

UNIVERSIDAD DE GUADALAJARA.

Centro Universitario de Ciencias Biológicas y Agropecuarias.

División de Ciencias Agronómicas.

Coordinación de Posgrado.



CUCBA



BIBLIOTECA CENTRAL

La Utilización de Técnicas Estadísticas Multivariable. Estudio de Caso.

Ing. José Pablo Torres Morán.

TESIS

Presentada como requisito parcial
para obtener el grado de
Maestro en Ciencias en
Manejo de Áreas de Temporal

Las Agujas, Zapopan, Jalisco a 28 de noviembre de 2000.

Esta tesis fue realizada bajo la dirección del Consejo Particular indicado, ha sido aprobada por el mismo y aceptada como requisito parcial para la obtención del grado de:

MAESTRO EN CIENCIAS
EN
MANEJO DE ÁREAS DE TEMPORAL

CUCBA



BIBLIOTECA CENTRAL

Consejo Particular

DIRECTOR:

DR. FRANCISCO JAVIER FLORES MENDOZA.

ASESOR:

DR. HUGO MORENO GARCÍA.

ASESOR:

M. C. RICARDO NUÑO ROMERO.

LAS AGUJAS, ZAPOPAN, JALISCO, NOVIEMBRE DE 2000.

“Las Horas que otros pasan dedicados a aquello que les interesa, en festividades o diversiones públicas, en diversos placeres o entregados mental y físicamente al descanso – que otros dedican a banquetes que empiezan a primera hora de la noche, a jugar a los dados o a la pelota- estas horas yo las he dedicado a la incesante práctica de estudios de esta clase. ¿Y quién podría criticarme por esto o indignarse conmigo con justicia?”

Cicerón.

Agradecimientos.

Con gran respeto agradezco a mi consejo particular sus atinadas observaciones al presente trabajo, su tiempo, profesionalismo y dedicación ameritan un reconocimiento profundo. Gracias al Dr. Francisco Javier Flores Mendoza, al Dr. Hugo Moreno García y al M. C. Ricardo Nuño Romero. Adicionalmente, agradezco a mis sinodales del examen de grado, los Señores Dres. Jesús Sánchez González y Mario Abel García Vázquez.

Dedicatoria.

Con todo mi amor dedico este humilde trabajo a Maximina, Andrea, José Pablo, Rodrigo y Josué.



CONTENIDO.

Resumen	v
1.- INTRODUCCIÓN	1
2.- OBJETIVOS	3
2.1.- Generales	3
2.2.- Particulares	3
3.- REVISIÓN DE LITERATURA	5
3.1.- Antecedentes	5
3.2.- Análisis de componentes principales	10
3.2.1.- Orígenes	10
3.2.2.- Utilización	10
3.2.3.- Principios	11
3.2.4.- Proceso	12
3.2.5.- Correlación entre variables y componentes principales	14
3.3.- Análisis de factores	16
3.3.1.- Orígenes	16
3.3.2.- Principios	18
3.3.3.- Supuestos y pruebas	20
3.3.4.- Solución del modelo de análisis de factores	21
3.3.5.- Evaluación de la solución	22
3.3.6.- Interpretación	25
3.4.- Análisis de conglomerados	26
3.4.1.- Orígenes	26
3.4.2.- Objetivos de la taxonomía numérica	26
3.4.3.- Principios	27
3.4.4.- Métodos en el análisis de conglomerados	27
3.4.5.- Proceso	30
3.4.6.- Evaluación de la solución	31
3.5.- Análisis múltiple discriminante	33
3.5.1.- Orígenes	33
3.5.2.- Supuestos y pruebas relacionadas	33
3.5.3.- Procedimientos de selección de variables	35
3.5.4.- Métodos para el análisis múltiple discriminante	37
3.5.5.- Número de funciones discriminantes a seleccionar	38
3.6.- Análisis multivariable de varianza	39
3.6.1.- Orígenes	39
3.6.2.- Utilización del MANOVA	39
3.6.3.- Supuestos en el MANOVA	41
3.6.4.- Pruebas de hipótesis	42
3.6.5.- Análisis <i>post hoc</i>	43

4.- Materiales y métodos.....	45
4.1.- Material genético.....	45
4.2.- Localidad.....	45
4.3.- Variables en estudio.....	47
4.4.- Análisis estadístico.....	49
4.5.- Material de soporte para el procesamiento.....	49
5.- Resultados y discusión.....	53
5.1.- Ajuste marginal a la distribución normal.....	53
5.2.- Correlaciones.....	54
5.3.- Análisis de componentes principales.....	56
5.4.- Análisis de factores.....	64
5.5.- Análisis de conglomerados.....	69
5.6.- Análisis múltiple discriminante.....	72
5.7.- Análisis multivariable de varianza.....	75
6.- Conclusiones.....	85
7.- Literatura citada.....	89

CUCBA



BIBLIOTECA CENTRAL

Índice de tablas.

Tabla 1	Valores de la prueba KMO y su interpretación	23
Tabla 2	Coefficientes de regresión para determinar el número de factores a extraer mediante el procedimiento paralelo	24
Tabla 3	Número asignado a cada variedad y genealogía correspondiente	47
Tabla 4	Características agroclimáticas de Zapopan.....	47
Tabla 5	Niveles de significación en la prueba de Kolmogorov Smirnov de ajuste a la distribución normal	54
Tabla 6	Matriz de correlaciones entre las variables en estudio	56
Tabla 7	Valores propios de la matriz de correlaciones	57
Tabla 8	Cargas de los primeros tres componentes principales.....	57
Tabla 9	Correlaciones entre las variables y los componentes principales	59
Tabla 10	Calificaciones de cada variedad en los tres componentes principales	61
Tabla 11	Descomposición de valores singulares de la matriz de correlaciones.....	63
Tabla 12	Cargas de los primeros tres componentes principales mediante la calculadora HP 48G.....	64
Tabla 13	Medida de adecuación de muestreo de Kaiser Meyer Olkin para las variables.....	65
Tabla 14	Cargas de factor y comunalidades por variable estudiada.....	66
Tabla 15	Prueba RMSR de cada variable	68
Tabla 16	Valores de factor para cada variedad	68
Tabla 17	Centroides por conglomerado e interpretación de los mismos.....	72
Tabla 18	Elección de funciones discriminantes canónicas	73
Tabla 19	Resultados de la clasificación mediante funciones discriminantes	75
Tabla 20	Valores propios de la matriz de correlaciones para suma de cuadrados entre individuos + residual	76
Tabla 21	Suma de cuadrados y productos cruzados de la matriz entre individuos + residual	77
Tabla 22	Matriz ajustada hipotética de cuadrados y productos cruzados para tratamientos.....	78
Tabla 23	Pruebas de significación multivariable	78
Tabla 24	Pruebas F univariable	79
Tabla 25	Valores propios y correlaciones canónicas para el efecto de los tratamientos	80
Tabla 26	Análisis de reducción de dimensiones	81
Tabla 27	Coefficientes de las funciones discriminantes estandarizadas.....	81
Tabla 28	Correlación entre las variables dependientes y canónicas posterior a la rotación varimax	82
Tabla 29	Prueba de Bonferroni para separar promedios con base en la primera función discriminante estandarizada	83

Índice de figuras

Figura 3.2-1	Primeros dos componentes principales	12
Figura 3.4-1	Tipos de dendrograma, según Manly	31
Figura 4.3-1	Algunas características del maíz bajo estudio	48
Figura 5.3-1	Calificación de las variedades en función de los dos primeros componentes principales	60
Figura 5.3-2	Cargas de variable en los primeros componentes principales	60
Figura 5.3-3	Diagrama de flujo que muestra el cálculo de componentes principales con calculadora HP 48G.....	63
Figura 5.4-1	Gráfica <i>Scree</i> para determinar el número de factores a extraer.....	65
Figura 5.4-2	Calificación de factores para las variedades	69
Figura 5.5-1	Dendrograma de las 30 variedades de maíz.....	70
Figura 5.5-2	Número de conglomerados a formar en función de r^2	71

Anexos

Base de datos	93
Vectores de medias por variedad.....	101
Matriz de distancias cuadradas de Mahalanobis	102

CINCUA



BIBLIOTECA CENTRAL

Resumen.

El objetivo de este estudio de caso fue aplicar las técnicas de análisis de componentes principales, análisis de factores, análisis de conglomerados, análisis múltiple discriminante y análisis multivariable de la varianza a un grupo de plantas de maíz, tanto razas como criollos e híbridos, a modo de análisis de datos multivariable, analizando las conclusiones de cada técnica y cómo se complementan unas y otras. Las variables analizadas fueron: altura de mazorca, altura de planta, diámetro de mazorca, diámetro de tallo, número de carreras, número de hojas, longitud de mazorca, peso de grano, peso de mazorca, porcentaje de humedad y profundidad de grano.

El análisis de componentes principales permitió reducir las dimensiones de todos los datos de 11 a 3. El análisis de componentes principales es un paso preliminar en varias técnicas multivariable. Se aplicó un procedimiento para efectuar este análisis mediante una calculadora HP 48G. En principio, la gráfica de los primeros dos componentes principales nos permite identificar en forma aproximada 7 grupos empleando para eso los primeros dos componentes principales. En el primer componente, la correlación entre la variable canónica y las variables originales son todas positivas y sólo se excluye al número de carreras, indicando que este componente se relaciona positivamente con las variables indicadoras de la biomasa. El segundo componente principal correlaciona positivamente con la altura de planta y la altura de mazorca y negativamente con el diámetro de mazorca, dando a entender que las plantas más altas tienen la mazorca a mayor altura y ésta es de diámetro más pequeño. En el último componente la correlación mayor a 0.5 entre las variables canónicas fue con el número de carreras.

En el análisis de factores resultaron significativas tres funciones canónicas. La única variable con un valor mediocre en la escala de Kaiser Meyer Olkin fue el número de carreras. La prueba de RMSR alcanzó un valor total de 0.0793956, por lo cual se considera que el análisis de factores fue adecuado. En el primer factor, las correlaciones más importantes fueron positivas, con diámetro y peso de mazorca, lo cual se interpreta como que entre las variedades estudiadas, el mayor peso de mazorca se relacionó con grosor, más que con longitud. El segundo factor tuvo correlaciones altas con las variables altura de planta, altura de

mazorca, longitud de mazorca y porcentaje de humedad, indicando que las plantas más altas tienen la mazorca más alta, de mayor longitud y mayor porcentaje de humedad en el grano. El tercer factor correlacionó positivamente con número de hojas, número de carreras y diámetro de tallo, así que las plantas con más hojas tuvieron mazorcas con mayor número de carreras y mayor diámetro de tallo.

El análisis de conglomerados empleó las calificaciones de factor de las variedades que se obtuvieron en el AF. Se construyeron 7 grupos, siguiendo las indicaciones preliminares que habían quedado tras el análisis de componentes principales. Los grupos que se conformaron se caracterizaron mediante sus centroides.

El análisis múltiple discriminante se aplicó a los 7 grupos construidos logrando un éxito de 63.39% de los casos.

Finalmente, el análisis multivariante de la varianza mostró que las diferencias entre los grupos son reales, aunque en este análisis se empleó la matriz de varianzas y covarianzas en lugar de la de correlaciones que se usó en los anteriores casos. Se hizo una discriminación de grupos empleando la prueba de Bonferroni.

1 Introducción.

La estadística tiene gran importancia en las ciencias agronómicas como base metodológica para la inferencia y la toma de decisiones. Entre las finalidades de la estadística se encuentra la descripción de grandes conjuntos de datos, la separación de éstos en grupos significativos, la medición de las relaciones entre variables y las pruebas de hipótesis. Las técnicas estadísticas más poderosas y sofisticadas permiten manejar simultáneamente más de dos variables, logrando mejorar sustancialmente el análisis de datos y las conclusiones que de él derivan. Además, ha tomado gran impulso el análisis exploratorio de datos, cuya mayor utilidad consiste en la investigación de diferencias observadas cuando las relaciones causales no son bien comprendidas.

La disposición de técnicas estadísticas multivariantes para el análisis de datos del mundo real, puede guiarnos en el descubrimiento de estructuras latentes de los mismos, cuando las técnicas univariantes fracasan en la búsqueda de diferencias significativas entre ellos. Por el contrario, series de pruebas estadísticas univariantes, con gran número de variables diferentes, pueden llevar al investigador a conclusiones erróneas, debido a la reducción de la probabilidad fiducial con cada variable agregada al estudio. De esta manera, tomando como ejemplo el análisis de varianza, la probabilidad de encontrar por azar diferencias significativas entre tratamientos se incrementa notablemente. Las técnicas multivariantes, tales como la T^2 de Hotelling o el análisis multivariante de varianza, son más poderosas y mantienen los niveles de significación prefijados al estudio.

La cantidad de artículos científicos recientemente producidos que emplean técnicas de análisis multivariante es impresionante. Esta situación permite considerar el auge en el empleo de técnicas multivariantes como algo diferente a una moda pasajera. El mejoramiento de la tecnología de cómputo, la disminución relativa del costo de la misma, la continua producción de software especializado en la ejecución de tareas complejas con un alto rendimiento y el desarrollo de calculadoras gráficas muy poderosas y de bajo precio, han sido factores que promueven un mayor empleo de métodos de análisis sofisticados como apoyo para el desarrollo científico y tecnológico de México.

En la revisión de literatura se mostrará cómo publicaciones recientes emplean la estadística multivariable: en la definición de relaciones taxonómicas entre individuos o poblaciones con objeto de precisar la filogenética entre grupos; en la discriminación de individuos pertenecientes a conjuntos diferentes, pero cuyo parecido hace difícil la clasificación; en la elección de las variables mejores para el establecimiento de métodos de selección o mejoramiento genético; en la identificación de la estructura de la información, para distinguir entre la relevante y la irrelevante.

El estudio de caso será con maíz *Zea mays* L. cultivado en Zapopan, durante 1994. Pocas plantas han sido tan estudiadas en México como el maíz. La idea de utilizar una base de datos de maíz en este trabajo, es mostrar cómo es posible ampliar las conclusiones a las que se puede llegar mediante análisis estadístico multivariable. Las técnicas seleccionadas son el análisis de componentes principales, el análisis de factores, el análisis de conglomerados, el análisis múltiple discriminante y el análisis multivariable de varianza, mismas que pueden complementarse entre sí.

El análisis multivariable en este estudio de caso se aplicará como consecuencia de haber elegido doce variables para cada una de quince plantas, las cuales han sido seleccionadas aleatoriamente de treinta variedades de modo que para cada individuo es factible obtener un vector que lo representa. A diferencia de las técnicas univariadas, que consideran a cada medida de una variable como representación del individuo, en la estadística multivariable cada individuo es representado por un vector de medidas.

CUCEBA



BIBLIOTECA CENTRAL

2. Objetivos.

2.1 Generales.

- Analizar un conjunto de datos de maíz tomando en consideración la totalidad de variables, cada una de las cuales es una nueva dimensión en el espacio R^n y se mide con la intención de lograr una descripción más completa de la muestra.
- Identificar influencias no fácilmente observables en la estructura de los datos.
- Separar objetivamente la muestra en categorías representativas.
- Estimar una función capaz de determinar con cierta precisión la inclusión de una planta individual dentro de cada conjunto.
- Determinar la influencia de las variedades de maíz sobre todas las variables de manera simultánea y separar los vectores de medias correspondientes a los grupos de variedades.

2.2 Particulares.

- Poner a disposición de los investigadores un documento de apoyo para analizar las principales características y métodos de las técnicas en estudio, ilustrando las semejanzas y diferencias entre estos métodos, así como los algoritmos correspondientes.
- Reducir la dimensión de los datos mediante el análisis de componentes principales, sin pérdida significativa de información. Esto es, se pretende

determinar índices compuestos por las variables originales, de modo que unos cuantos de éstos reflejen la mayor variación posible.

- Definir un procedimiento para desarrollar el algoritmo de componentes principales con una calculadora gráfica Hewelett Packard 48G.
- Utilizar los valores de factor obtenidos en el paso 5 con los datos originales estandarizados y destinarlos a la búsqueda de agrupaciones significativas mediante análisis jerárquico de conglomerados.
- Dividir la muestra en los grupos correspondientes del paso 6. Seleccionar mediante muestreo estratificado aleatorio con asignación proporcional la mitad de los datos para construir la función múltiple discriminante, y la otra mitad para verificar los resultados de dicha función.
- Determinar la diferencia de las variedades con todas las variables en estudio mediante el análisis multivariable de varianza.

CUCBA



BIBLIOTECA CENTRAL



3 Revisión de Literatura.

3.1 Antecedentes.

Los fundamentos de muchas técnicas de análisis multivariable de datos empleadas exitosamente en las ciencias biológicas han sido propuestos desde principios del siglo XX. Algunos ejemplos de utilización más recientes se muestran a continuación.

Doeygard S. [11] empleó el análisis de componentes principales para estudiar los caracteres morfológicos en el género *Pachyrhizus* al cual pertenece la jícama. Doscientos treinta y seis valores fueron analizados mediante un conjunto de 23 caracteres cuantitativos. Se encontró que tres componentes principales explicaban el 59% de la variación total de la estructura de interdependencia. Los caracteres florales parecieron ser la mayor fuente de diversidad. El análisis demostró que varias especies formaron grupos homogéneos.

Zheng-Yong Zhan y Wang Hong [37], agruparon 48 líneas mejoradas de maíz con diferentes orígenes en doce grupos a partir de sus distancias genéticas y análisis de conglomerados. El primer grupo fue formado por líneas exóticas o por líneas mejoradas con genes exóticos. Los conglomerados no tuvieron relación con el origen geográfico de los materiales. La distancia no reflejó la divergencia genética entre las líneas mejoradas, sino sus diferencias morfológicas. Existió heterosis para rendimiento de grano, pero ésta no tuvo correlación lineal significativa con las distancias genéticas entre las líneas mejoradas. Se concluyó que las distancias genéticas tienen valor limitado en la predicción de heterosis en maíz.

Bernhardt, Allen y Helmers [3], presentan el análisis de conglomerados como opción para resolver el problema de clasificar unidades de observación para la investigación agrícola multidisciplinaria de sistemas de producción en convencionales o alternativos. Se presenta el problema de las metas específicas de la clasificación, las aproximaciones de los conglomerados al logro de dichas metas y otros varios problemas encontrados. La categorización final se basó en 59 variables de cultivo analizadas y agrupadas en cinco

conjuntos cuyo rango varió desde sistema de monocultivo de maíz en irrigación, hasta agricultura casi orgánica. Los datos proceden de 990 granjas de Nebraska.

Eritja [12] utilizó el análisis múltiple discriminante AMD para estudiar la expresión sexual en el tórax de un individuo dañado, ginandromorfo silvestre, capturado en octubre de 1992 en Baix Llobregat, España. El sexo genotípico de cada ala fue comparado con una gran base de datos de 1,490 individuos de *Culex pipiens* para cuatro parámetros de dimorfismo sexual en alas. Ambas alas fueron clasificadas como hembras. Entonces se concluyó que la línea del límite del tórax no era oblicua, sino transversal. La identificación del sexo en individuos ginandromorfos constituye un problema en las especies con dimorfismo sexual, dado que presentan cualidades de ambos sexos. El análisis MDA probó ser una técnica eficaz cuando se enfrenta un caso tan difícil como este. Además, el análisis logró distinguir la posición de un surco en el tórax que puede ser valioso auxilio morfológico en la solución del problema. El análisis múltiple discriminante, en nuestro caso, sería un método con potencial relevante para distinguir, entre otras cosas, plantas fuera de tipo.

