-COORDENADAS PARA CARACTERES (L1/2*V')E

CORDX	CM1	CM2	CM3
ALT	-0.8304	0.0839	0.2094
FM	0.7914	-0.3696	0.4492
NH	0.8702	-0.1756	0.4067
AH	0.0954	0.8432	0.3554
HIJ	0.4753	-0.7485	0.0438
RAM	0.8588	0.3698	0.0686
LPR	0.5908	0.7398	-0.0624
LRP	-0.8691	0.2273	-0.2568
AES	-0.9164	-0.2009	0.1790
LES	-0.9046	-0.1145	0.0445
ROY	0.7799	0.1433	-0.3291
PSE	-0.7191	0.2312	0.5208

Continúa...

Cuadro 3. Continuación...

-COORDENADAS PARA UNIDADES TAXONOMICAS (MATRIZ U=W)

CORDU	U1	U2	U3	U4
JSG Y LOS-95	-0.5206	0.2620	0.1724	0.0199
JSG Y LOS-99	-0.5374	0.1656	0.1000	0.0028
JSG Y LOS-103	-0.4407	0.1545	0.0444	0.0064
JSG Y LOS-135	-0.5638	0.2170	0.1652	-0.0200
JSG Y LOS-48	-0.2345	0.1118	-0.1568	0.0801
JSG Y LOS-53	-0.5165	0.0583	-0.0388	-0.0555
JSG Y LOS-55	-0.6721	0.1821	0.0045	-0.1034
C-19-78	-0.3733	0.2005	-0.0028	0.0375
JSG Y LOS-86	-0.4185	-0.0565	-0.1683	0.1047
JSG Y LOS-88	-0.3629	-0.0456	-0.2147	0.1082
JSG Y LOS-83	-0.4590	-0.2802	-0.3392	0.0187
JSG Y LOS-106	0.4501	-0.1434	0.0373	0.1672
JSG Y LOS-109	0.4639	-0.1836	-0.0254	0.1069
JSG Y LOS-111	0.0727	0.3818	0.1354	0.0222
JSG Y LOS-117	0.1720	0.1478	-0.0339	-0.0371
JSG Y LOS-124	0.3953	0.1904	0.0407	0.0185
JSG Y LOS-130	0.4000	0.0689	-0.1489	0.0805
JSG Y LOS-172	0.3157	0.2391	0.0003	0.0853
JSG Y LOS-165	0.5544	0.0518	0.1134	-0.0287
C-17-78	0.5016	-0.0726	-0.0991	-0.1021
JSG Y LOS-15	-0.2366	-0.6868	0.1440	0.1648
JSG Y LOS-17	-0.2948	-0.5888	0.1800	-0.0175
JSG Y LOS-68	-0.3208	-0.3329	0.1841	-0.1727
MAS-13	-0.2524	-0.3495	0.2204	-0.0777
U DE G	-0.3015	-0.3625	0.0830	-0.0908
JSG Y LOS-128	-0.5290	0.1534	0.0942	-0.0547
JSG-194	-0.4891	0.2794	0.2184	-0.0229
JSG Y LOS-134	-0.0817	0.0147	-0.2187	-0.1055
JSG Y LOS-75	-0.2542	-0.2050	-0.3264	0.0924
JSG Y LOS-5	-0.0136	-0.0129	-0.2746	-0.0413
JSG Y LOS-142	0.3636	0.1604	0.0518	0.0229
JSG Y LOS-159	-0.0371	0.0414	-0.1945	0.1497
JSG-183	0.1934	0.0941	0.0270	0.2094
JSG-205	0.4163	0.1154	0.0413	0.0090
JSG Y LOS-40	0.2677	0.1038	-0.2681	-0.2589
JSG Y LOS-43	0.1840	0.2004	-0.1963	-0.1898
JSG Y ALA-197	0.2645	-0.1329	0.0709	0.0313
JSG-196	0.5025	-0.0719	0.0692	0.0991
JSG-202	0.6406	-0.1901	0.0370	-0.1189
JSG Y LOS-74	0.5999	-0.2281	-0.0187	-0.1686
MAS-11	0.4859	0.0190	0.1589	-0.0618
JSG-200	0.2551	0.3192	0.1454	0.0951
MAS-15	0.4106	0.0106	0.1861	-0.0045

Continúa.....