Forveille, Vercauteren y Rutledge [16] aplicaron el análisis estadístico multivariable a datos bidimensionales de extractos polifenólicos (método de resonancia magnética nuclear, NMR por sus siglas en inglés) de vid *Vitis vinifera* L. cosechadas en la región de Burdeos, Francia. Un análisis de varianza detectó los volúmenes de correlación espectral más discriminantes, mismos que fueron sometidos al análisis de componentes principales, análisis jerárquico de conglomerados y análisis discriminante. Los resultados mostraron que los clones estaban divididos en tres grupos de acuerdo al cultivar y que estos podían ser perfectamente diferenciados. Actualmente el NMR bidimensional es la única herramienta analítica que puede diferenciar clones de vid, especies y cultivares y puede ser utilizado en otras plantas que producen polifenoles.

Como puede verse, el empleo conjunto de métodos químicos sofisticados y las técnicas del análisis estadístico multivariable, podrían ser auxiliares en el planteamiento y resolución de problemas complejos. La citada investigación pone de manifiesto la necesidad de que el investigador se empape del conocimiento de las características de su objeto de estudio.

Mallarino *et al.* [20], usaron el análisis de factores para estudiar las relaciones entre muchas variables de sitio y rendimientos de maíz en la faja maicera de los Estados Unidos de América. Las variables correlacionadas se agruparon en: fertilidad de suelo, control de maleza y factores para crecimiento temprano. Su importancia en la explicación de la variabilidad del rendimiento difirió grandemente entre campos.

Radloff [27] utilizó el análisis multivariable como herramienta para identificar poblaciones selectas de abeja melífera, tomando caracteres morfométricos y analizando las feromonas presentes en el aguijón de las obreras, a lo largo de cinco transectos del África. Se emplearon dos formas de agrupación mediante conglomerados y se encontraron coincidencias entre los conjuntos formados con base en las feromonas y mediante morfometría. Se identificaron transectos de alta variabilidad, los cuales indicaron al mismo tiempo una alta diversidad genética y la realización de apareamientos híbridos. Los centroides de alta varianza intercolonial ocurrieron entre y dentro de los límites de los conglomerados y fueron típicos de zonas de inestabilidad o transición ecológico-climatológica por discontinuidades temperatura-lluvia.

El análisis de componentes principales y el análisis discriminante *Stepwise* permitieron la identificación de 3 conglomerados morfométricos correspondientes a *Apis mellifera saharensis* y *A. mellifera intermissa* en Marruecos y *A. m. iberica* (con tres poblaciones biométricas) en España, pero no permitieron identificar conglomerados de feromonas. Los espectros de varianza morfométrica y de feromonas, indicaron regiones de alta hibridación natural a lo largo del transecto Sahara-Pirineos. En el cuerno de África se identificaron poblaciones discretas y estadísticamente homogéneas: *A. m. jemenitica*, *A. m. bandasii* y *A. m. sudanensis* en Etiopía y un grupo no identificado en el suroeste de Somalia. Las áreas de alta variación intercolonial fueron interpretadas como zonas de hibridación entre las poblaciones.

Puede apreciarse que las técnicas multivariadas fueron suficientemente poderosas como para identificar subpoblaciones de abejas, determinar si las causas de variación son genéticas o ambientales y definir lugares con alto grado de hibridación entre grupos.

Lafitte, Edmeades y Taba [19] aplicaron análisis de componentes principales y análisis de conglomerados en la tarea de comparar razas locales de maíz y variedades mejoradas. Partieron de la idea de emplear razas locales de maíz como parte de poblaciones fuente para lograr la adaptación a los ambientes deficientes en nitrógeno (N), porque tradicionalmente, éstas han sido manejadas en ambientes de fertilidad de suelos más restrictivos que los cultivares mejorados. Se probaron durante el verano en Poza Rica, Veracruz, México, comparando los patrones de rendimiento de grano, consumo y tipo de N de 38 razas procedentes del banco de germoplasma del CIMMYT (Centro Internacional de Mejoramiento de Maíz y Trigo) y 26 variedades tropicales mejoradas, tanto en ambientes adecuados como deficientes en N. Encontraron que los cultivares mejorados siempre superaron en rendimientos a las razas locales, en 56% como promedio. Sin embargo, bajo condiciones de N limitado, las variedades mejoradas no fueron consistentemente mejores para recuperación total de N, la fracción del mismo en grano o la biomasa sobre el suelo. Las razas locales tuvieron mayores concentraciones de N en el grano en las dos condiciones del experimento. Se probó que la selección sobre la única base de rendimiento del grano puede ser insuficiente, concluyéndose que las razas locales pueden aportar rasgos útiles para alcanzar una producción estable en condiciones de ambiente limitadas en N. El análisis de componentes principales se empleó para identificar características relacionadas con la comprensión de los patrones de localización del N. Tomando como base estas particularidades, el análisis de conglomerados identificó seis grupos dentro de cada estación de cultivo. Los conglomerados se diferenciaron en grupos de floración temprana y tardía y el análisis de componentes principales se aplicó de nuevo sobre cada subgrupo. Algunos conglomerados se formaron bien bajo N adecuado, pero no cuando éste fue limitante, mientras que otros mostraron una respuesta a la inversa. Esto se interpretó como una adaptación específica a condiciones de N en el ambiente, donde se observó que acumulación de N total y N en el grano variaron independientemente. En algunos casos, los conglomerados revelaron una asociación con las áreas geográficas de las colectas, pero no se encontró asociación entre la precipitación pluvial y los sitios de colecta para un conglomerado y el desempeño del cultivo. Cuando se mejora maíz para rendimientos altos y estables bajo condiciones de N limitado, una estrategia razonable sería desarrollar poblaciones fuente de maduración precoz y tardía desde razas locales que exhiban una gran capacidad de absorción de nitrógeno, enviando grandes cantidades del mismo a la materia seca y el grano y

manteniendo grandes concentraciones de N en el grano bajo condiciones limitantes de abastecimiento de este elemento.

Ribeiro y Nielsen [28] emplearon el análisis múltiple discriminante AMD en registros procedentes de suelos contaminados con Cu, Cr y As, en los cuales se determinó la presencia de estos elementos mediante la extracción secuencial para identificar la distribución de los mismos en la fase sólida del suelo. La presencia de Cu, Cr y As en la fase sólida se indicaba mediante el análisis químico por vía húmeda. Se emplearon técnicas analíticas cualitativas y microfotográficas sobre secciones delgadas de agregados del suelo y se efectuó el análisis sobre áreas marcadas mediante la luz plana polarizada, guardándolas en CD. Grupos selectos de imágenes se sometieron al análisis discriminante para encontrar patrones de reconocimiento de las características del suelo correspondientes a la contaminación identificada efectivamente mediante técnicas cualitativas. Fue posible plantear un procedimiento para estudiar la distribución espacial de los contaminantes considerando las imágenes tomadas en secciones delgadas.

Shaw *et al.* [31], emplearon el análisis de componentes principales para clasificar jugos comerciales de naranja sobre la base del monto de constituyentes volátiles de éstos. Paneles de 10 o más catadores han sido empleados por años para juzgar las diferencias de sabor entre varios tipos de jugo de naranja. Tales paneles consumen tiempo, dinero, son subjetivos y especialmente difíciles si se usan procesadores pequeños de jugo. Este estudio trabajó con 25 componentes del sabor en dos tipos de jugo de naranja, seguidos por un análisis de computadora para mostrar que los diferentes tipos pueden ser separados en dos categorías por este método más rápido y objetivo, que no requiere panel de catadores. Se distinguieron dos grupos: el de jugos pasteurizados no procedentes de concentrados y el de jugos preparados de concentrados congelados. Se probó que los procesadores de jugo ejercen alguna influencia sobre el sabor del mismo. El análisis de componentes principales ayudará a preparar jugos de mejor sabor para el consumidor.

CUCBA



BIBLIOTECA CENTRAL

3.2 Análisis de componentes principales.

3.2.1 Orígenes.

Karl Pearson fue el creador de la metodología del análisis de componentes principales ACP en 1901, la cual habría sido propuesta en principio como un medio de ajustar planos ortogonales por el método de mínimos cuadrados, pero fue en el año de 1930 cuando Hotelling planteó usarla para estudiar estructuras de datos correlacionados. En el ACP no se requiere que los datos conformen una distribución normal multivariante [23].

Pearson consideró que los componentes principales eran la solución a muchos de los problemas biométricos de su tiempo, aunque no propuso algún método práctico de cálculo para más de dos o tres variables. El ACP es una de las técnicas multivariantes más sencillas. Consiste en tomar p variables, X_1, X_2, \dots, X_p y encontrar combinaciones de las mismas tales que se produzcan índices $\xi_1, \xi_2, \dots, \xi_n$ no correlacionados entre sí. Esta condición resulta útil, pues implica que dichos valores miden diferentes dimensiones en los datos. No obstante, los índices están ordenados de manera que la variación del primero supera la del segundo y ésta la del tercero y así sucesivamente: $\text{Var}(\xi_1) > \text{Var}(\xi_2) > \dots > \text{Var}(\xi_n)$. Cada uno de los índices es un componente principal [21].

3.2.2 Utilización.

Wulder [36] menciona dos usos de los componentes principales: el uso directo consiste en la identificación de grupos de variables interrelacionadas para lograr reducir el número de variables; el uso indirecto es como método para la transformación de datos. Es como reemplazar los datos con otros cuyas propiedades no están presentes en los datos originales. Los datos serán eficientemente simplificados mientras se resuelven problemas tales como la multicolinealidad. Los componentes principales frecuentemente sirven como paso intermedio para otras investigaciones, tales como regresión múltiple, análisis de conglomerados o para la factorización de la matriz de covarianzas en el modelo de análisis de factores [18].

3.2.3 Principios.

Los componentes principales consisten en la transformación de los vectores de las variables originales mediante una rotación del eje de coordenadas, de manera que se obtiene un nuevo sistema de coordenadas con propiedades estadísticas inherentes. Los componentes principales son los vectores característicos de la matriz de covarianzas, por lo cual el ACP puede considerarse como poner en términos estadísticos el desarrollo usual de raíces y vectores característicos de matrices positivas semidefinidas [1].

Desde el punto de vista teórico, los componentes principales conforman un conjunto conveniente de coordenadas y la varianza de los componentes principales caracteriza sus propiedades estadísticas. En la práctica, el ACP se usa para encontrar las combinaciones lineales con mayor varianza. Esto es muy útil en los estudios exploratorios con una gran cantidad de variables. Uno de los aspectos relevantes en estos casos es la desviación, por lo que una forma de reducir la cantidad de variables consiste en descartar las combinaciones lineales que tienen varianzas pequeñas y estudiar exclusivamente aquellas con varianzas mayores [1].

Las varianzas de cada componente principal son los valores propios obtenidos de la ecuación característica de la matriz de varianzas y covarianzas, o bien, de la matriz de correlaciones. El proceso consiste en una rotación de ejes, de manera que cada nuevo eje va en la dirección de máxima variación y es ortogonal al eje consecutivo, mismo que explica la mayor variación no esclarecida por el primero. El proceso continúa hasta un máximo de "p" componentes principales, igual al número de variables en estudio [8] (figura 3.2.1).



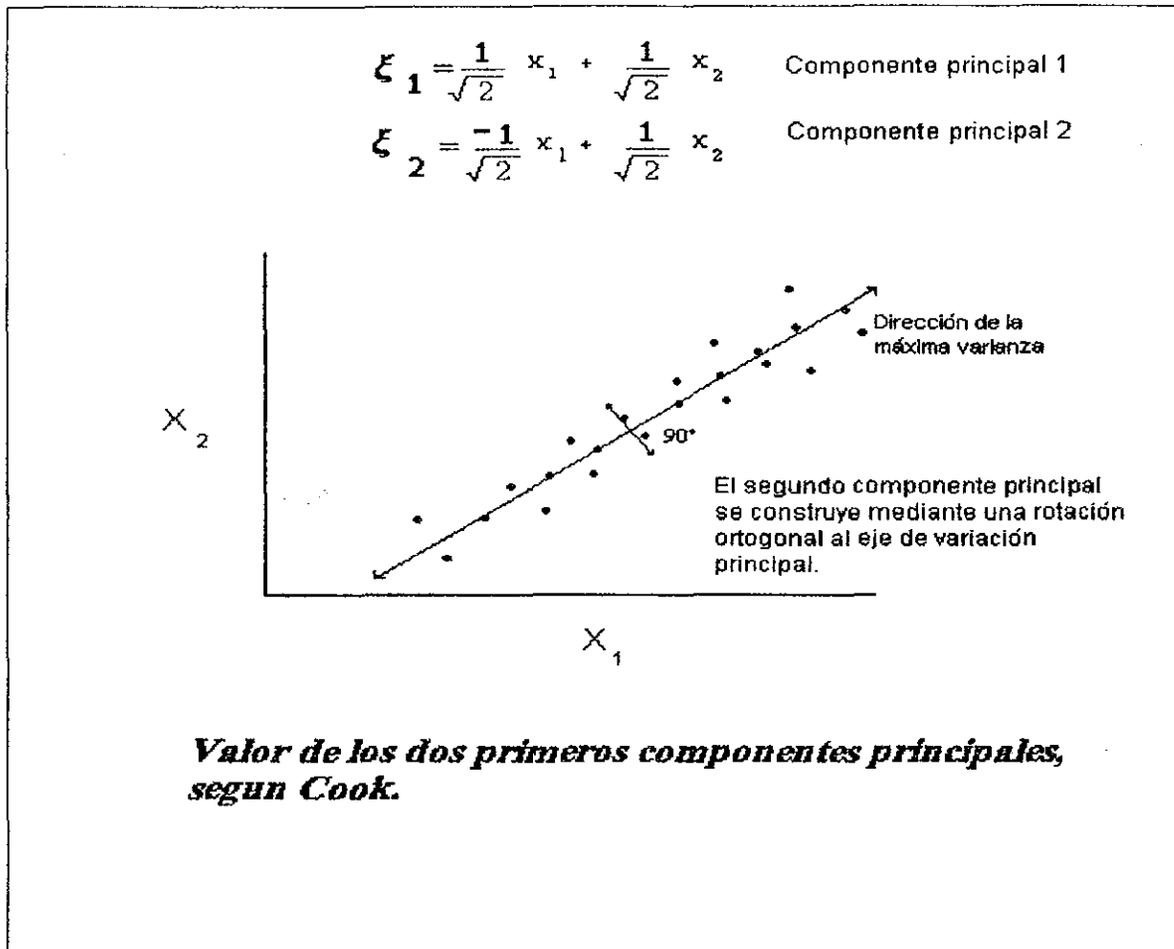


Figura 3.2-1 Primeros dos componentes principales

3.2.4 Proceso.

Al construir los componentes principales, se tiene siempre la esperanza de que las varianzas de la mayoría de ellos sean tan pequeñas que puedan despreciarse. En ese caso, la variación en el conjunto de datos puede ser adecuadamente descrita por los pocos componentes principales cuya varianza no puede despreciarse. El grado de economía del proceso es lograr que las p variables originales sean explicadas por un número más pequeño de componentes principales. No obstante, el análisis no siempre trabaja de este modo: si las variables originales no están correlacionadas entre sí, el método no consigue absolutamente nada [8]. Sharma [30] refuerza: el análisis de componentes principales es más apropiado si las variables están interrelacionadas, porque sólo entonces es posible reducir el número de variables sin pérdida significativa de información.

La sustitución de las variables originales por los componentes principales conduce a cierta pérdida de información, pero se pretende mantenerla a un nivel ínfimo. Los componentes principales tienen importancia teórica por su relación con las distribuciones elípticas y las distancias estandarizadas [14].

Los propósitos de un ACP, desde un punto de vista geométrico, consisten en la identificación de un nuevo conjunto de ejes ortogonales [30] de modo que:

- Las coordenadas de las observaciones con respecto a cada eje, den los valores a las nuevas variables. Los nuevos ejes son los **componentes principales** y los valores de las nuevas variables se llaman **valores, calificaciones o pesos** de los componentes principales.
- Cada nueva variable es una combinación lineal de las variables originales.
- La primera nueva variable representa la variación máxima del conjunto de datos.
- La segunda nueva variable representa la variación máxima del conjunto de datos que no fue considerada en la primera variable.
- La p-ésima nueva variable representa la variación máxima del conjunto de datos que no fue considerada en las primeras p-1 variables.
- Todas las p variables no están correlacionadas entre sí.

Una vez identificados los propósitos del ACP, es fácil indicar el proceso de cálculo correspondiente. Partamos de la definición de los componentes principales como una combinación lineal tan grande en varianza como sea posible [14]:

Debido a que los coeficientes α_{ij} podrían aportar arbitrariamente cualquier peso a la varianza de un componente principal, simplemente haciéndolos múltiplos de cualquier

cantidad mayor que uno, es necesario aplicar una restricción a los coeficientes: $\alpha_{i1}^2 + \alpha_{i2}^2 + \dots + \alpha_{ip}^2 = 1$. Esta restricción implica que consideraremos únicamente combinaciones lineales estandarizadas [14].

Los pesos (o cargas) de cada factor estimados por los coeficientes, se estiman de manera que $\alpha_{i1}\alpha_{j1} + \alpha_{i2}\alpha_{j2} + \dots + \alpha_{ip}\alpha_{jp} = 0$ para todo $i \neq j$. Esta última condición asegura que los nuevos ejes son ortogonales entre sí y por tanto ξ_1 no está correlacionado con ξ_2 [30].

Las varianzas de los componentes principales son los valores propios de la matriz de covarianzas. Hay p de tales valores propios y algunos de ellos pueden ser iguales a cero. No es posible obtener valores propios negativos de una matriz de covarianzas. Si asumimos que los valores propios están ordenados como $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ entonces λ_i corresponde al i -ésimo componente principal:

Ecuación 1 Valor del primer componente principal.

$$\xi_1 = \alpha_{i1} x_1 + \alpha_{i2} x_2 + \dots + \alpha_{ip} x_p$$

En particular, $Var(\xi_i) = \lambda_i$ y también $\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{ip}$, son los elementos del correspondiente vector propio, tales que $\alpha_{i1}^2 + \alpha_{i2}^2 + \dots + \alpha_{ip}^2 = 1$. Una propiedad importante es que la suma de los valores propios es igual a la traza de la matriz de covarianzas (o correlaciones) [21].

3.2.5 Correlación entre variables y componentes principales.

En la elección del uso de la matriz de covarianzas o correlaciones en el ACP, Morrison [23] señala que si las respuestas se han registrado en varias unidades diferentes, un arreglo lineal de las mediciones originales podría tener escaso significado y entonces es mejor emplear la matriz de correlaciones. Por el contrario, si las unidades de medición y las respuestas son razonablemente homogéneas, la matriz de covarianza resulta estadísticamente

atractiva, pues el i -ésimo componente principal extractado de ella será el arreglo lineal que contribuya a explicar la mayor parte de la varianza total y la maximización de tal medida mediante la estandarización será bastante artificial. La teoría de muestreo de componentes extraídos de matrices de correlación es más compleja que la procedente de matrices de covarianza.