Cuadro 3. Continuación...

VALORES -CP- (CORRELACIONES ENTRE LOS COMPONENTES PRINCIPALES Y LAS VARIABLES ORIGINALES)

CP13	CP1	CP2	CP3
ALT	0.6896	0.6966	0.7405
FM	0.6263	0.7629	0.9647
NH	0.7573	0.7881	0.9535
AH	0.0091	0.7200	0.8464
HIJ	0.2259	0.7862	0.7881
RAM	0.7375	0.8743	0.8790
LPR	0.3490	0.8963	0.9002
LRP	0.7553	0.8070	0.8729
AES	0.8397	0.8801	0.9121
LES	0.8184	0.8315	0.8335
ROY	0.6082	0.6287	0.7370
PSE	0.5171	0.5705	0.8417

altamente susceptibles a roya de la hoja (ROY), mayor ciclo vegetativo FM), mayor número de hojas (NH) y más hijos por planta (HIJ). Ancho de la hoja (AH) tiende a presentar los mayores valores en la raza Chalco y algunas colecciones de Balsas, con menor expresión en *Zea perennis*.

Lo anterior se confirma calculando los promedios por raza (datos no presentados), los cuales indican lo siguiente:

Nobogame se distribuye a 1850 msnm, es la raza más precoz de las estudiadas con 55 días a floración, pocas hojas (9), pocos hijos por planta (4), tamaño medio de semilla (5 g/100 semillas), pocas ramas de la espiga (7), rama principal de la espiga de 13 cm, espiguilla mediana (2.0 por 8.3 mm), y resistente a roya. **Chalco y Mesa Central** (incluyendo **Durango**) son de ciclo intermedio (65 días a floración), número de hojas (11-12), ancho de hoja (5-6 cm), número de hijos (1-3), tamaño grande de semilla (9.5-12 g/100 semillas), número de ramas de la espiga (20-23), rama principal de la espiga de 15-17 cm, tamaño de espiguilla similar que Nobogame y resistentes a roya.

Como se mencionó anteriormente, **Balsas** presenta gran variación, sin embargo, algunas de las características que en promedio los distinguen sobre el resto son: tardíos (94 a 107 días a floración), gran número de hojas por planta (19-22), gran número de hijos (4 a 14), grano pequeño a mediano (3.2 a 6.5 g/100 semillas), ramas de la espiga desde 22 hasta 94 en promedio, con plantas de hasta más de 200, rama principal de la espiga de 5 a 14 cm y espiguillas pequeñas de 5.6 a 6.7 mm de largo y 1.5 a 1.8 mm de ancho y altamente susceptibles a roya de la hoja.

Con respecto a los teocintles perennes, se puede mencionar que son tardíos (104 días a floración) con la particularidad de que aparecen primero los estigmas que las anteras; número alto de hijos promedio (7-18), hojas angostas (3-4 cm), pocas ramas de la espiga (4-7), espiguillas grandes (2.2 por 9 mm) y resistentes a roya.