La determinación de la cantidad de variación explicada por cada componente principal se determina dividiendo el valor propio correspondiente entre la traza de la matriz de covarianzas o correlaciones empleada en el análisis [21]. Si los componentes fueron extraídos de la matriz de correlaciones en lugar de la de covarianzas, la suma de las raíces características [23] será:

Ecuación 2 La traza de la matriz de correlaciones es igual al número de variables y a la suma de las raíces características de esta matriz.

$$Tr R = p$$

y la proporción de la “varianza” en el diagrama de dispersión atribuible al j -ésimo componente será λ_j/p . Si los primeros r componentes, extraídos explican una gran parte de la varianza de muestra, podemos determinar un nuevo valor para cada unidad de muestreo y emplearlo en análisis posteriores en lugar de las variables originales. Las calificaciones de componentes principales procedentes de la matriz de covarianzas para el i -ésimo sujeto son:

Ecuación 3 Calificaciones de componente principal del i -ésimo sujeto, cuando se extraen de la matriz de varianzas y covarianzas.

$$y_{i1} = a_1'(x_i - \bar{x}), \dots, y_{ir} = a_r'(x_i - \bar{x})$$

donde x_i es el vector de la i -ésima observación y \bar{x} es el vector de promedios de la muestra. Los valores pueden escribirse como la matriz $N \times r$.

Ecuación 4 Proceso de cálculo de las calificaciones de componente principal

$$N = \left(I - \frac{1}{n} E \right) X A$$

En donde X es la matriz de datos, E es la matriz $N \times N$ con unos en toda posición y A es la matriz cuyas columnas son los primeros r vectores característicos. Si α_i se extrajo de la matriz de correlaciones, los valores pueden calcularse a partir de las observaciones estandarizadas. Entonces, los calificaciones de un componente en el i -ésimo sujeto serán:

Ecuación 5 Calificaciones del i ésimo sujeto si las α_i se obtuvieron de la matriz de correlaciones.

$$y_{i1} = \alpha'_1 z_i, \dots, y_{ir} = \alpha'_r z_i$$

Para determinar la importancia relativa de cada variable y su contribución a la variación total, se determinan las correlaciones entre las calificaciones de variable y los componentes principales correspondientes. La manera de determinar esta correlación [18], es mediante las fórmulas siguientes, la primera si se ha empleado en el cálculo la matriz de varianzas y covarianzas, la segunda, si se utilizó la matriz de correlaciones:

Ecuación 6 Correlaciones entre las calificaciones de componente principal y las variables.

$$r_{x_i, z_k} = \alpha_{ki} \sqrt{\frac{\hat{\lambda}_i}{s_{kk}}} \quad r_{x_i, \xi_k} = \alpha_{ik} \sqrt{\hat{\lambda}_i}$$

Esta prueba de las correlaciones debe tomarse más en cuenta para la interpretación que las calificaciones de Componente Principal.

3.3 Análisis de factores.

3.3.1 Orígenes.

El análisis de factores AF fue propuesto por Charles Spearman en 1904 como una metodología para explicar el desempeño de estudiantes en varios cursos y entender la

conexión entre los grados escolares y la inteligencia de los sujetos. Spearman hipotetizó que el rendimiento de los estudiantes en varios cursos estaría interrelacionado y que esto último podría explicarse por el nivel de inteligencia general de los sujetos [30].

El análisis de factores AF es el nombre genérico dado a una clase de métodos estadísticos cuyo propósito primario es la reducción y resumen de datos. Las funciones del AF son: identificar en un gran conjunto de variables las dimensiones que son latentes (no fácilmente observables) lo cual se refiere como AFR, e idear un método para combinar o condensar un gran número de datos en grupos distintos dentro de una población grande, lo cual se conoce como AFQ; identificar variables apropiadas para subsiguiente regresión, correlación o análisis múltiple discriminante dentro de un conjunto de variables mayor; crear un conjunto enteramente nuevo con menor número de variables que el original para reemplazar parcial o totalmente a las variables originales en un análisis de regresión, correlación o análisis discriminante [17].

La idea de aplicar el análisis de componentes principales y de factores a un conjunto de datos, consiste en formar subconjuntos coherentes, relativamente independientes entre sí. Los factores agrupan variables fuertemente correlacionadas, permitiendo separarlas de otros grupos de variables, mismas que tendrán como característica una fuerte asociación. Los factores son representativos de los procesos que han ocurrido para crear las correlaciones entre las variables [36].

Los ACP y AF pueden ser exploratorios por su naturaleza. El AF puede ser utilizado como herramienta para reducir un gran conjunto de variables a otro más pequeño, pero significativo. Debido a que ambos análisis son sensibles a la magnitud de las correlaciones, deberían efectuarse comparaciones robustas para asegurar la calidad del proceso [36].

Los métodos del AF pueden lograr sus propósitos tanto desde la perspectiva de un análisis exploratorio como confirmatorio. El segundo caso es cuando el investigador quiere probar la hipótesis de que ciertas variables deberían ser fusionadas en un número preciso de factores, basándose en algún soporte teórico o en investigación anterior. En otras palabras, se mide el grado en el cual los datos cumplen con la estructura esperada por el analista [17].

3.3.2 Principios.

Para el estudio de estructuras de “dependencia” multinormal, podríamos preferir una técnica para explicar las covarianzas de las respuestas. Aunque el ACP efectúa esto en cierta medida a través de la factorización de la matriz de covarianzas, debemos recordar que consiste más bien en una transformación y no en un modelo fundamental de la estructura de la matriz de covarianzas [23].

Hay semejanzas y diferencias entre ACP y AF: ambas técnicas pretenden reducir las dimensiones de un conjunto de p variables, encontrando índices que ayuden a elucidar las interrelaciones de las variables. Sin embargo, el ACP no se basa en algún modelo estadístico particular, mientras el AF tiene como base un modelo lineal, de modo que [21]:

Ecuación 7 Modelo del análisis de factores.

$$X_i = \alpha_i F + \varepsilon_i$$

en donde X_i es el i -ésimo valor estandarizado con media 0 y varianza 1, α_i es una constante, F es un valor de “factor”, el cual tiene media 0 y varianza 1 considerando a todos los individuos y ε_i es la parte de X_i específica del i -ésimo valor. Además, la varianza de cada x puede considerarse como $\text{Var}(x_i) = \text{Var}(\alpha_i F + \varepsilon_i) = \alpha_i^2 + \text{Var}(\varepsilon_i)$, esto último porque F y ε son independientes y “ α ” una constante. La varianza de F es unitaria, pero ocurre lo mismo con la varianza de x_i , de modo que el modelo puede representarse como: $\alpha_i^2 + \text{Var}(\varepsilon_i) = 1$.

La diferencia principal entre el ACP y el AF consiste en que el AF hace distinción entre partes comunes y específicas de las variables. El ACP se orienta más a la varianza, en tanto que el AF, analiza el monto de cada variable que está relacionado con las demás, por lo que el análisis de factores se orienta hacia la covarianza. Los registros observados en el análisis de factores, deben ser ajustados en cierta medida para reflejar, precisamente, sus partes covariantes. Este monto de ajuste es una corrección de los elementos de la diagonal de la matriz de correlaciones R . Cuando la diagonal de la matriz de correlaciones se reemplaza con comunalidades, se tiene una **matriz de correlaciones reducida**. Esencialmente, la matriz de

correlaciones reducida es una estimación de la matriz de correlaciones de factor común, la cual será factorizada en Λ y Φ . Una comunalidad puede tomar valores entre 0 y 1. Un valor de 0 significa que la variable dada no tiene correlación con ninguna otra variable, mientras que una comunalidad de 1 significa que la varianza de la variable está perfectamente explicada por los factores comunes [10].

Aquí la constante α_i denominada carga del factor es tal que su cuadrado resulta proporcional a la varianza de x_i explicada por el factor correspondiente [21].

Spearman formuló la teoría de los dos factores en los *test* mentales: cada uno de éstos sería formado por dos partes, una común a todos (inteligencia general) y otra específica de cada uno de los *test*. Esto condujo al modelo general para el análisis de factores, en donde x_i es el valor del i -ésimo *test* con media 0 y varianza 1. $\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{im}$ son las calificaciones de factor para el i -ésimo *test*. F_1, F_2, \dots, F_m son m factores comunes no correlacionados con media 0 y varianza 1 ε_i es el factor específico para el i -ésimo *test*, el cual no está correlacionado con ninguno de los factores y tiene media cero.

Ecuación 8 Modelo general de análisis de factores

$$x_i = \alpha_{i1} F_1 + \alpha_{i2} F_2 + \dots + \alpha_{im} F_m + \varepsilon_i$$

Con este modelo,

Ecuación 9 Varianza de X_i como suma de comunalidades y especificidades.

$$Var(x_i) = 1 = \alpha_{i1}^2 Var(F_1) + \dots + \alpha_{im}^2 Var(F_m) + Var(\varepsilon_i) = \alpha_{i1}^2 + \dots + \alpha_{im}^2 + Var(\varepsilon_i)$$

en donde $\alpha_{i1}^2 + \alpha_{i2}^2 + \dots + \alpha_{im}^2$ son llamadas **comunalidades** de x_i (la parte de la varianza relacionada con los factores comunes), mientras $Var(\varepsilon_i)$ se llama **especificidad o singularidad** de x_i (la parte de la varianza no relacionada con los factores comunes) [21].

Además, la correlación entre x_i y x_j es $r_{ij} = \alpha_{i1} \alpha_{j1} + \alpha_{i2} \alpha_{j2} + \dots + \alpha_{im} \alpha_{jm}$. De esta última ecuación se deduce que dos valores correlacionados tendrán calificaciones altas en los mismos factores. También, $-1 \leq \alpha_{ij} \leq 1$ por lo que la comunalidad no excede a 1.

3.3.3 Supuestos y pruebas.

Antes de trabajar datos mediante el ACP y el AF, debe revisarse gráficamente la presencia de *outliers* (valores extremos), datos perdidos y correlaciones bajas entre las variables, debido al riesgo de un pobre ajuste de cada una de ellas a la distribución normal (distribuciones marginales). En ocasiones, es necesario transformar los datos. Debido a que los coeficientes de correlación son poco confiables cuando proceden de muestras pequeñas, se recomienda tener un mínimo de cinco casos para cada variable observada. La estimación de valores perdidos no es un asunto trivial, pues si esto se realiza mediante regresión, resultarán correlaciones entre variables artificialmente altas y en consecuencia, factores prefabricados [36].

En el ACP, la multicolinealidad no es un problema, pues no es necesario invertir ninguna matriz. No obstante, el caso es distinto en el AF. Si el determinante de R^{-1} y los valores propios asociados con algunos factores se aproximan a cero, la multicolinealidad o la singularidad podrían estar presentes. En ese caso, se requiere eliminar las variables que provocan el problema correspondiente [36].

La multicolinealidad es una alta correlación entre dos variables independientes². La multicolinealidad puede causar serias perturbaciones en el ajuste de mínimos cuadrados y volver inútil el modelo lineal, puesto que afecta seriamente la estimación de los coeficientes. Puede analizarse la matriz de correlaciones y obtener sus valores propios (*valor propio*). Si algunos de éstos son muy pequeños, indican que la multicolinealidad está presente. Algunos analistas han propuesto emplear la razón [22]:

¹ La matriz de correlaciones.

² No obstante, hay que tomar en cuenta que el AF es una técnica de interdependencia y la multicolinealidad simplemente indica que alguna variable está proporcionando información redundante.

Ecuación 10 Prueba de multicolinealidad mediante valores propios de la matriz de correlaciones.

$$k = \frac{\lambda_{m\acute{a}xima}}{\lambda_{m\acute{i}nima}}$$

En donde: λ máxima es el mayor valor propio obtenido de la descomposición de la matriz de varianzas y covarianzas o correlaciones, y λ mínima el menor de los valores propios de dicha matriz. Para la interpretación del valor de k como método de diagnóstico, la multicolinealidad será: si $k < 100$, leve; con valores de $100 < k < 1,000$, moderada a fuerte; si $k > 1,000$, severa;

3.3.4 Solución del modelo de análisis de factores.

Los pasos en el análisis de factores son tres: se comienza determinando valores provisionales para las calificaciones de los factores α_{ij} . Una forma de conseguirlo es empleando el ACP, rechazando todos los componentes después de los primeros m , los cuales son factores por sí mismos. Los factores encontrados no están correlacionados entre sí, lo cual significa que uno de los supuestos del análisis de factores no se ha cumplido; en cambio, se habrán conseguido comunalidades altas [21].

Una vez determinadas las calificaciones provisionales de los factores, se puede mostrar que estos no son únicos: si F_1, F_2, \dots, F_m son los factores transitorios, habrá un número ilimitado de soluciones para el modelo de AF, puesto que es posible formar combinaciones lineales no correlacionadas que “expliquen” los datos igualmente bien, de forma:

Ecuación 11 Rotación de factores.

$$F'_m = d_{m1}F_1 + d_{m2}F_2 + \dots + d_{mm}F_m$$

Este segundo estado en la solución del AF se llama **rotación de factores** y con ella se pretende obtener factores más sencillos de interpretar. Rotar es encontrar los valores de las d_{ij} en la ecuación 11.

La rotación de factores puede ser ortogonal u oblicua. Con la rotación ortogonal, los nuevos factores no están correlacionados y con la rotación oblicua sí. Cualquiera que sea el tipo de rotación empleado, es deseable que una carga en los nuevos factores sea muy cercana o muy diferente de cero. Una carga muy cercana a cero de α_{ij} significa que x_i no está fuertemente relacionada con el factor F_j , mientras que un gran valor (positivo o negativo) de α_{ij} indica que x_i está determinada por F_j en gran medida. Si los valores están fuertemente relacionados con algún factor, pero no con los otros, los factores resultarán fáciles de identificar e interpretar [21].

Una rotación ortogonal corresponde a una rotación rígida (reflexión) de los ejes de coordenadas. Debido a que las calificaciones originales no son fáciles de interpretar, es usual rotarlas hasta lograr una estructura más simple [18]. Kaiser sugirió la rotación varimax, la cual es la opción estándar en SPSS y SAS; ésta se basa en el supuesto de que para medir la capacidad de interpretar el factor “j” deben tomarse los cuadrados de sus calificaciones de factor, es decir, las de $\alpha_{ij}^2, \alpha_{2j}^2, \dots, \alpha_{mj}^2$. Mientras mayor sea esta varianza, los α_{ij}^2 tenderán a cero o uno. La rotación varimax maximiza la suma de esas varianzas para todos los factores, siendo entonces la rotación más popular [21]. La segunda rotación ortogonal en popularidad es la quartimax, cuyo objetivo es obtener un patrón de calificaciones tales que todas las variables tengan justamente una alta carga en un factor y casi cero en los factores restantes. Es una rotación adecuada cuando el investigador sospecha la presencia de un solo factor general. La rotación varimax destruye o suprime dicho factor general y por tanto no debe ser empleada cuando se sospeche esta condición [30]. Esta rotación es opcional en SPSS y SAS.

El último paso del AF implica el cálculo de los valores de factor, que son los valores de los factores F_1, F_2, \dots, F_m para cada uno de los individuos.

3.3.5 Evaluación de la solución.

La manera de decidir si los datos son apropiados para el análisis de factores adopta medidas, la mayoría de las cuales son “reglas de pulgar” y tienen naturaleza heurística. Primero, se sugiere ver la matriz de correlaciones R , pues alta correlación entre las variables será indicio de grupos homogéneos de éstas, que miden las mismas construcciones o

dimensiones, mientras que las variables pobremente correlacionadas serán indicio de conglomerados heterogéneos. Se puede ver al AF como una técnica que trata de agrupar variables que forman factores comunes. Después será posible examinar las correlaciones parciales controladas por todas las variables, las cuales aparecen en una matriz especial llamada de correlaciones negativas anti-imagen, misma que debe consistir en valores pequeños para ser apropiada para factorizarse. La matriz de correlaciones anti-imagen es la matriz del negativo de los coeficientes de correlación parcial. Si las variables comparten factores comunes, los coeficientes de correlación parcial deben ser pequeños, debido a que se han eliminado los efectos lineales de otras variables. Si hay una gran porción de coeficientes altos en esta matriz, deberá reconsiderarse el uso de un modelo de análisis de factores para las variables. Sin embargo, decidir cuando un valor es “pequeño” es cuestión de juicio. Finalmente, debe examinarse la medida de adecuación de muestreo de Kaiser, o KMO (Kaiser-Meyer-Olkin), la cual es una medida de la homogeneidad de variables. Kaiser sugiere las siguientes guías para los valores de la prueba [30]:

Tabla 1 Valores en la prueba KMO y su interpretación.

Valor KMO	Interpretación
> 0.9	Maravilloso
> 0.8	Meritorio
> 0.7	Regular
> 0.6	Mediocre
> 0.5	Miserable
< 0.5	Inaceptable

En la prueba KMO se sugiere emplear los mayores a 0.8, aunque 0.6 parece tolerable. En ocasiones, el valor de la prueba KMO puede aumentar si eliminamos las variables inadecuadas para el análisis. Un valor matricial de 0.813 o más para la prueba indica que la matriz es apropiada para la factorización [30].

El siguiente paso consiste en decidir el número de factores necesarios para explicar las correlaciones entre las variables. El enfoque resulta similar al de la selección del número de componentes principales que deben extraerse en ACP. Las dos técnicas heurísticas más

populares son la del valor propio mayor que 1 y el gráfico *scree*³. La regla del valor propio >1 se basa en que los valores menores a 1 contribuyen a explicar la varianza menos que una sola de las variables originales. Esta es la opción inicial en programas como SPSS y SAS. El gráfico *scree* muestra en uno de los ejes de coordenadas los valores propios y en otro el número de componentes. El gráfico debe ser examinado en la búsqueda de un “codo” o punto de inflexión notable, mismo que indicará el número de factores a extraer. Hay bastante subjetividad en la identificación del “codo”. También se ha propuesto un procedimiento paralelo por Allen y Hubbard en 1986, donde el número de valores propios a extraer se estima por la ecuación [30]:

Ecuación 12 Procedimiento paralelo de Allen y Hubbard.

$$\ln \lambda_k = a_k + b_k \ln(n-1) + c_k \ln\{(p-k-1)(p-k-2)/2\} + d_k \ln(\lambda_{k-1})$$

donde λ_k es la estimación del k-ésimo valor propio, p es el número de variables, n el número de observaciones, a_k , b_k , c_k y d_k son coeficientes de regresión (estimados y tabulados por Allen y Hubbard mediante simulación) y λ_0 se supone tener valor de 1. La tabla 2 muestra los valores de los coeficientes para la estimación de las primeras k = 12 raíces, por ser éste el número máximo de valores propios que pueden extraerse de la matriz de correlaciones con la cual trabajaremos.

Tabla 2 Coeficientes de regresión para determinar el número de factores a extraer mediante el proceso paralelo.

Raíz k	a	b	c	D	R ²
1	0.9794	-0.2059	0.1226	0.0000	0.931
2	-0.3781	0.0461	0.0040	1.0578	0.998
3	-0.3306	0.0424	0.0003	1.0805	0.998
4	-0.2795	0.0364	-0.0003	1.0714	0.998
5	-0.2670	0.0360	-0.0024	1.0899	0.998
6	-0.2632	0.0373	-0.0040	1.1039	0.998
7	-0.2580	0.0329	-0.0039	1.1173	0.998
8	-0.2544	0.0310	-0.0064	1.1421	0.998
9	-0.2111	0.0288	-0.0079	1.1229	0.998
10	-0.1964	0.0276	-0.0083	1.1320	0.998
11	-0.1858	0.0266	-0.0073	1.1284	0.998
12	-0.1701	0.0229	-0.0090	1.1534	0.998

³ Scree = Montón de guijarros en la ladera de una montaña. La gráfica recibe este nombre por su apariencia que tiene y fue propuesta por Cattell en 1966.