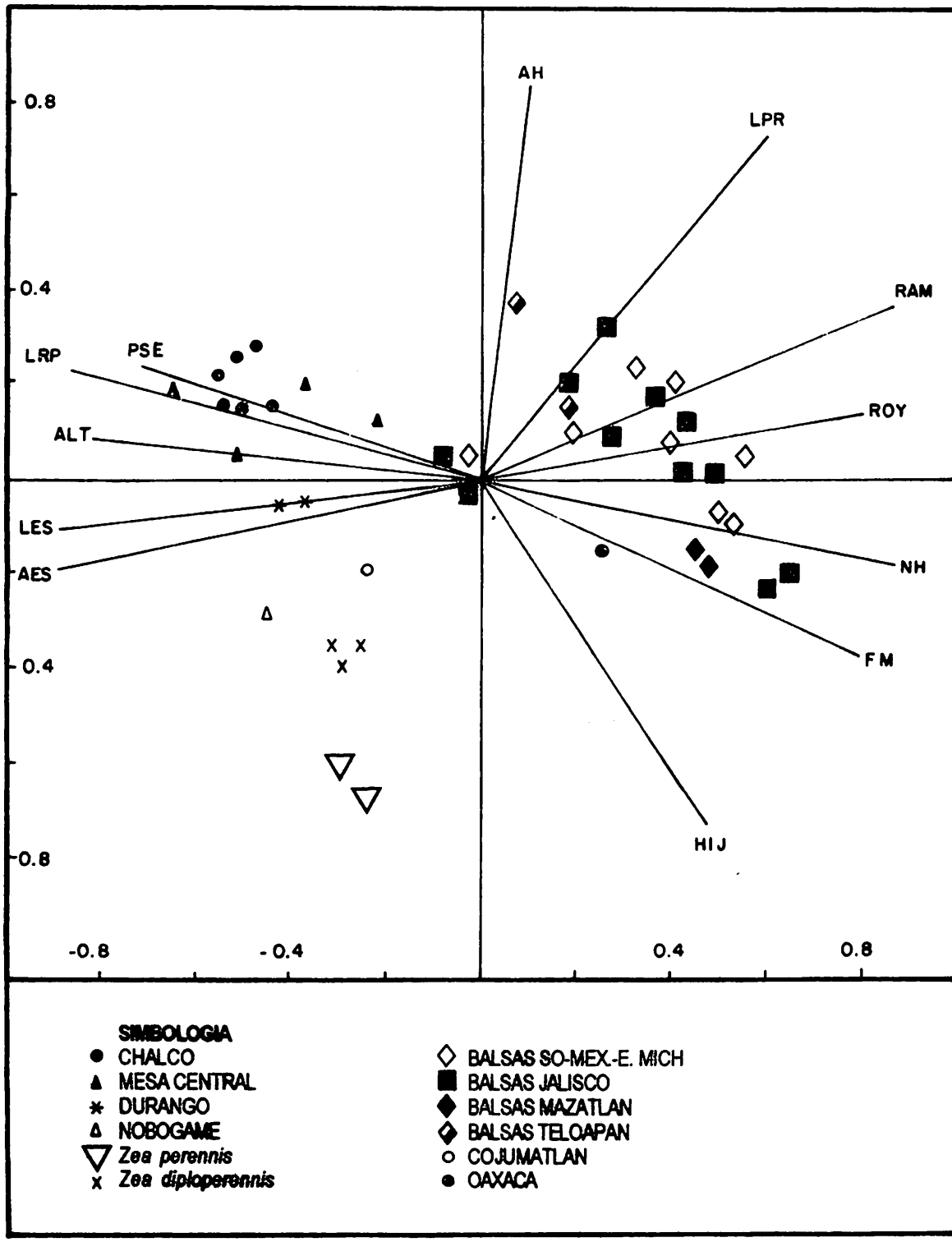

Figura 1. Biplot de datos morfológicos y altitud en teocintle mexicano.

Finalmente, con respecto a correlaciones se pueden identificar en la Figura 1 las siguientes:

(1). Positivas: peso de 100 semillas (PSE) con longitud de la rama principal de la espiga (LRP); longitud (LES) con ancho de espiguilla (AES); número de hojas (NH) con días a floración masculina (FM).

(2). Negativas: peso de 100 semillas (PSE) y longitud de la rama principal de la espiga (LRP) con número de hojas (NH) y días a floración masculina (FM) ; altitud del sitio de recolección (ALT) con número de hojas (NH).

COMENTARIOS GENERALES

En este trabajo se muestra el método Biplot como una técnica multivariada exploratoria en el análisis de la estructura de una matriz de datos compuesta de unidades taxonómicas y los valores de las variables que los caracterizan. La utilidad del método será más evidente si se toma en cuenta que puede reemplazar la necesidad de usar gráficas múltiples, análisis de agrupamiento y cuadros de coeficientes de correlación. La representación gráfica permite determinar visualmente con cierto grado de confiabilidad si existen patrones entre las unidades taxonómicas como resultado de los valores de sus variables, qué variables separan los grupos definidos y qué relación existe entre las variables.