Una prueba adicional para evaluar la solución del AF es la de esfericidad de Bartlett. Este proceso determina cuando una matriz es ortogonal. Dicha matriz tendrá un determinante unitario, indicando que las variables no están correlacionadas. Si un par de variables están perfectamente asociadas, el determinante será cero. Un valor altamente significativo en el *test* de Bartlett indica que la matriz no es ortogonal y por tanto es apropiada para factorizarse; la prueba, sin embargo, es sensible al tamaño de muestra y muestras grandes tienden a parecer no ortogonales aún con valores de correlación pequeños [30].

Además, es necesario evaluar la bondad de la solución de factores. Para esto se determina la raíz cuadrada de la matriz de correlaciones reproducida RMSR, misma que debe ser pequeña. Podemos comparar diferentes soluciones de factores mediante RMSR y quedarnos con la que tenga un valor menor como la posible mejor solución. También podemos aplicar esta técnica a cada una de las variables, determinando así cuales son bien explicadas por la solución de factores encontrada. La fórmula es:

Ecuación 13 Fórmula de residuales de la matriz de correlaciones reproducida.

$$RMSR = \sqrt{\frac{\sum_{i=1}^p \sum_{j=1}^p residual_{ij}^2}{p(p-1)/2}}$$

en donde el $residual_{ij}^2$ es el residuo de la correlación entre las variables j e i y p es el número de variables [30].

3.3.6 Interpretación.

Para interpretar los factores debemos decidir cuáles cargas de los mismos tienen valor considerable. Una regla de pulgar para un análisis preliminar sugiere considerar las cargas superiores a 0.3 (en valor absoluto) como significativas, las mayores a 0.4 como importantes y las más altas que 0.5 como muy significativas. Esta guía es útil para tamaños de muestra mayores a 50 [17].

3.4 Análisis de conglomerados.

3.4.1 Orígenes.

Según se ha visto en los antecedentes, el análisis de conglomerados AC⁴ es una de las técnicas más ampliamente utilizadas en la estadística multivariable.

La necesidad de clasificar las especies tal y como se entiende en la actualidad fue iniciada por Linneo en el siglo XVIII. Sin embargo se considera que la obra que más ha influido en el enfoque numérico de la taxonomía es la publicada por Sokal y Sneath en 1963, en la cual se exponen las bases, procedimientos y reglas de la clasificación [9].

3.4.2 Objetivos de la taxonomía numérica.

A grandes rasgos, la taxonomía numérica intenta construir clasificaciones “naturales” basadas en la semejanza fenotípica de individuos o clases, la cual se valora partiendo de una adecuada elección de un coeficiente de similaridad. Cuando se utilizan variables cuantitativas se trabaja con una matriz de correlaciones (producto-momento de Pearson), o de distancias (euclidiana, de Mahalanobis, etc.) [9].

El AC puede aceptar una gran variedad de datos. Estos son llamados comúnmente medidas de “similaridad”, pero también podrían denominarse “proximidad”, “parecido” o “asociación”. Algunos autores recomiendan emplear datos estandarizados, mientras uno podría desear agrupar datos con diferentes escalas [15].

La mayor utilidad del AC consiste en detectar grupos naturales en los datos. La clasificación se basa en la ubicación de objetos en categorías más o menos homogéneas, de manera tal que se revelen las relaciones entre los grupos. El AC carece de un cuerpo teórico-estadístico y es de naturaleza heurística, requiriendo que el usuario decida cómo determinar

⁴ *Cluster analysis* en inglés.

los conglomerados, pero tales decisiones tienen gran influencia en los resultados de la clasificación. A pesar de ello, la agrupación será más objetiva que subjetiva [36].

3.4.3 Principios.

La fundamentación matemática de la taxonomía numérica son los principios de jerarquía indexada y geometría ultramétrica. Si $\Omega = \{1, 2, \dots, n\}$ es un conjunto finito, se dice que $H \subset P(\Omega)$ es una jerarquía de partes de omega si [9]:

- Dados dos elementos de H , o son disjuntos o uno de ellos está contenido en el otro.
- H es el resultado de la reunión de los elementos h , o bien carece de elementos.
- Omega es una jerarquía total si contiene a H y a todas las partes formadas por un solo elemento.
- Los H se llaman clases o conglomerados si, $\Omega = \{h_1 + h_2 + \dots + h_n\}$ de donde $\{h_1, h_2, \dots, h_n\}$ es una partición o conglomerado.
- Un índice de la jerarquía H es una aplicación d que hace corresponder a cada clase h un número real no negativo $d(h)$, tal que: 1) $d(\{i\}) = 0 \forall i \in \Omega$ y 2) $h \subset h' \Rightarrow d(h) < d(h')$; se dice entonces que H es una jerarquía indexada.

3.4.4 Métodos en el análisis de conglomerados.

Los métodos se clasifican en: aglomerativos y divisivos, según partan de objetos que se fusionan progresivamente o del conjunto total, que se divide paulatinamente; jerárquicos y no jerárquicos, si los grupos se forman mientras decrece la homogeneidad entre los mismos o se forman grupos homogéneos sin establecer relación entre los mismos; monotéticos y politéticos, si se basan en una sola característica muy relevante, o en muchas, buscando analogías entre miembros de una misma clase [9].

Los métodos de agrupación jerárquica pueden ir de arriba hacia abajo y emplear la división lógica, o bien de abajo hacia arriba y conformar agregación. Este último proceso se basa en la agrupación de casos a través de similitudes y es el método más común [36].

Los métodos para formar los conglomerados por el principio de jerarquía son: el criterio de varianza mínima (MVAR), la varianza mínima jerárquica (MNVR), la ligazón simple (SLNK), la ligazón completa (CLNK), la ligazón promedio (ALNK), la ligazón promedio ponderado (WLNK), la ligazón mediana (MEDN) y el método de centroide (CNTR) [2, 14].

El método de varianza mínima (MVAR) construye grupos con mínima varianza interna (compactos) y máxima varianza externa (aislados). Es útil para agrupación de sinopsis y para todos los trabajos donde otros métodos carezcan de justificación explícita. El método se conoce también como agrupación de Ward.

Con excepción del método de varianza mínima jerárquica (MNVR), todos los métodos de conglomerados construyen distancias euclidianas para todos los “n” vectores, requiriendo una gran capacidad interna de memoria. La opción MNVR funciona con principios idénticos a la MVAR, pero sin necesitar almacenar distancias euclidianas, ahorrando memoria y tiempo de proceso.

Cón SLNK frecuentemente se logra una jerarquía muy sesgada o “eslabonada”. Por eso no es muy útil para resumir datos, pero podría indicar objetos muy anómalos o extremos (*outliers*), los cuales serán los últimos en incorporarse en la jerarquía. Se basa en la distancia mínima. Se localizan los dos individuos separados por la distancia más corta y se colocan juntos en el mismo grupo; después, se encuentra la próxima distancia menor y un tercer individuo se une al grupo, o bien, se forma un nuevo grupo de dos individuos, siguiendo con el proceso hasta que todos los individuos formen un solo conjunto. Este proceso también se denomina el enfoque “vecino más cercano”⁵. La distancia entre dos grupos es la más pequeña

⁵Nearest-neighbor approach, como se designa en el software.

entre cualquier punto de un conjunto, a cualquier punto del otro. Dos grupos son fusionados en cualquier etapa mediante la ligazón más corta o más fuerte entre ellos.

Al emplear CLNK, a menudo se consiguen grupos que no difieren impropriamente de los construidos mediante MVAR, pero sus criterios restrictivos no son útiles si los datos son "ruidosos". El procedimiento es parecido al de SLNK, con la diferencia de trabajar con base en distancias máximas. Por ello se ha denominado el enfoque "vecino más lejano". Es un método en donde la distancia máxima entre dos individuos dentro de un grupo representa la más pequeña hiperesfera capaz de encerrar al conjunto. Todos los objetos dentro de un grupo están ligados uno al otro en alguna distancia máxima (similaridad mínima).

Si usamos ALNK emplearemos un criterio razonablemente equilibrado entre SLNK (laxo) y CLNK (rígido): todos estos métodos son interesantes si queremos una representación gráfica de los resultados de la agrupación. El criterio de agrupación es la distancia⁶ promedio de los individuos dentro de un grupo respecto de los individuos en otro. La partición se basa en todos los elementos del grupo, más que en algún par de miembros extremos. ALNK tiende a formar grupos con varianzas más pequeñas y a producir conjuntos sesgados, con aproximadamente la misma varianza.

El método WLNK no toma en cuenta los tamaños relativos de los conglomerados para aglomerarlos. Este y los dos siguientes métodos no se recomiendan para uso general.

Con MEDN se reemplaza un grupo con el valor de la mediana. No se puede garantizar que el criterio de esos valores será uniforme y eso podría presentar dificultades con la interpretación del dendrograma.

El método CNTR reemplaza un conglomerado con el valor del centroide. Al igual que en el caso anterior, es posible invertir la jerarquía. La distancia entre dos conglomerados es la

⁶ Típicamente, la distancia euclidiana, aunque podría ser cualquiera otra entre las ya mencionadas.

misma que hay entre sus centroides (promedios). En este método, cada vez que los individuos se agrupan se calcula un nuevo centroide. Este método es muy popular en biología, pero podría producir confusión o desorden en los resultados. La ventaja de este método consiste en ser más robusto contra los *outliers*. Requiere datos métricos para funcionar adecuadamente.

La capacidad del AC para reproducir bien la estructura jerárquica de grupos previamente conocidos depende de las diferencias entre éstos; los grupos muy contrastantes que no presenten traslapes tienen mayores posibilidades de ser bien reproducidos mediante un análisis de conglomerados. El análisis puede dificultarse también por la forma de los conglomerados y bajo algunos patrones de agrupación de los datos, todos los algoritmos fallarán en la identificación de los grupos. Por supuesto, los grupos se basarán en las variables medidas y por tanto el estudio resulta sensible a la elección de las mismas. Una elección de variables que aparenten ser igualmente razonables puede originar conglomerados diferentes [21].

3.4.5 Proceso.

El primer paso en el AC consiste en seleccionar una medida de similaridad, después hay que decidir cuál técnica de agrupación emplear (jerárquica o no jerárquica), en tercer lugar, el método dentro de la técnica (p. ej. técnica jerárquica y método de centroide), seguido del número de conglomerados que deben construirse, para culminar con una interpretación de los grupos. Las medidas de similaridad pueden ser las distancias euclidianas cuadradas, la matriz de distancias de Mahalanobis o cualquier otra medida que se considere conveniente [30].

Los procedimientos jerárquicos comienzan con cada objeto fuera de su propio conglomerado. En pasos subsecuentes, los dos grupos (o individuos) más cercanos se combinan en un nuevo agregado reduciendo la cantidad de conjuntos en una unidad en cada paso. Eventualmente, todos los individuos formarán un solo grupo. El procedimiento puede representarse gráficamente mediante un dendrograma o gráfico de árbol [17] (figura 3.4-1).

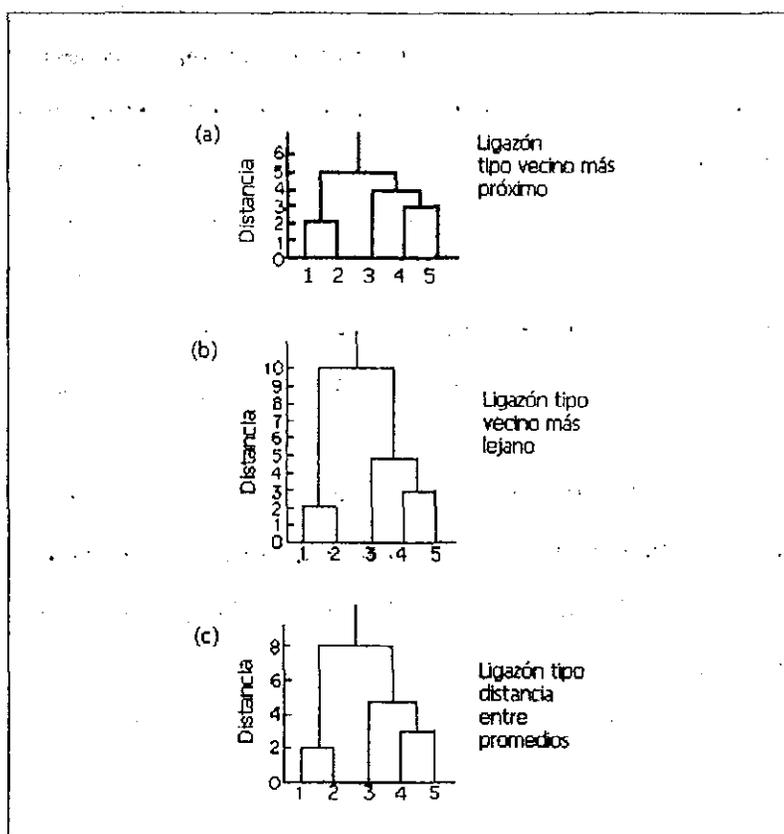


Figura 3.4-1 Tipos de dendrograma según Manly [21].

3.4.6 Evaluación de la solución.

En cualquier situación práctica en que se utilice el AC, debemos responder algunas preguntas respecto de una solución dada: ¿cómo difieren los conglomerados?, ¿Cuál es el número óptimo de conglomerados a formar?, ¿Qué tan bueno es el ajuste de la solución para un nivel previamente especificado de conglomerados? La primera pregunta es concerniente a las diferencias entre los perfiles de cada aglomerado; la segunda, implica un trueque entre la parsimonia, con la acepción de pocos aglomerados y una medida del incremento de la homogeneidad dentro de un grupo, considerando el aumento de la cantidad de grupos en la solución; la tercera, considera la recuperabilidad del grupo, la cual puede verse como pérdida de ajuste, en el sentido de que entre los datos originales y la solución resultante de conglomerados debiera darse una alta correlación de rangos u orden [10].

Hay muchos criterios y guías para enfocar el problema del número de conglomerados a ser formados. Infortunadamente, no existe un proceso de selección estandarizado para la determinación objetiva de tal número. Las distancias entre los conglomerados en pasos sucesivos podrían servir como guía útil y el analista podría elegir detenerse cuando ésta exceda un valor especificado o cuando las distancias sucesivas entre conglomerados en diferentes pasos hagan un salto inesperado. Tales medidas son referidas a veces como apreciaciones de variabilidad del error. Además, algunas relaciones conceptuales intuitivas o teóricas podrían sugerir un número natural de conglomerados. En el análisis final, sin embargo, es probablemente mejor calcular las varias soluciones por diferente número de conglomerados (por ejemplo, dos, tres, cuatro) y entonces decidir entre las soluciones alternativas basados en un criterio *a priori*, un juicio práctico, el sentido común o fundamentos teóricos. También, se podría comenzar este proceso determinando si el problema de los grupos sería más manejable con una cantidad de conglomerados, digamos entre tres y seis, luego resolver para este número de conglomerados y seleccionar la mejor opción después de evaluarlos a todos ellos [17].

Entre los estadísticos más ampliamente utilizados para evaluar la solución del AC están: la raíz del cuadrado medio del nuevo conglomerado (RMSSTD), la r^2 semiparcial, la r^2 y la distancia entre dos conglomerados. Dichos estadísticos proveen información acerca de la solución de conglomerados en cualquier paso dado, en relación con el nuevo conglomerado formado y las consecuencias de ello [30].

La raíz del cuadrado medio del nuevo conglomerado debe ser tan pequeña como sea posible; la r^2 es la razón entre la suma de cuadrados entre grupos y la suma de cuadrados totales. Podríamos decir que es una medida de la homogeneidad de los grupos, donde 0 implica que no hay diferencia entre los grupos y 1 indica la diferencia máxima.

3.5 Análisis múltiple discriminante.

3.5.1 Orígenes.

El AC y el AMD son técnicas multivariadas relacionadas con la separación de objetos en diferentes grupos y con la localización de nuevos objetos (u observaciones) en grupos previamente definidos. La discriminación es exploratoria por naturaleza y requiere resolver menos problemas estructurales que la clasificación. Los objetivos primordiales de la discriminación y la clasificación son, respectivamente [18]:

- Describir gráfica o algebraicamente las características diferenciales de objetos (observaciones) de varias colecciones (poblaciones) conocidas. Tratamos de encontrar “discriminantes” cuyos valores numéricos sean tales que las colecciones sean separadas tanto como sea posible.
- Clasificar objetos (observaciones) en dos o más grupos marcados. Se enfatiza en la derivación de una regla que pueda emplearse para asignar óptimamente un nuevo objeto a dichos grupos.

Se empleó el término discriminación tal como lo planteó R. A. Fisher en el primer tratado moderno de problemas sobre separación, publicado en 1936 [18]. En la práctica una función de clasificación podría sugerir un proceso discriminatorio y viceversa, con lo cual puede empañarse la diferencia entre separar y clasificar.

3.5.2 Supuestos y pruebas relacionadas.

El AMD asume que los datos proceden de una distribución normal multivariable y las matrices de covarianza son iguales. La violación del primero de los supuestos invalida el análisis, aunque una ligera desviación de la normalidad multivariable puede ser aceptable. Esto último, sin embargo, implica que la clasificación de algunos grupos puede sobreestimarse, mientras en otros se subestima. Si hay razones para creer que se ha violado la

distribución normal multivariable, debemos optar por la regresión logística como proceso alternativo [21].

Por otra parte, la violación del supuesto de igualdad de las matrices de covarianza afecta las pruebas de significación y los resultados de la clasificación. El grado en que esto es afectado depende de las variables discriminatorias y del tamaño de muestra de cada grupo; la hipótesis nula de igualdad de vectores de medias es rechazada más a menudo cuando el número de variables discriminatorias es grande o los tamaños de muestra de los grupos son diferentes, por lo cual el nivel de significación estaría inflado. Sin embargo, el AMD es bastante robusto a las violaciones de ambos supuestos [17]. La prueba M de Box es el estadístico apropiado para determinar el último de los supuestos, pero es muy sensible a las diferencias de tamaño entre muestras. Si se rechaza la hipótesis de igualdad de matrices de covarianza, debe optarse por la técnica alterna de función cuadrática discriminante [17]. En el paquete SPSS, la opción de emplear la matriz de varianzas y covarianzas de cada grupo en lugar de la matriz de covarianzas común a todos los grupos fue especialmente diseñada para emplearse cuando se viola el supuesto de igualdad de las matrices de varianza y covarianza entre grupos.

El AMD pretende estimar las interrelaciones entre una variable categórica y un conjunto de variables métricas en la forma general [17]:

Ecuación 14 Ecuación básica del análisis múltiple discriminante.

$$Y_i = x_1 + x_2 + \dots + x_n$$

donde Y_i es categórica. El AMD exige derivar la combinación lineal de dos o más variables independientes como apoyo para distinguir elementos entre grupos definidos a priori. Esto se logra mediante la aplicación de la regla estadística de maximizar la varianza entre grupos en relación con la varianza dentro de grupos. La función discriminante es una ecuación lineal de forma [17]:

Ecuación 15 Valor múltiple discriminante.

$$Z = w_1x_1 + w_2x_2 + \dots + w_nx_n$$

donde: Z es el valor discriminante, y w_i el peso de la i-ésima variable independiente x_i .

Promediando los valores discriminantes de todos los individuos dentro de un grupo particular, obtenemos el promedio de grupo denominado **centroide**, mismo que indica la localización más típica de un individuo dentro de un grupo particular, mientras la comparación entre centroides muestra cuán diferentes son los grupos a lo largo de las dimensiones que están siendo probadas. La prueba de significación estadística de la función discriminante es una medida generalizada de la distancia entre centroides de grupo [17].

3.5.3 Procedimientos de selección de variables.

Es necesario distinguir cuáles son las variables que discriminan mejor a los grupos y esa debe ser la primera meta del AMD. A esas variables se les denomina **variables discriminantes**. La segunda meta del AMD es identificar nuevos ejes tales que las nuevas variables Z, dadas por las observaciones proyectadas (valores discriminantes) hacia este nuevo eje provean la máxima separación o discriminación entre los grupos. El nuevo eje identificado se denomina **función discriminante**, mientras la proyección de un punto hacia tal función se designa **valor discriminante**. La tercera meta de AMD es la clasificación de futuras observaciones en alguno de los grupos existentes [30].

Entre los criterios comunes de selección estadística empleados para definir la adición o supresión de una variable, están la lambda de Wilks, la V de Rao, la distancia cuadrada de Mahalanobis y la razón F entre grupos [30].

A la razón entre las sumas de cuadrados dentro de grupos y la suma de cuadrados totales se le conoce como Λ (lambda) de Wilks. El proceso de cálculo es iterativo y en cada paso se incluye la única variable con menor valor Λ después que el efecto de las variables incluidas ya en la función discriminante se ha removido total o parcialmente. El procedimiento

puede ser aproximado por una prueba de F. Debido a que la Λ de Wilks implica la minimización de sumas de cuadrados dentro de grupos y la maximización de la suma de cuadrados entre grupos; considera que entre grupos hay separación y dentro de grupos homogeneidad. Este criterio es el más ampliamente utilizado en la práctica [30].

Ecuación 16 Lambda de Wilks.

$$\Lambda = \frac{SS_w}{SS_t} = \frac{SS_w}{SS_b + SS_w}$$

La V de Rao se basa en las distancias de Mahalanobis y se concentra en separar los grupos a través de la diferencia entre el centroide de cada grupo y el centroide de la muestra completa. La V de Rao y el cambio que ocurre al adicionar o segregarse una variable de la función discriminante pueden ser aproximados como un estadístico que sigue dicha distribución. Sin embargo, concentrado en la separación entre grupos, Rao no toma en consideración la homogeneidad de los mismos, de modo que la función discriminante podría no tener una máxima homogeneidad dentro de grupos [30].

Tanto los criterios de Rao como de Wilks maximizan la separación total entre los grupos, pero en el caso de más de dos grupos, podría darse una separación no óptima entre un par de éstos. La D^2 de Mahalanobis asegura que habrá una separación entre todos los pares de grupos. En cada paso el proceso agrega la variable con máximo aumento en la distancia de separación entre los pares de grupos más cercanos o remueve la de mínimo decremento [30].

La D^2 de Mahalanobis se determina considerando un modelo de muestreo de datos e individuos: Se obtienen N vectores de observaciones independientes x_1, \dots, x_n de un vector p-dimensional aleatorio multinormal con media vectorial μ y matriz de covarianzas Σ . Estos parámetros son estimados por un vector de medias μ y una matriz de covarianzas S. Si decimos que x_0 es uno de los vectores específicamente observados mediante algún criterio *a priori*, o bien, seleccionados aleatoriamente, podemos expresar la distancia de x_0 a x por la medida [23]:

Ecuación 17 Distancias cuadradas de Mahalanobis.

$$D^2 = (x_0 - \bar{x})' S^{-1} (x_0 - \bar{x})$$

Esta es la distancia más correcta entre puntos en el hiperespacio, representados por vectores contruidos con variables correlacionadas entre sí. Entre sus propiedades están: es invariante a cambios de escala; siempre es mayor o igual que cero, siendo éste el caso de dos puntos $x_i = x_j$; cumple la desigualdad del triángulo, está expresada en unidades de desviación estándar y tiene en cuenta la redundancia entre variables. Si la distribución de las variables es $N(\mu, \Sigma)$, podemos considerar a D^2 como variable aleatoria que sigue la distribución χ^2 [9].

Finalmente, el criterio de razón de F entre grupos toma en cuenta lo que la D^2 de Mahalanobis ha pasado por alto, provocando quizá que no todos los pares de grupos tengan una separación óptima: el tamaño de cada grupo. En este criterio, los grupos con tamaño de muestra más grandes reciben mayor peso que los grupos pequeños y la razón de F mide la separación entre cualquier par de grupos dado [30]:

Ecuación 18 Criterio de razón F entre grupos.

$$F = \left(\frac{1 - \Lambda}{\Lambda} \right) \left(\frac{n_1 + n_2 - p - 1}{p} \right)$$

Este procedimiento puede seguirse para asegurar la significación de la lambda de Wilks, donde p es el número de variables, n_1 y n_2 el tamaño de la muestra en el grupo 1 y 2 respectivamente. Con más de 2 variables, la F será sólo aproximada. La H_0 es que esta medida sigue una distribución de F con p grados de libertad para el numerador y $n_1 + n_2 - p - 1$ en el denominador.

3.5.4 Métodos para el análisis múltiple discriminante.

El AMD puede ser desarrollado por tres métodos: *forward*, *backward* y *stepwise*. En el método *forward* se inicia sin ninguna variable en la función discriminante; la variable que entra primero a la función es la que provee mayor discriminación entre los grupos, medida por

un criterio estadístico dado. Le sigue la variable que, excluyendo la primera, provee el mayor monto adicional de poder discriminatorio a la función de acuerdo con el mismo criterio estadístico. En el método *backward* se inicia con todas las variables en la función discriminante y en cada paso se remueve la variable con la menor cantidad de poder discriminatorio de acuerdo con el criterio estadístico. En el proceso *stepwise* hay un equilibrio entre los métodos *forward* y *backward*. En cada paso se agrega una variable con poder discriminatorio o se remueve una variable que no lo tenga. El proceso se detiene cuando ninguna variable se agrega o remueve de la función. Si las variables no están correlacionadas entre sí, los tres métodos encuentran la misma función. Si la multicolinealidad está presente los resultados se pueden ver afectados [30].

3.5.5 Número de funciones discriminantes a seleccionar.

Para seleccionar k funciones discriminantes debemos retener sólo aquellas que representen adecuadamente las diferencias entre los grupos. La decisión del número de funciones significativas puede tomarse empleando el estadístico χ^2 en un proceso iterativo hasta que el valor de la r -ésima función discriminante no sea significativo [30]:

Ecuación 19 Criterio de χ^2 para selección de funciones discriminantes.

$$\chi^2 = [n - 1 - (p + G) / 2] \sum_{k=r}^k \ln(1 + \lambda_k)$$

donde λ_k es el valor propio de la k -ésima función discriminante, G el número de grupos, n el tamaño de la muestra, r la r -ésima función que se está analizando y p el número de variables que conforman las funciones. Los grados de libertad son $(p+r+1)(G-r)$. En el primer cálculo, intervienen todas las funciones discriminantes y si χ^2 es significativa se supone que al menos la primera de ellas tiene poder de discriminar los grupos. La prueba se repite excluyendo el valor propio de la primera función y así sucesivamente.

3.6 Análisis multivariable de varianza.

3.6.1 Orígenes.

En el año de 1932, Wilks formuló la teoría del análisis multivariable de varianza (MANOVA) como una generalización del análisis de varianza de una variable. Se define el MANOVA como una técnica de dependencia que mide simultáneamente las diferencias entre dos o más variables métricas, tomando como base un conjunto de predictores categóricos [17].

El análisis multivariable de varianza es una generalización de la teoría de mínimos cuadrados (p. ej. el análisis de regresión) y el análisis de varianza al caso de variables vectoriales. El álgebra para el caso multivariable esencialmente coincide con la del método de una variable. Esto significa que la distribución teórica es la misma en ambos casos y el criterio de prueba es semejante al estadístico F. Sin embargo, el caso multivariable es más amplio en la elección de las pruebas de significación [1].

3.6.2 Utilización del MANOVA.

El análisis multivariable de varianza MANOVA trata con la separación de grupos sobre la base de variables dependientes interrelacionadas. Por esta razón, es mejor considerar el MANOVA como una forma de análisis múltiple discriminante [33]. No hay razón para emplear el análisis multivariable de varianza cuando las variables dependientes no están correlacionadas [24]. En contraposición, una serie de pruebas de F (ANOVA) trata con cada variable aisladamente, aunque algunas de ellas puedan estar correlacionadas y brindar información redundante [33].

Los analistas de datos han procedido con el MANOVA sobre el supuesto de que un efecto significativo multivariable protege contra errores tipo I. En pruebas secuenciales. Hubert y Morris citados por Smith [33], indican que las evidencias con modelos de simulación no soportan este argumento. Además, los mismos autores puntualizan que es posible la ocurrencia de una separación significativa en el espacio multidimensional sin la presencia de una separación unidimensional. Esto es, hay un efecto multivariable, pero ninguna diferencia

univariable. Adicionalmente, ellos señalan que el MANOVA se basa en el supuesto de que las variables dependientes forman un conjunto o conjuntos de variables sistemáticamente relacionadas y por tanto, sólo se recomienda realizar un MANOVA cuando este supuesto sea razonable. Deben tenerse en mente las preguntas que tratan de responderse con los datos, sin únicamente pretender protegerse contra errores de tipo I. Ellos recomiendan el empleo de varias pruebas univariadas cuando:

- Las variables dependientes no están interrelacionadas.
- Se propone determinar cuáles efectos generales (o manipulación experimental) han tenido efecto. Esto puede visualizarse como una fase exploratoria en el desarrollo de un programa de investigación.
- Las variables dependientes han sido previamente estudiadas a través del uso de pruebas F.
- El investigador está interesado en probar la capacidad de comparar grupos (como en una evaluación de la investigación).

Por otra parte, el MANOVA estaría plenamente justificado si tenemos en cuenta estas tres preguntas genuinamente multivariadas:

- ¿Puede un subconjunto de las variables dependientes separar los grupos con un grado de eficiencia aceptablemente cercano al del conjunto completo? Esto podría denominarse el problema de selección de variable y se relaciona con el AMD.
- Dado un subconjunto de variables dependientes que separa los grupos, ¿cuáles son sus contribuciones relativas?
- ¿Qué construcciones lógicas subyacentes afectan a los grupos?

3.6.3 Supuestos en el MANOVA.

Resulta muy importante revisar los datos para la búsqueda y eliminación de *outliers*⁷, pues el MANOVA es muy sensible a tales valores, los cuales impactan desproporcionadamente la tasa de error de tipo I [17].

El modelo lineal del MANOVA puede expresarse matricialmente como $y = A\beta + \epsilon$, en donde las ϵ son independientes, están normalmente distribuidas con media 0 y varianza común. También las y_i serían independientes, por lo que en forma matricial tenemos:

Ecuación 20 Esperanza matemática del vector de variables respuesta.

$$E(y) = A\beta \quad [9]$$

Las observaciones deben ser independientes; la más básica y seria violación de los supuestos del análisis multivariable de varianza, se da porque hay una gran variedad de situaciones de naturaleza tanto experimental como aleatoria, las cuales provocan observaciones dependientes. Un ejemplo, son las series ordenadas en el tiempo, o correlación en serie, donde empleamos los mismos individuos en un análisis antes-después. El investigador debe explorar todas las formas posibles de dependencia y corregirlas mediante covariables cuando sea posible [17].

El segundo supuesto es la igualdad de matrices de varianza y covarianza entre grupos. Al igual que con el problema de la heterocedasticidad en la regresión, podríamos encontrar varianzas diferentes en distintos grupos cuando se mide una misma variable. Afortunadamente una violación de este supuesto tiene un impacto mínimo si los tamaños de los grupos son aproximadamente iguales⁸. Para probar la igualdad de matrices de varianza y covarianza, se emplea el *test* de Box y la mayoría de los programas estadísticos dan niveles de significación para el mismo. Cuando se encuentra en el *test* un valor significativo que requiere remedio,

⁷ Los *outliers* son valores "situados fuera", es decir, valores atípicos, extremos y alejados del centro.

⁸ Si la razón entre el grupo mayor y menor no supera 1.5.

deben ensayarse transformaciones tendientes a estabilizar la varianza. La prueba de Box es muy sensible a fallas en la normalidad multivariable, por lo cual debe verificarse la normalidad de todas las variables dependientes antes de aplicar el *test* [17].

El tercer supuesto es el de normalidad multivariable, aunque afortunadamente, este supuesto es en el cual tienen menos impacto las violaciones. Su único efecto es crear problemas en la aplicación de la prueba de Box, pero las transformaciones usualmente corregirán ese problema [17].

3.6.4 Pruebas de hipótesis.

El análogo multivariable para probar las diferencias entre grupos es una función del valor propio de la matriz $SSCP_b \times SSCP_w^{-1}$, donde $SSCP_b$ es la matriz de sumas de cuadrados y productos cruzados entre grupos y $SSCP_w^{-1}$ la matriz inversa de sumas de cuadrados y productos cruzados dentro de grupos. Las medidas más populares son [30]:

Ecuación 21 Traza de Pillai.

$$\text{Traza} = \sum \frac{\lambda_i}{1 + \lambda_i}$$

Ecuación 22 Traza de Hotelling.

$$\sum \lambda_i$$

Ecuación 23 Lambda de Wilks.

$$\Lambda = \prod \frac{1}{1 + \lambda_i}$$

Ecuación 24 Raíz mayor de Roy.

$$\text{Raíz mayor} = \frac{\lambda_{\text{máx}}}{1 + \lambda_{\text{mín}}}$$

CUCA



BIBLIOTECA CENTRAL

donde λ es el i -ésimo valor propio y k el número de valores propios. La lambda de Wilks se puede aproximar a la distribución de F . El MANOVA toma en consideración la correlación entre las variables, mientras el ANOVA ignora esta información. Los programas estadísticos tales como SPSS y SAS dan una aproximación de F para cada prueba[30].

Además, cuando se ha probado que hay diferencias en varias dimensiones entre los grupos, las pruebas en orden decreciente de poder y robustez en el análisis son: Traza de Pillai, lambda de Wilks, traza de Hotelling y raíz mayor de Roy [24].

3.6.5 Análisis *post hoc*.

En el análisis univariable de varianza es necesario efectuar pruebas que permitan identificar cuáles promedios son diferentes, teniendo en cuenta que en tal procedimiento carecen de validez las pruebas de "t" pareadas, puesto que la probabilidad de cometer un error de tipo I aumenta con el número de comparaciones intergrupo efectuadas. De modo análogo, en el MANOVA debe realizarse un análisis que permita identificar las diferencias específicas, manteniendo bajo control las tasas de error de Tipo I. Entre los procedimientos más comunes están 1) La prueba de Scheffe; 2) La diferencia significativa honesta de Tukey (HSD, por sus siglas en inglés); 3) La extensión de Tukey del método de diferencia mínima significativa de Fisher (LSD); 4) La prueba de rango múltiple de Duncan y 5) La prueba de Newman-Keuls. Estas pruebas han sido catalogadas respecto a su poder; Las conclusiones muestran que la prueba de Scheffe es más conservadora respecto al error tipo I, seguida por las pruebas de Tukey (HSD), Tukey (LSD), Newman-Keuls y Duncan [17].

CUCBA



BIBLIOTECA CENTRAL

CUCBA



BIBLIOTECA CENTRAL

UNIVERSIDAD DE GUADALAJARA

4 Materiales y Métodos.

4.1 *Material genético.*

Se trabajó con la colección de maíces del banco de germoplasma de la División de Ciencias Agronómicas de la Universidad de Guadalajara. En total se evaluaron 30 variedades en un diseño de bloques completos al azar con tres repeticiones, tomando aleatoriamente 5 plantas en cada parcela experimental y determinando en las mismas los valores de cada una de las variables.

Entre las variedades evaluadas hubo criollos, razas e híbridos. La tabla 3 muestra el número asignado en el ensayo a cada variedad.

Tabla 3 Número asignado a cada variedad en el ensayo y genealogía correspondiente.

Número de variedad	Nombre de variedad	Número de variedad	Nombre de variedad
7	Amarillo oaxaca	27	Purísima guanajuato
20	Ancho negro	4	Santa Ana Tepetitlán
9	Blanco cristalino	12	Tequila
8	Blanco subtropical	15	Tesistán
16	Cuquío blanco	24	Tirándaro Michoacán amarillo
29	D 877 (Dekalb)	23	Tirándaro Michoacán blanco
30	D 880 (Dekalb)	1	Trejos
19	Gordo blanco	21	Turicato blanco
18	Güino blanco	22	Turicato negro
17	Güino negro	11	Tuxpeño
6	Jaleño	10	Tuxpeño caribe
13	La Mora, Teuchitlán	14	Ubalano
5	MGG1E	26	Valles altos blanco
28	Olote colorado	25	Valles altos negro
3	Población CIMMYT	2	Zapolote chico

El banco de datos que se formó tras el experimento, se muestra en el anexo 1.

4.2 *Localidad.*

El trabajo se desarrolló en el verano de 1994 en la hacienda "La Soledad", del Municipio de Zapopan, Jalisco. El predio se ubicó a 2.3 km al WNW del CUCBA, y sus coordenadas geográficas son 20° 45' 24" de latitud norte, 103° 31' 51" de longitud oeste y 1,640 msnm de altitud [6].

La siembra se efectuó el 1o. de Julio, 10 días después de la fecha recomendada por el INIFAP, lo cual podría sesgar los resultados del experimento, pero dado que el estudio de caso tiene fines ilustrativos, no se considera relevante.

El suelo de "La Soledad" es regosol éútrico de textura media [7]. En la nomenclatura de la clasificación de suelos FAO-UNESCO, los regosoles son suelos delgados de material no consolidado. Los regosoles son suelos sueltos derivados de cenizas volcánicas, con horizontes indiferenciados, muy permeables y en ocasiones de baja fertilidad [32].

Por ser suelos de textura media contienen menos de 65% de arena y menos de 35% de arcilla. No obstante, si la fracción de arcilla no es inferior al 18%, el porcentaje de arena podría ser hasta de 82%. La pendiente de "La Soledad" es ligeramente ondulada y no sobrepasa el 5%. Por la naturaleza poco consolidada de los agregados del suelo, hay una erosión laminar moderada, concierta degradación química debido a la percolación de bases intercambiables. Los regosoles éútricos tienen un horizonte A ócrico y un porcentaje de saturación con bases de 50% entre 20 y 50 cm de profundidad [13]. El denominado epipedón ócrico es un horizonte superficial de colores claros con valores de color mayores a 5.5 en seco y mayores a 3.5 en húmedo, cuyo contenido de materia orgánica es inferior al 1% [4]. Estos suelos tienen la tendencia de tomar una consistencia dura y masiva cuando están secos. Esta circunstancia provoca frecuentes resiembras o la necesidad de romper la capa masiva para facilitar la germinación de las semillas [25].

Los regosoles son equiparables a los entisoles. Estos suelos presentan problemas de ingeniería debido a su erodabilidad. En "La Soledad", los entisoles son fértiles y arables. Es notable su tendencia a inundarse cuando están en zonas bajas [4]. Entre las características climatológicas importantes para el cultivo del maíz en la zona, tenemos las determinadas por Villalpando y García [35], las cuales se sintetizan en la Tabla 4.

CUCEA



BIBLIOTECA CENTRAL

Tabla 4 Características agroclimáticas de Zapopan.