Otra de las aplicaciones del método Biplot se encuentra en el área del análisis de la interacción genotipo-ambiente (Zobel *et al.*, 1988). En estos análisis, la matriz de datos es representada por los efectos de interacción genotipo-ambiente, para una variable en particular. En este tipo de aplicación, la magnitud de la interacción con la que contribuye cada ambiente se representa por la longitud de cada vector a partir del centro del Biplot (Murillo, 1988); la similitud entre ambientes puede determinarse por la magnitud del ángulo entre pares de vectores. Los vectores altamente correlacionados indican que las unidades taxonómicas mostraron respuesta similar en dichos ambientes; para los casos de ausencia de correlación o alta correlación negativa, las unidades taxonómicas responden de manera diferente en dichos ambientes. Por otra parte, la mayor distancia del centro del Biplot de las unidades taxonómicas, indica mayor interacción genotipo-ambiente. La cercanía de las unidades taxonómicas a determinados vectores, indicará aquellos ambientes en donde dichas unidades exhibieron mayores cambios de rango; lo opuesto ocurre para los vectores en dirección contraria a las unidades taxonómicas.

LITERATURA CITADA

Chatfield, C. and A.J. Collins. 1980. Introduction to Multivariate Analysis. Chapman and Hall, London. 246 pp.

Doebley, J.F. 1983. The maize and teosinte male inflorescence: a numerical taxonomic study. Ann. Missouri Bot. Gard. 70: 32-70.

Doebley, J.F. 1990. Molecular evidence and the evolution of maize. Econ. Bot. 44: 6-27.

Doebley, J.F., M. M. Goodman and C.W. Stuber. 1984. Isoenzymatic variation in *Zea* (Gramineae). Syst. Bot. 9: 203-218.

Gabriel, K.R. 1972. Analysis of meteorological data by means of canonical decomposition and biplots. J. Appl. Meteor. 11: 1071-1077.

Gabriel, K.R. 1981. Biplot display of multivariate matrices for inspection of data and diagnosis. In: V. Barnett (ed.). Interpreting multivariate data. pp. 147-173.

Green, P.E. and J.D. Carroll. 1978. Mathematical tools for applied multivariate analysis. Academic Press, New York. 376 pp.

Murillo, O. 1988. The performance of *Pinus oocarpa* Schiede provenances across environments in South America. Tesis de M.Sc. North Carolina State University, Department of Forestry, Raleigh, N.C. 69 pp.

Rawlings, J.O. 1988. Applied Regression Analysis. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, California. 553 pp.

Rohlf, F.J. 1993. NTSYS-pc, Numerical taxonomy and multivariate analysis system. Exeter Software, New York.

SAS Institute Inc. SAS/IML User's Guide for personal computers, Version 6 Edition. Cary, NC: SAS Institute Inc., 1985. 243 pp.

Shafii, B., K.A. Mahler, W.J. Price and D.L. Auld. 1992. Genotype x environment interaction effects on winter rapeseed yield and oil content. Crop Sci. 32: 922-927.

Shy-Modjeska, J.S., J.E. Riviere and J.O. Rawlings. 1984. Application of Biplot methods to the multivariate analysis of toxicological and pharmacokinetic data. Toxicology and Applied Pharmacology 72: 91-101.

Sneath, P.H.A. and R.R. Sokal. 1973. Numerical Taxonomy. W.H. Freeman and Company, San Francisco. 573 pp.

Tatsuoka, M.M. 1971. Multivariate analysis: Techniques for educational and psychological research. John Wiley & Sons, New York. 310 pp.

Zobel, R.W., M.J. Wright and H.G. Gauch Jr. 1988. Statistical analysis of a yield trial. Agronomy Journal 80: 388-393.