Característica	Valor
Temperatura media junio-octubre	18 a 22° C
Temperatura máxima junio-octubre	26 a 30° C
Temperatura mínima junio-octubre	<14° C
Amplitud térmica	14 a 16° C
Temperatura diurna	22 a 26° C
Temperatura nocturna	14 a 18° C
Unidades calor (T base 10° C)	1,600 a 2,000
Temperatura media anual	18 a 20° C
Precipitación pluvial junio-octubre (1)	600 a 800 mm
Días con lluvia apreciable junio-octubre	60 a 75
Evaporación junio-octubre	800 a 900 mm
Evapotranspiración potencial junio-octubre	600 a 700 mm
Índice de humedad	0.8 a 1.2 (2)
Fecha de inicio de la estación de crecimiento	16 a 30 de junio (1)
Fecha de término de la estación de crecimiento	16 a 31 de octubre (1)
Duración de la estación de crecimiento	120 a 150 días (1)
Fecha de inicio del periodo húmedo	julio 16 a 31 (1)
Fecha de término del periodo húmedo	Septiembre 16 a 30 (1)
Duración del periodo húmedo	60 a 90 días

(1) Para ésta característica, se consideró una probabilidad de lluvia al 70%.

(2) Se considera adecuada.

El maíz es una planta termófila que requiere para su crecimiento de 1,800 a 2,600 unidades calor (base 10° C). La temperatura óptima para el crecimiento de la planta es de 28 a 35° C, y las temperaturas que inhiben el mismo oscilan entre 45 a 47° C. La humedad del suelo más favorable para el crecimiento de la planta es de 60 a 70% de la capacidad de retención de humedad. Se estima que una planta adulta de maíz evapora de 2 a 4 kg de agua en un día. Con densidades de 40,000 a 60,000 plantas/ha la evaporación sería de 80 a 240 ton. Se requieren de 60 a 100 ton de agua para formar un quintal de grano de maíz [34].

4.3 Variables en estudio.

Las variables en estudio fueron: altura de planta, altura de mazorca, diámetro de tallo a la altura de la primera mazorca, número de mazorcas, número de hojas, longitud de mazorca, diámetro de mazorca, peso de mazorca, profundidad de grano, número de carreras, peso de grano y porcentaje de humedad (figura 4.5.1).

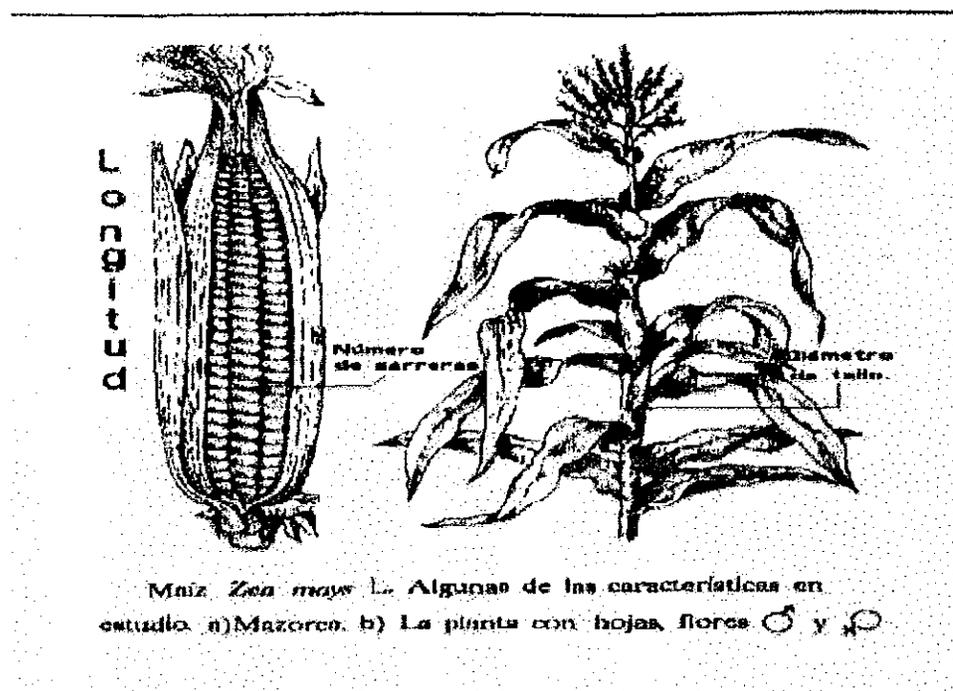


Figura 4.3-1 Algunas características del maíz bajo estudio.

La altura de planta se determinó mediante un estadal para topografía, desde la zona del cuello (transición entre tallo y raíz) hasta la espiga. Tomando la misma referencia, se midió la altura de mazorca. El diámetro de tallo se tomó a la altura de la primera mazorca. Debido a la forma oval del tallo, se determinó el diámetro mayor, mediante un pié de rey con aproximación a décimas de milímetro. Con el mismo instrumento se midió el diámetro de mazorca, en el tercio medio de la misma dada su forma aproximadamente cónica. Cuando una planta tuvo más de una mazorca, el diámetro representó el promedio aritmético de las mazorcas. La longitud de mazorca se midió con regla de madera de 400 mm.

Se contó el número de mazorcas y de hojas. De igual modo se procedió con las carreras de las mazorcas. En las plantas con más de una mazorca, el número de carreras fue el promedio aritmético de las mazorcas.

Los pesos de mazorcas y grano se determinaron mediante una balanza granataria "Ohaus" con precisión de ± 0.1 gr. En la determinación de la humedad se contó con el apoyo del personal del laboratorio de semillas de la División de Ciencias Agronómicas del CUCBA.

4.4 Análisis estadístico

La justificación de los estudios de maíz en nuestro medio obedece al creciente consumo aparente del grano en México, el cuál da fe de la necesidad de incrementar la producción mediante métodos científicos: Éste valor alcanzó la cifra de 15.66 millones de toneladas en 1991; 18.22, en 1992; 18.13, en 1994; 20.91, en 1994 [26]. En adición a lo anterior, se tienen datos preliminares para los siguientes dos años: 20.93 millones de toneladas en 1995; 23.80, en 1996 [29].

Antes de iniciar los análisis multivariables, se revisará que los datos no tengan problemas de falta de normalidad, multicolinealidad, valores extremos, o correlaciones bajas. Para lograr esta meta se procederá con gráficas de caja y bigote múltiples para las 12 variables en estudio, con la gráfica de ajuste a la normal, la prueba de χ^2 y la de Kolmogorov, utilizando Statgraphics versión 5.0. En caso de falta de ajuste a la normal, se efectuarán las transformaciones necesarias con los datos.

El resumen adecuado de las estadísticas de datos multivariables se obtiene mediante los vectores de medias y la matriz de covarianzas (o correlaciones). Estas cantidades describen la localización, variabilidad y dependencia en la población o en una distribución de probabilidades. Se creará la matriz de varianzas y covarianzas y el vector de promedios para cada una de las variedades. El próximo paso contempla la síntesis de la matriz de correlaciones, construida con todas las plantas registradas en la base de datos. Se sigue con la construcción de la matriz de distancias de Mahalanobis, luego con la prueba de normalidad multivariable (prerrequisito de los análisis multivariable de varianza y múltiple discriminante) [1].

El análisis de componentes principales se llevará a cabo partiendo de la matriz de correlaciones, debido a la diversidad de escalas de medición de las variables. Una vez conocidos los componentes, se elegirá el número de ellos mediante las pruebas de valor propio mayor que 1, el proceso alternativo y la gráfica *scree*. Una vez identificado el número de componentes principales, se aplicarán sus valores a los vectores de medias de las variedades para graficar los componentes principales elegidos y tratar de identificar conglomerados

preliminares. Con estos mismos valores de componentes principales se trabajará el AF, para determinar las dimensiones latentes del conjunto de datos de maíz, mediante el análisis de factores con rotación varimax y normalización de Kaiser. Determinar el valor de factor para cada variedad, nombrar los factores significativos identificando la información que cada cual aporta, y graficando los valores de cada variedad para facilitar la agrupación de las mismas por su afinidad multidimensional.

El análisis de conglomerados será jerárquico y se llevará a cabo utilizando la matriz de distancia cuadrada de Mahalanobis, puesto que este valor es el más representativo de la distancia entre variables correlacionadas. Con fines de comparación, se utilizarán como datos los valores de componentes principales, y los procedentes de la matriz de correlaciones, pues se espera la conformación de grupos consistentes. El método para identificar los grupos será el de centroides.

Se buscará una función capaz de discriminar las 450 plantas en los grupos determinados por el análisis de conglomerados, empleando el método *stepwise*. Se dividirá la muestra en dos conjuntos mediante muestreo estratificado aleatorio con asignación proporcional, y se utilizará el primero de ellos para construir la función discriminante y el otro para validarla. Se espera que la función discriminante supere por lo menos 25% en eficiencia a la designación por azar de las plantas en grupos.

Se probará la hipótesis de igualdad de vectores de medias para todas las variedades empleando el MANOVA, utilizando como función de prueba la lambda de Wilks y la F. Para la separación de vectores de medias o análisis *post hoc*, se utilizará la prueba de Bonferroni.

4.5 Material de soporte para el procesamiento.

El trabajo será efectuado mediante el programa SPSS versión 3.1 de 1989, junto con su módulo *advanced procedures*, en una computadora AT 80286 con 1Mb de memoria RAM. Las bases de datos y los procesos adicionales se harán en el mismo equipo, con el programa Quattro Pro versión 5.0 de 1993, y Statgraphics versión 5.0, de 1991. En vista de que SPSS

versión 3.1 no obtiene la matriz de distancias de Mahalanobis, se describirán los procedimientos y Macros para generarla en Quattro pro.

En adición, es importante identificar cuáles procesos pueden desarrollarse mediante una calculadora gráfica, pues el precio de uno de estos dispositivos es muy inferior al de las computadoras con software, y en algunas regiones eso puede representar la diferencia entre utilizar o no las técnicas. Se describirá la manera de aplicar la técnica de componentes principales mediante una calculadora Hewelett Packard 48G.





5 Resultados y Discusión.

5.1 Ajuste marginal a la distribución normal.

El resumen multivariable de la información se obtiene mediante los vectores de medias, la matriz combinada de varianzas y covarianzas, y la matriz de correlaciones. Estas matrices pueden verse en los anexos. La prueba de ajuste a la distribución normal cuando todas las plantas se analizaron simultáneamente fue muy pobre, principalmente por el sesgo de los mismos. Esta situación obligó a efectuar las pruebas de Kolmogorov-Smirnov para cada variedad en lugar de la Ji cuadrada.

Todas las plantas carentes de mazorcas mostraban fuertes valores extremos (*outliers*); considerando que por su naturaleza dichas plantas no dejarán descendencia, la variable denominada número de mazorcas fue eliminada del análisis por dos razones: contribuía significativamente a la pérdida de normalidad por aportar la mayor cantidad de datos extremos, y la variable en sí no pudo ser normalizada mediante transformaciones. Éste procedimiento obligó a eliminar 33 plantas, por lo cual el análisis se desarrolló con una base de datos con 417 registros en lugar de 450.

Otra variable que se aparta de la normalidad en algunos casos es el número de carreras, sobre todo cuando los valores registrados son muy uniformes. Las variedades más afectadas fueron la 7, 13 y 26. Todas las plantas de la variedad 7 tuvieron 14 carreras, 14 plantas de la variedad 13 tuvieron 14 carreras y otra tuvo 16, mientras que en la variedad 26 hubo seis plantas con 12 carreras t ocho con catorce. Se decidió considerar que la variable es normal, ya que en el 90% de los casos (27 variedades) eso fue verdadero. Los resultados se muestran en la tabla 5.

CUCBA



BIBLIOTECA CENTRAL

Tabla 5 Niveles de significación en la prueba de Kolmogorov Smirnov de ajuste a la distribución normal.⁸

Variable	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
ALTPL	.85	.64	.54	.56	.94	.80	.23	.16	.98	.67
ALTMZ	.83	.92	.16	.73	.79	.11	.89	.61	.91	.88
DMTLL	.62	.92	.50	.98	.82	.16	.16	.56	.78	.08
NO_MAZ	.01	.01	.00	.05	.00	.00	.00	.00	.00	.00
NO_HOJ	.07	.45	.34	.80	.94	.33	.86	.91	.58	.46
LONGMAZ	.77	.74	.99	.73	.99	.77	.84	.85	.99	.89
DIAMAZ	.22	.56	.89	.09	.93	.39	.20	.77	.39	.86
PESOMAZ	.82	.92	.91	.98	.41	.99	.91	.88	.34	.94
PROFG	.97	.58	.56	.53	.97	.87	.45	.42	.93	.98
NO_CARR	.11	.18	.67	.14	.13	.09	.02	.38	.48	.47
PESGR	.74	.89	.92	.97	.41	.99	.81	.99	.78	.74
P_HUM	.12	.87	.84	.44	.84	.99	.40	.32	.95	.84
Variable	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20
ALTPL	.99	.51	.99	.81	.79	.89	.91	.98	.93	.88
ALTMZ	.98	.09	.99	.94	.87	.69	.28	.61	.51	.95
DMTLL	.58	.78	.97	.98	.46	.62	.99	.79	.85	.73
NO_MAZ	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
NO_HOJ	.15	.89	.94	.53	.44	.83	.59	.75	.85	.75
LONGMAZ	.66	.79	.85	.88	.99	.96	.46	.91	.99	.91
DIAMAZ	.39	.89	.99	.37	.39	.88	.37	.76	.60	.76
PESOMAZ	.85	.77	.93	.96	.80	.96	.93	.88	.99	.88
PROFG	.75	.29	.92	.93	.50	.97	.55	.67	.99	.67
NO_CARR	.26	.15	.00	.30	.25	.21	.08	.05	.23	.05
PESGR	.98	.72	.74	.99	.70	.99	.68	.82	.74	.82
P_HUM	.20	.98	.88	.59	.47	.62	.12	.53	.60	.53
Variable	V21	V22	V23	V24	V25	V26	V27	V28	V29	V30
ALTPL	.99	.81	.61	.84	.70	.18	.78	.80	.99	.95
ALTMZ	.86	.87	.27	.91	.15	.21	.90	.77	.93	.99
DMTLL	.39	.42	.63	.77	.95	.37	.80	.88	.62	.99
NO_MAZ	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
NO_HOJ	.52	.71	.38	.14	.69	.77	.46	.45	.93	.79
LONGMAZ	.96	.97	.74	.97	.96	.39	.68	.90	.98	.09
DIAMAZ	.97	.76	.87	.98	.90	.95	.56	.44	.81	.38
PESOMAZ	.22	.99	.86	.95	.51	.57	.37	.99	.98	.03
PROFG	.99	.09	.76	.96	.60	.93	.55	.80	.47	.80
NO_CARR	.62	.34	.28	.08	.07	.04	.42	.36	.08	.26
PESGR	.45	.95	.99	.99	.77	.73	.18	.49	.97	.67
P_HUM	.17	.79	.37	.93	.66	.99	.81	.72	.43	.91

5.2 Correlaciones.

La tabla 6 es la matriz de correlaciones de las variables y muestra que todos los coeficientes de correlación simple son positivos, indicando con eso una relación directa entre

⁸ Esta prueba se realizó eliminando los datos con valores extremos. El nivel de significación se aproximó a centésimos.

las variables. La determinante de la matriz es 0.0037273, por lo cual no es singular y puede invertirse. La traza de la matriz de correlaciones es 11, lo cual comprueba la afirmación hecha respecto a la determinante.

Las variables peso de mazorca y peso de grano mostraron correlaciones con el 90.91% de las restantes variables, apoyando con ello la idea que tienen los fitomejoradores en el sentido de que los rendimientos son afectados por la mayor cantidad de genes. En orden de importancia, fueron seguidas por las variables: número de hojas, diámetro de tallo, longitud de mazorca, profundidad de grano y porcentaje de humedad, cada una de las cuales mostró correlaciones significativas con el 81.82% de las restantes variables. Las variables altura de planta y altura de mazorca tuvieron un 72.73% de correlaciones significativas con otras variables. Les siguieron las variables diámetro de mazorca con 50.54% de correlaciones significativas y número de carreras, con 45.45%.

Para que se considere significativa la correlación entre dos variables con una muestra de 417 plantas, se requiere que alcance un valor de $r = 0.126014070829$ ($\alpha = 0.01$) ó $r = 0.160585985325$ ($\alpha = 0.001$). Estos niveles fueron los elegidos para la significación debido a que la correlación es sensible al tamaño de muestra.

De las 55 correlaciones posibles entre 11 variables, fueron significativas 83.6364% de las mismas, lo cual hace de esta matriz un buen candidato al análisis de componentes principales.

CUCBA



BIBLIOTECA CENTRAL

Tabla 6 Matriz de correlaciones entre las variables en estudio.⁹

ALTPL	1.0000					
ALTMZ	0.8704**	1.0000				
DMTTLL	0.3806**	0.3591**	1.0000			
NO HOJ	0.4259**	0.3726**	0.4680**	1.0000		
LONGMAZ	0.5012**	0.4535**	0.2816**	0.3481**	1.0000	
DIAMAZ	0.0633	-0.12	0.1361	0.1399	0.3583**	1.0000
PESOMAZ	0.3381**	0.2735**	0.2727**	0.2852**	0.5656**	0.6990**
PROFG	0.2197**	0.1623**	0.2496**	0.2527**	0.2987**	0.4992**
NO CARR	0.0886	0.0389	0.1423*	0.2014**	0.0754	0.3716**
PESOGR	0.2847**	0.2363**	0.2455**	0.2148**	0.4400**	0.6233**
P HUM	0.4367**	0.3944**	0.1669**	0.3038**	0.3807**	0.2233**
	PESOMAZ	PROFG	NOCARR	PESOGR	P HUM	
PESOMAZ	1.0000					
PROFG	0.4920**	1.0000				
NO CARR	0.2026**	0.1198	1.0000			
PESOGR	0.8465**	0.4030**	0.1532**	1.0000		
P HUM	0.4405**	0.2588**	0.0986	0.3451**	1.0000	

5.3 Análisis de componentes principales.

La tabla 7 muestra los resultados de los valores propios de la matriz de correlaciones (valores propios) y el porcentaje de la variación total o conjunta que explica cada uno de ellos. Éste último se determina dividiendo cada valor propio de la matriz de correlaciones y permite identificar las variables más importantes que contribuyen a una varianza común.

Antes de efectuar el análisis, se aplicó la prueba propuesta por Montgomery y Peck [22] para determinar si existen problemas con la multicolinealidad¹⁰: el mayor valor propio es de 4.331320 y el menor 0.122931. La razón k entre ambas cantidades es 35.23375, por lo cual se considera leve. La tabla 7 muestra que solo tres de los componentes principales tienen valores propios mayores a la unidad y son capaces de explicar conjuntamente el 66.473% de la variación del conjunto de datos.

⁹ * = Nivel de significación 0.01. ** = Nivel de significación 0.001.

¹⁰ Ver ecuación 8, página 19

Tabla 7 Valores propios de la matriz de correlaciones.

Componente número	Valor propio	% de la varianza explicado	% de varianza explicado acumulado
1	4.331320	39.37564	39.37564
2	1.891970	17.19973	56.57537
3	1.088690	9.89718	66.47255
4	0.854148	7.76498	74.23753
5	0.682767	6.20697	80.44450
6	0.632913	5.75375	86.19825
7	0.548699	4.98817	91.18642
8	0.454241	4.12946	95.31588
9	0.267256	2.42960	97.74548
10	0.125064	1.13695	98.88243
11	0.122931	1.11755	100.00000
Total	11.00000		

Las calificaciones para los tres primeros componentes principales se muestran en la tabla 8.

Tabla 8 Cargas de los primeros tres componentes principales.

Variable	Z1	Z2	Z3
ALTPL	0.328305	0.4335660	-0.0906603
ALTMZ	0.294395	0.4707570	-0.1193270
DMTLL	0.248574	0.2022110	0.4351840
NO HOJ	0.271115	0.2291670	0.4215620
LONGMAZ	0.344544	0.0628211	-0.2057140
DIAMAZ	0.290833	-0.4838580	0.0665072
PESOMAZ	0.398992	-0.2845580	-0.1718780
PROFG	0.279154	-0.2193560	0.0296040
NO CARR	0.138664	-0.1816020	0.6634980
PESOGR	0.356182	-0.2932590	-0.1984560
P HUM	0.289390	0.1042200	-0.2316640

Se puede observar que las cargas en el primer componente principal tienen todas el mismo signo y, salvo por la variable "número de carreras", aproximadamente el mismo valor. De acuerdo con lo que se presenta en la tabla 8, los caracteres con calificación mayor en este componente son: longitud de mazorca, profundidad de grano, peso de grano y porcentaje de humedad. Todas estas variables están relacionadas con la producción de grano y como puede verse, tienen signo positivo, confirmando las tendencias que se observan en la matriz de correlaciones: Las variedades con mazorcas más largas tienden a poseer mazorcas más pesadas, con mayor peso y profundidad de granos, y mayor contenido de humedad. Un valor alto del primer componente principal identificaría a las buenas variedades para la producción de grano. Denominaremos al componente principal 1 "producción de grano", de manera que es factible pensar que el 39.37564% de las variaciones entre las variedades de maíz estudiadas consisten en diferencias para producir grano. Adicionalmente, las plantas con mayor

calificación en el componente principal 1 tendrán mayor tamaño en relación con las que tengan baja calificación. Para la visualización de las cargas por componente principal, véase la figura 5.3-2.

El segundo componente principal muestra un contraste interesante: con signo positivo están las variables altura de planta, altura de mazorca, diámetro de tallo, número de hojas, casi con la misma calificación, todas ellas relacionadas con el crecimiento vegetativo de las plantas. Con signo negativo están las variables: diámetro de mazorca, peso de mazorca, peso de grano y número de carreras, todas ellas relacionadas con la producción de grano. El segundo componente principal explica el 17.19973% de la variación total. Después de la producción de grano (componente principal 1), es importante distinguir entre plantas que producen grano y las que producen follaje. Los dos componentes juntos explican 56.57537% de la variación conjunta de todas las variables. Las plantas con alto valor en el componente principal dos serán más productoras de follaje y menos eficientes para la producción de grano.

El tercer componente principal muestra un contraste entre diámetro de tallo, número de hojas y número de carreras por una parte, y por la otra peso de grano, porcentaje de humedad, peso y longitud de mazorca. Las plantas con alto valor para este componente tienen tallos gruesos y muchas hojas, pero son relativamente poco eficientes para producir grano, pues aunque sus mazorcas tienen muchas carreras son relativamente cortas y no proporcionalmente más gruesas. Este componente explica el 9.887% de la variación total y en conjunto con los otros dos, 66.045% de la misma.

Las correlaciones entre las variables y los componentes principales que se muestran en la tabla 9, en donde se han marcado con asterisco los coeficientes de correlación mayores que 0.5, indican que a excepción de la variable número de carreras todas las demás correlacionan positivamente con el componente principal 1. El componente principal 2 correlaciona positivamente con las variables altura de planta y de mazorca y negativamente con diámetro de mazorca indicando que las plantas más altas cuyas mazorcas se insertan a mayor distancia del piso tienen mazorcas más delgadas. La correlación del tercer componente principal es alta

sólo con la variable número de carreras, con lo cual resulta ser este el componente con menor correlación de todos, lo cual también cobra sentido al revisar las correlaciones simples¹¹.

Tabla 9 Correlaciones entre las variables y los componentes principales.

Variable	Z1	Z2	Z3
ALTPL	0.683263*	0.5963653*	-0.0945952
ALTMZ	0.612690*	0.6475211*	-0.1245062
DMTILL	0.517328*	0.2781390	0.4540723
NO HOJ	0.564240*	0.3152167	0.4398591
LONGMAZ	0.717059*	0.0864097	-0.2146426
DIAMAZ	0.605277*	-0.6655414*	0.0693938
PESOMAZ	0.830375*	-0.3914064	-0.1793380
PROFG	0.580970*	-0.3017218	0.0308889
NO CARR	0.288585	-0.2497916	0.6922959*
PESOGR	0.741280*	-0.4033745	-0.2070696
P HUM	0.602273*	0.1433535	-0.2417189

El siguiente paso en el análisis consiste en calcular las calificaciones de componente principal para los vectores de medias estandarizados de cada variedad, lo cual permite elaborar un gráfico para agrupar variedades afines. Gracias al procedimiento de componentes principales, hemos conseguido reducir el problema de 11 a sólo tres dimensiones, aunque para eso hemos tenido que sacrificar una parte de la información correspondiente a un 33.955% de la varianza. La figura muestra la representación gráfica de las variedades en función de los dos primeros componentes principales. Arbitrariamente, se formaron siete grupos de variedades en función de esta agrupación. Las calificaciones de cada variedad se muestran en la tabla 10.

La variedad Trejos¹² no puede compararse con ninguna y conforma un grupo ella sola. La gráfica, muestra que es una variedad equilibrada en tamaño y producción de grano. El segundo grupo está formado por las variedades Blanco Cristalino, Blanco Subtropical y Tuxpeño Caribe, mismas que tienen un tamaño superior al promedio, pero son relativamente ineficientes para producir grano.

¹¹ Ver página 56.

¹² Ver Tabla 3, página 45.



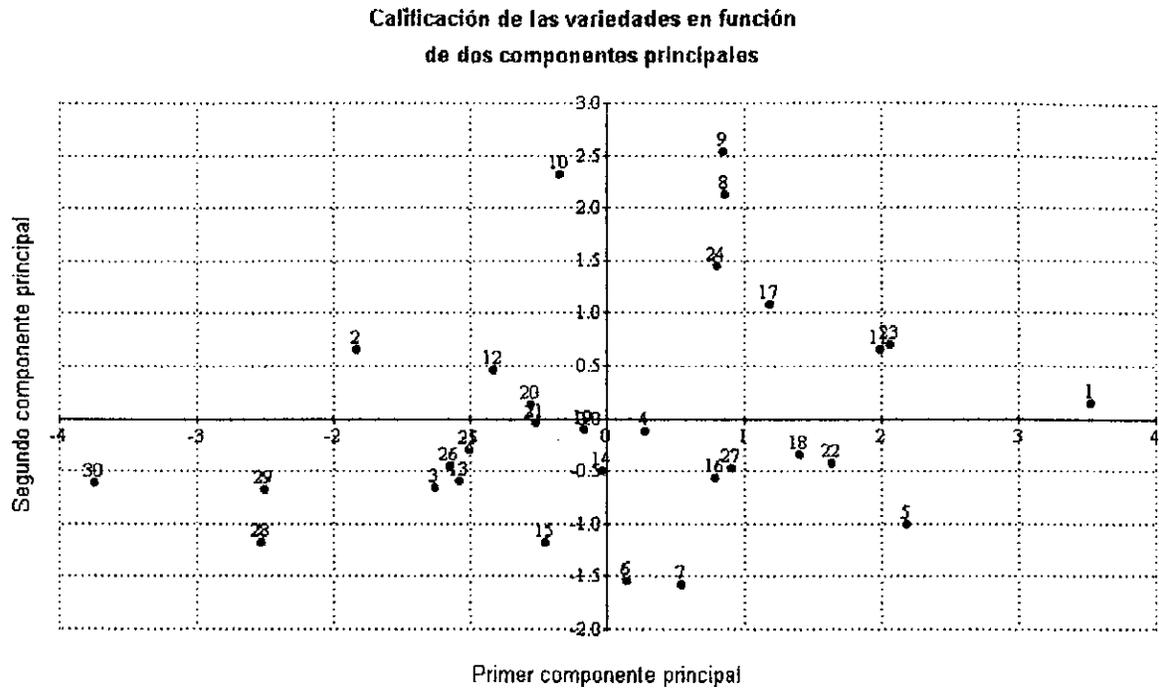


Figura 5.3-1 Calificación de las variedades en los primeros componentes principales

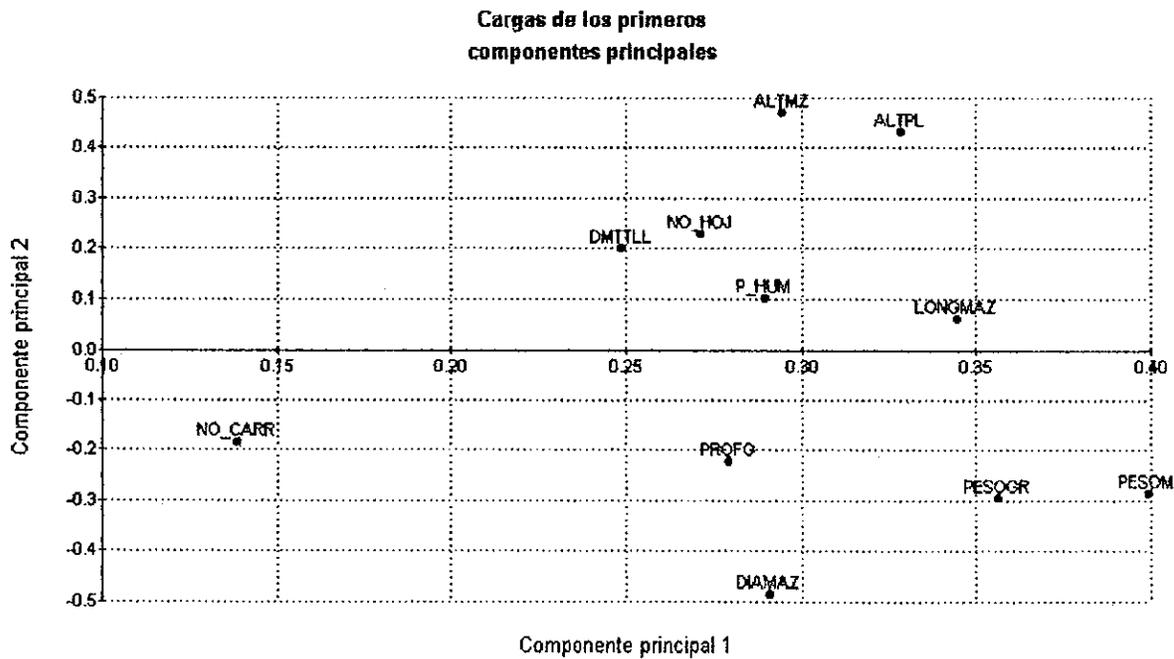


Figura 5.3-2 Cargas de variable en los primeros componentes principales

El tercer grupo lo forman las variedades Tirándaro Michoacán Blanco, Tuxpeño, Güino Negro y Tirándaro Michoacán Amarillo, que son variedades con tamaño superior al promedio y tendencia a producir más follaje que grano, sin que la producción de grano sea mala. El cuarto grupo está formado por las variedades MGG1E, Turicato Negro, Güino Blanco, Purísima Guanajuato y Cuquíó Blanco, las cuales son variedades con tamaño superior al promedio y relativamente eficientes para producir grano. Si el interés es la producción de grano, este grupo fue el más atractivo después de la variedad trejos. El quinto grupo es el más abundante y está formado por las variedades Santa Ana Tepetitlán, Ubalano, Gordo Blanco, Turicato Blanco, Ancho Negro, Tequila, Valles Altos Negro, la mora teuchitlán, Valles Altos Blanco, población CIMMYT y Zapalote Chico, las cuales tienen aproximadamente un tamaño promedio y una eficiencia promedio para la producción de granos. El sexto grupo está formado por las variedades Amarillo Oaxaca, Jaleño y Tesistán, que tienen un tamaño promedio pero son eficientes en producir grano. El séptimo grupo lo forman las variedades Olote Colorado, D877 y D880, mismas que tienen un tamaño inferior al promedio y producen poco grano. Debido a su baja biomasa total, la producción de estas variedades no es buena ni en follaje ni en grano

Tabla 10 Calificaciones de cada variedad en los tres componentes principales.

Variedad	Z1	Z2	Z3
1	3.5218	0.1448	-0.0020
2	-1.8303	0.6511	-1.1131
3	-1.2585	-0.6528	0.6518
4	0.2744	-0.1234	0.4482
5	2.1773	-1.0078	-0.6718
6	0.1435	-1.5468	-0.5515
7	0.5308	-1.5702	-0.5190
8	0.8629	2.1236	-0.2470
9	0.8465	2.5343	0.7305
10	-0.3498	2.3212	0.2448
11	1.9937	0.6595	0.4542
12	-0.8382	0.4676	-1.1145
13	-1.0786	-0.5954	0.5315
14	-0.0350	-0.4987	-0.2066
15	-0.4515	-1.1789	1.2598
16	0.7884	-0.5661	-0.1785
17	1.1748	1.0840	-0.7415
18	1.3883	-0.3462	-0.0622
19	-0.1628	-0.0954	0.4236
20	-0.5641	0.1307	1.0229
21	-0.5214	-0.0435	0.7390
22	1.6311	-0.4163	0.1748
23	2.0617	0.7050	-0.2845
24	0.7969	1.4546	-0.2073
25	-1.0101	-0.2852	0.8715
26	-1.1493	-0.4525	0.0296

Variedad	Z1	Z2	Z3
27	0.9062	-0.4618	0.3244
28	-2.5395	-1.1792	-0.8588
29	-2.5112	-0.6745	-0.4370
30	-3.7489	-0.6121	-0.6050

El proceso de análisis de componentes principales se puede desarrollar en una calculadora HP48G o superior. Para ello es necesario construir la matriz de correlaciones y utilizar como algoritmo básico la descomposición espectral de valores singulares¹³, proceso que factoriza una matriz no singular, cuadrada y simétrica¹⁴ en tres [5].

Ecuación 18 Descomposición espectral de valores singulares.

$$A = TDT^{-1}$$

El procedimiento se basa en la diagonalización ortogonal de la matriz, pues **T** es una matriz ortogonal. La matriz **D** es diagonal, siendo los valores ubicados en su diagonal los valores propios de la matriz **A**. Con la matriz de varianzas y covarianzas o con la de correlaciones, el orden de las matrices es $p \times p$, donde p es el número de variables en estudio [5]. El procedimiento se ilustra en la figura 5.3-3.

CUCBA



BIBLIOTECA CENTRAL

¹³ SVD por sus siglas inglesas.

¹⁴ Tales son los casos de las matrices de varianzas y covarianzas y correlaciones.

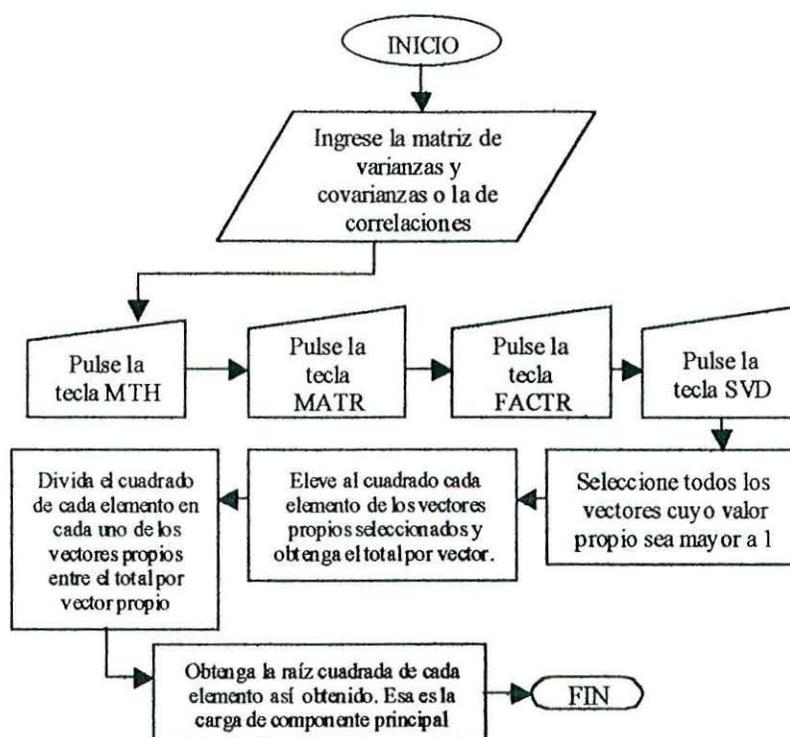


Figura 5.3-3 Diagrama de flujo que muestra el cálculo de componentes principales con calculadora HP48G

La tabla 11 muestra la descomposición de valores singulares de la matriz de correlaciones de acuerdo al procedimiento señalado.

Tabla 11 Descomposición de valores singulares de la matriz de correlaciones (Matriz T).

Vector1	Vector2	Vector3	Vector4	Vector5	Vector6	Vector7	Vector8	Vector9	Vector10	Vector11
-0.32731	-0.44607	0.100496	0.177424	-0.21576	-0.20188	-0.11286	-0.05057	0.152624	0.332357	-0.64494
-0.29224	-0.48223	0.121234	0.144450	-0.30963	-0.19996	-0.15592	-0.10817	0.101499	-0.30114	0.611253
-0.24988	-0.21928	-0.40844	-0.50452	0.050247	0.318229	-0.39459	0.452779	0.059201	-0.00965	0.002212
-0.27245	-0.24287	-0.39810	-0.16756	0.355025	0.144090	0.426425	-0.58996	0.054610	0.006801	0.022318
-0.34662	-0.07449	0.223656	-0.05907	-0.08292	0.222287	0.679301	0.507902	-0.17045	0.095435	0.086642
-0.29546	0.475351	-0.08300	0.024320	-0.14217	-0.04080	0.072401	0.025325	0.791779	0.114601	0.107039
-0.40314	0.273938	0.170607	-0.01188	-0.08733	0.188152	-0.08474	-0.14101	-0.13750	-0.71431	-0.36599
-0.28232	0.209715	-0.02646	-0.41844	0.112485	-0.78072	0.061033	0.079372	-0.25780	0.034576	0.025648
-0.14063	0.174833	-0.69461	0.574318	-0.16704	-0.10567	-0.02726	0.126485	-0.28512	-0.00931	0.011745
-0.36000	0.283175	0.196167	-0.05546	-0.17244	0.298818	-0.28227	-0.32222	-0.36577	0.514013	0.220122
-0.26786	-0.01081	0.206844	0.391320	0.789939	-0.02767	-0.25960	0.167520	0.059146	0.027289	0.088047

La tabla 12 muestra las cargas de componentes principales obtenidas con la calculadora, y aunque se advierte una pequeña diferencia en los decimales la solución es prácticamente igual que la obtenida mediante la computadora. Los signos sólo se interpretan como valores contrarios, así que podemos respetar los signos que proporcionó la matriz T o multiplicar ésta matriz por -1 .

Tabla 12 Cargas de los primeros tres componentes principales mediante la calculadora HP 48G.

Variable	Z 1	Z 2	Z 3
ALTPL	-0.327314	-0.4460710	0.1004956
ALTMZ	-0.292242	-0.4822323	0.1212340
DMTILL	-0.249890	-0.2192806	-0.4084385
NO HOJ	-0.272446	-0.2428709	-0.3981046
LONGMAZ	-0.344662	-0.0744916	0.2223655
DIAMAZ	-0.295457	0.4753511	-0.0829974
PESOMAZ	-0.403139	0.2739378	0.1706074
PROFG	-0.282318	0.2097160	-0.0264614
NO CARR	-0.140630	0.1748336	-0.6946119
PESO GR	-0.359996	0.2831755	0.1961670
P HUM	-0.267858	-0.0181456	0.2068438

5.4 Análisis de factores.

El análisis de factores se hizo con la matriz de correlaciones, pues las variables de medición tienen escalas de medición muy diferentes. La prueba de Kaiser-Meyer-Olkin de adecuación de muestreo de la matriz de correlaciones tiene un valor de 0.79084, el cual se considera más que regular. La prueba de esfericidad de Bartlett tuvo un valor de 2,342.3215 y es altamente significativa, por lo cual la matriz puede factorizarse.

Los valores propios de la matriz de correlaciones que se mostraron en la tabla 7 indican la conveniencia de extraer 3 factores, debido a que sólo los primeros tres valores propios de la matriz de correlaciones tienen un valor propio mayor que 1. La gráfica *scree* presenta un codo en el tercer factor extraído, coincidiendo con la regla de valor propio mayor que 1 en recomendar la extracción de tres factores.

CUCBA



BIBLIOTECA CENTRAL

Gráfica scree para análisis de factores

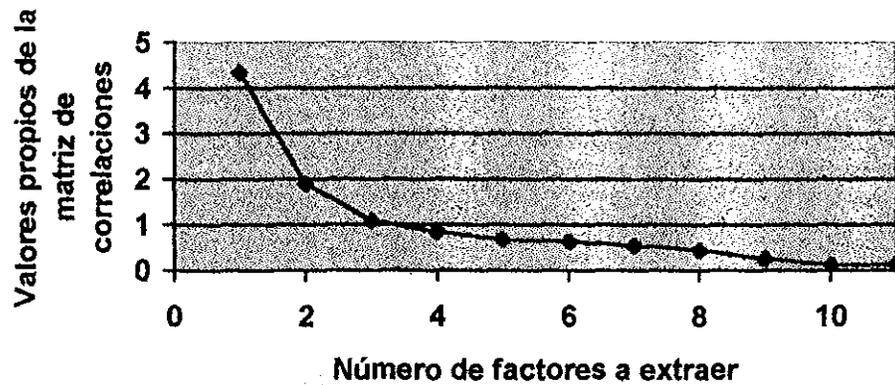


Figura 5.4-1 Gráfica scree para determinar el número de factores a extraer.

La matriz de correlaciones anti-imagen muestra doce elementos (equivalentes al 10.9% de las correlaciones parciales posibles) fuera de la diagonal principal con valores superiores a 0.09, por lo cual la solución por factores puede ser adecuada. La tabla 13 muestra la medida de adecuación de muestreo para las variables, donde las variables que más contribuyeron a una varianza común fueron el porcentaje de humedad y la longitud de mazorca; seguidas por la profundidad de grano, diámetro de tallo y número de hojas; diámetro de mazorca, peso de mazorca, altura de planta y altura de mazorca. Finalmente, la puntuación más baja la tuvo la variable número de carreras, y eso se debió a que mazorcas con número suficiente de carreras tuvieron problemas con el llenado de grano, dando entonces esta variable un resultado inconsistente con fuertes singularidades.

Tabla 13 Medida de adecuación de muestreo de Kaiser-Meyer-Olkin para las variables.

Variables	Adecuación de muestreo	Calificación KMO
Altura de planta	0.71914	Regular
Altura de mazorca	0.70054	Regular
Diámetro de tallo	0.84727	Meritorio
Número de hojas	0.84584	Meritorio
Longitud de mazorca	0.90515	Extraordinario
Diámetro de mazorca	0.79097	Regular
Peso de mazorca	0.76714	Regular
Profundidad de grano	0.87191	Meritorio
Número de carreras	0.60088	Mediocre
Peso de grano	0.78446	Regular
Porcentaje de humedad	0.92141	Extraordinario

Después de la rotación de factores, la cual como se realizó mediante el método varimax se encontraron las cargas de factor y las comunalidades que se muestran en la tabla 14.

Tabla 14 Cargas de factor y comunalidades por variable estudiada.

Variable	Comunalidad	Factor 1	Factor 2	Factor 3
Altura de planta	0.83145	0.10546	0.89097*	0.16280
Altura de mazorca	0.81017	0.02747	0.89246*	0.11375
Diámetro de tallo	0.55117	0.08332	0.41315	0.61117*
Número de hojas	0.61120	0.09481	0.47341	0.61489*
Longitud de mazorca	0.56771	0.49724	0.56487*	0.03717
Diámetro de mazorca	0.81412	0.86123*	-0.13332	0.23372
Peso de mazorca	0.87488	0.89058*	0.27277	0.08570
Profundidad de grano	0.42952	0.60824	0.13062	0.20617
Número de carreras	0.62495	0.22385	-0.18044	0.73640*
Peso de grano	0.75508	0.84130	0.21545	0.02971
Porcentaje de humedad	0.44171	0.38336	0.54240*	-0.02348
Total ¹⁵		3.0847	2.7866	1.4406
Porcentaje		28.0427	25.3327	13.0964
Porcentaje acumulado		28.0427	53.3755	66.4718

Las comunalidades permiten identificar las variables que mejor contribuyen a explicar la variación común, utilizando nuevamente para ello el criterio de Kaiser-Meyer-Olkin. Altura de planta, altura de mazorca, diámetro de mazorca, y peso de mazorca tienen un valor meritorio en su contribución a la variación común. El peso de grano es una variable con valor regular. Quizá en un posterior análisis de factores confirmatorio, puedan desecharse las variables con valor mediocre, miserable o inaceptable: Diámetro de tallo, número de hojas, longitud de mazorca, profundidad de grano, número de carreras y porcentaje de humedad, que tienen fuertes singularidades. Los totales de la tabla muestran que tras la rotación de factores, la variación identificada por cada uno de estos es diferente a la de los componentes principales, pero los tres factores acumulados explican la misma proporción de variación que los tres primeros componentes principales.

En la tabla 14 se destacaron con un asterisco las cargas de factor mayores a 0.5, permitiendo identificar en cual factor tiene mayor importancia cada variable y ayudando así a darles nombre a los factores. En el primer factor, las variables con mayor carga son diámetro de mazorca, peso de mazorca, profundidad de grano y peso de grano, todas ellas relacionadas con plantas con buen rendimiento de grano. Este factor se denominará "Producción de grano".

¹⁵ Se calcula elevando al cuadrado las cargas de factor y sumando después.

Las plantas con valores bajos en este factor serán pobres productoras de grano, y aquellas con alto valor en este factor serán buenas productoras de grano.

En el segundo factor las variables con mayor carga son altura de planta, altura de mazorca, longitud de mazorca y porcentaje de humedad, todas ellas con carga positiva; estas variables también están relacionadas con la longevidad de las plantas de maíz, por lo que será denominado "Temporalidad". Las plantas con valores pequeños en este factor serán más precoces, y aquellas con valor alto, más tardías.

En el tercer factor las variables con mayor carga son diámetro de tallo, número de hojas y número de carreras, relacionadas con un aspecto vigoroso de las plantas, y se denominará "Vigor". Las plantas con alto valor en esta variable tienen muchas hojas, muchas carreras y tallo vigoroso, mientras aquellas con bajo valor tendrán pocas hojas, tallos delgados y pocas carreras.

Como puede apreciarse, la solución de factores difiere algo de la de componentes principales, es más fácil de interpretar y tiene mayor soporte teórico. Para revisar la solución de factores, empleamos la matriz de correlaciones reproducida para determinar las diferencias entre las correlaciones originales y aquellas que pueden reconstruirse a partir de los factores. Estas diferencias son los residuales, con los que se efectúa la prueba RMSR (raíz cuadrada de residuales), cuyos resultados se muestran en la tabla 15. Los residuales de todas las variables no son muy grandes, dejando así la certeza de que todas las variables han contribuido en cierta medida a construir una buena solución de factores.

Se estandarizaron las medias de las variables y con las cargas de los factores, se procedió a encontrar los valores de factor para cada variedad, realizando después el análisis de conglomerados. Las cargas de factor para cada variedad se muestran en la tabla 16. Una gráfica de los valores de variedad permite suponer la posibilidad de conformar 8 grupos, y ese fue el número de grupos propuesto en inicio para el análisis de conglomerados. Los datos se muestran en orden creciente para el primer factor, que como se mencionó, contribuye a explicar la mayor variación conjunta de los factores. La gráfica de los valores de factor por variedad se muestra en la figura 8.

Tabla 15 Prueba RMSR de cada variable

Variable	RMSR
Altura de planta	0.0295822
Altura de mazorca	0.0322491
Diámetro de tallo	0.0452320
Número de hojas	0.0370481
Longitud de mazorca	0.0286014
Diámetro de mazorca	0.0192089
Peso de mazorca	0.0189385
Profundidad de grano	0.0290900
Número de carreras	0.0560091
Peso de grano	0.0308990
Porcentaje de humedad	0.0345816
Total	0.0793956

Tabla 16 Valores de factor para cada variedad.

Variedad	Producción de grano	Temporalidad	Vigor
1.- Trejos	5.1483	4.9167	2.2686
2.- Zapalote Chico	-3.1851	-1.4370	-2.3434
3.- Población CIMMYT	-1.4894	-2.5548	-0.0188
4.- Santa Ana Tepetitlán	1.4208	0.0380	0.9715
5.- MGG1E	4.3639	2.0614	0.5767
6.- Jaleño	1.7184	-1.2956	-0.2995
7.- Amarillo Oaxaca	2.3192	-0.8196	0.0236
8.- Blanco Subtropical	-0.7572	3.3947	0.1444
9.- Blanco Cristalino	-1.3609	3.5495	1.0801
10.- Tuxpeño Caribe	-2.8388	1.8686	-0.1728
11.- Tuxpeño	2.2342	3.1897	1.6544
12.- Tequila	-1.4990	-0.3512	-1.5876
13.- La Mora, Teuchitlán	-1.2327	-2.2148	-0.0616
14.- Ubalano	0.4225	-0.4886	-0.2525
15.- Tesistán	0.1166	-2.1985	1.0751
16.- Cuquío Blanco	1.7352	0.5250	0.2852
17.- Gúino Negro	1.5137	2.8277	0.1712
18.- Gúino Blanco	2.3930	1.4961	0.7502
19.- Gordo Blanco	-0.3286	-0.4332	0.3351
20.- Ancho Negro	-1.2051	-0.8302	0.4953
21.- Turicato Blanco	-0.9155	-0.9049	0.2967
22.- Turicato Negro	2.7863	1.6929	1.1544
23.- Tirándaro Michoacán Blanco	2.4487	3.5880	0.8821
24.- Tirándaro Michoacán Amarillo	-0.1689	2.6505	0.1340
25.- Valles Altos Negro	-1.4888	-1.9142	0.3343
26.- Valles Altos Blanco	-1.3397	-1.9941	-0.6905
27.- Purísima, Guanajuato	1.6977	0.6393	0.8638
28.- Olote Colorado	-2.5615	-4.3428	-2.4063
29.- D 877	-3.0994	-3.9172	-1.9798
30.- D 880	-5.0137	-5.4135	-2.9064

Si empleamos el análisis de factores como medio para determinar cuáles son las variables más influyentes en las comunalidades, y por tanto, más importantes para determinar la variabilidad total, nos deberemos quedar con aquellas que en la prueba de Kaiser-Meyer-

Olkin sean regulares o más. La única variable que no coincidiría con ese criterio, y por tanto podría ser desechada, es el número de carreras.

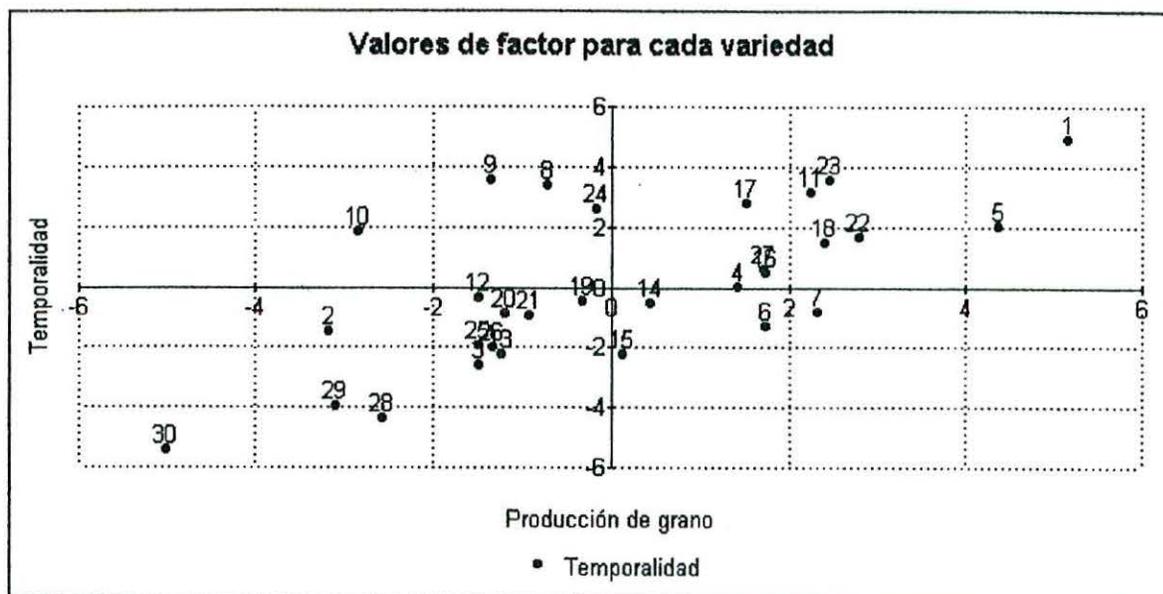


Figura 5.4-2 Calificación de factores para las variedades.

5.5 Análisis de conglomerados.

Con los valores de factor para cada variedad, se desarrolló el análisis de conglomerados (*cluster*), el cual inició determinando una medida de disimilaridad. La opción de una medida de disimilaridad en SPSS consiste en la distancia euclidiana cuadrada, misma que considera los valores de factor como coordenadas. Se determinaron entonces tres coordenadas para cada promedio varietal y con ellas se procedió a definir las variedades más parecidas por su composición en cuanto a producción de grano, temporalidad y vigor (Tabla 15), mediante el método de centroides. El empleo de valores de factor en éste cálculo protege contra posibles fallas originadas por la multicolinealidad.

CUCBA



BIBLIOTECA CENTRAL

Se conformaron 7 grupos de acuerdo con los codos identificados en los gráficos de r^2 y RMSSTD (figura 5.5-2). El dendrograma correspondiente se muestra en la figura 5.5-1, en la cual las distancias se han reescalado a un máximo de 25 unidades.

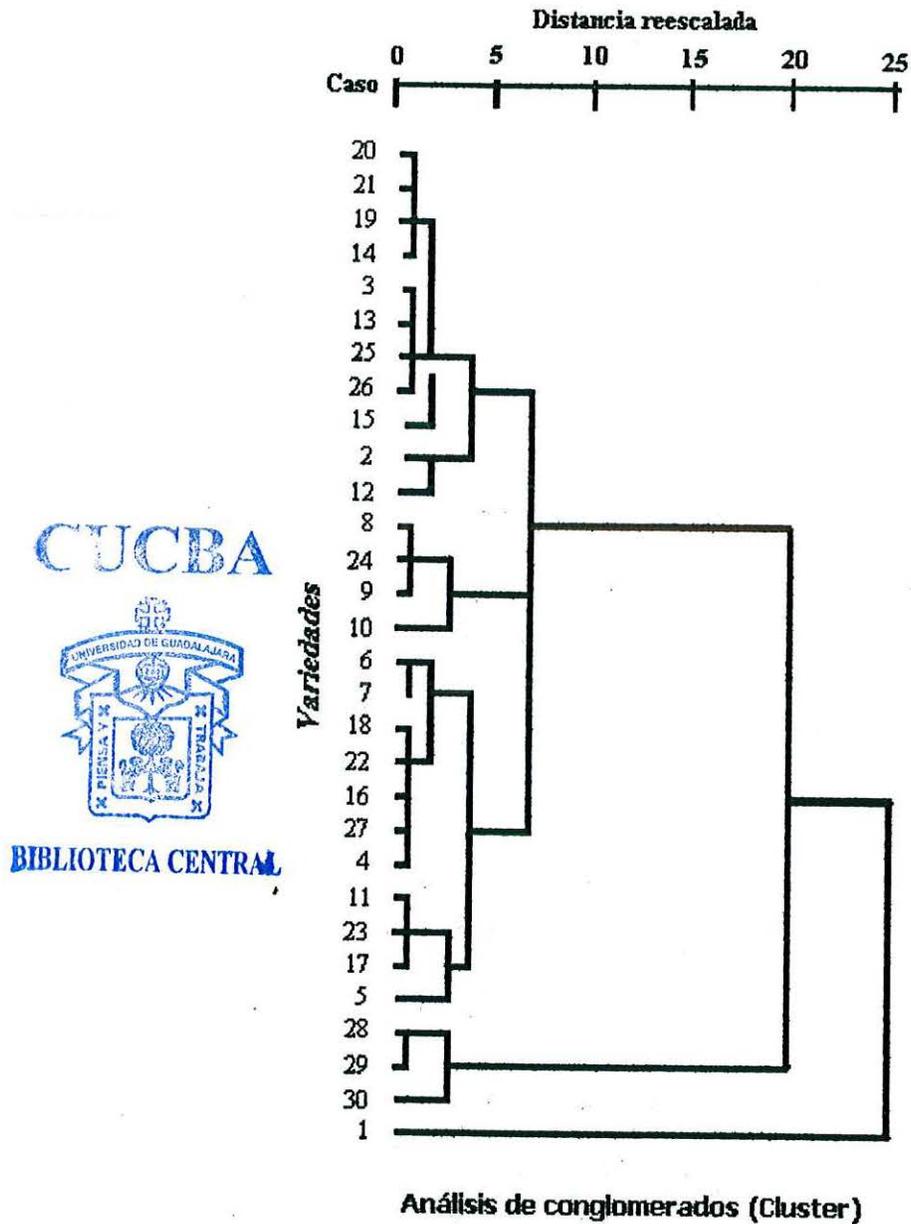


Figura 5.5-1 Dendrograma de las 30 variedades de maíz estudiadas.

Los 7 conglomerados definidos en la solución son: I) La variedad Trejos; II) Las variedades Zapalote Chico y Tequila; III) Las variedades población CIMMYT, La Mora Teuchitlán, Ubalano, Tesistán, Gordo Blanco, Ancho Negro, Turicato Blanco, Valles Altos Negro, Valles Altos Blanco; IV) Santa Ana Tepetitlán, Jaleño, Amarillo Oaxaca, Cuquío Blanco, Güino Blanco, Turicato Negro, Purísima Guanajuato; V) MGG1E, Tuxpeño, Güino Negro, Tirándaro Michoacán Blanco; VI) Blanco Subtropical, Blanco Cristalino, Tuxpeño Caribe, Tirándaro Michoacán Amarillo; VII) Olote Colorado, D877, D880.

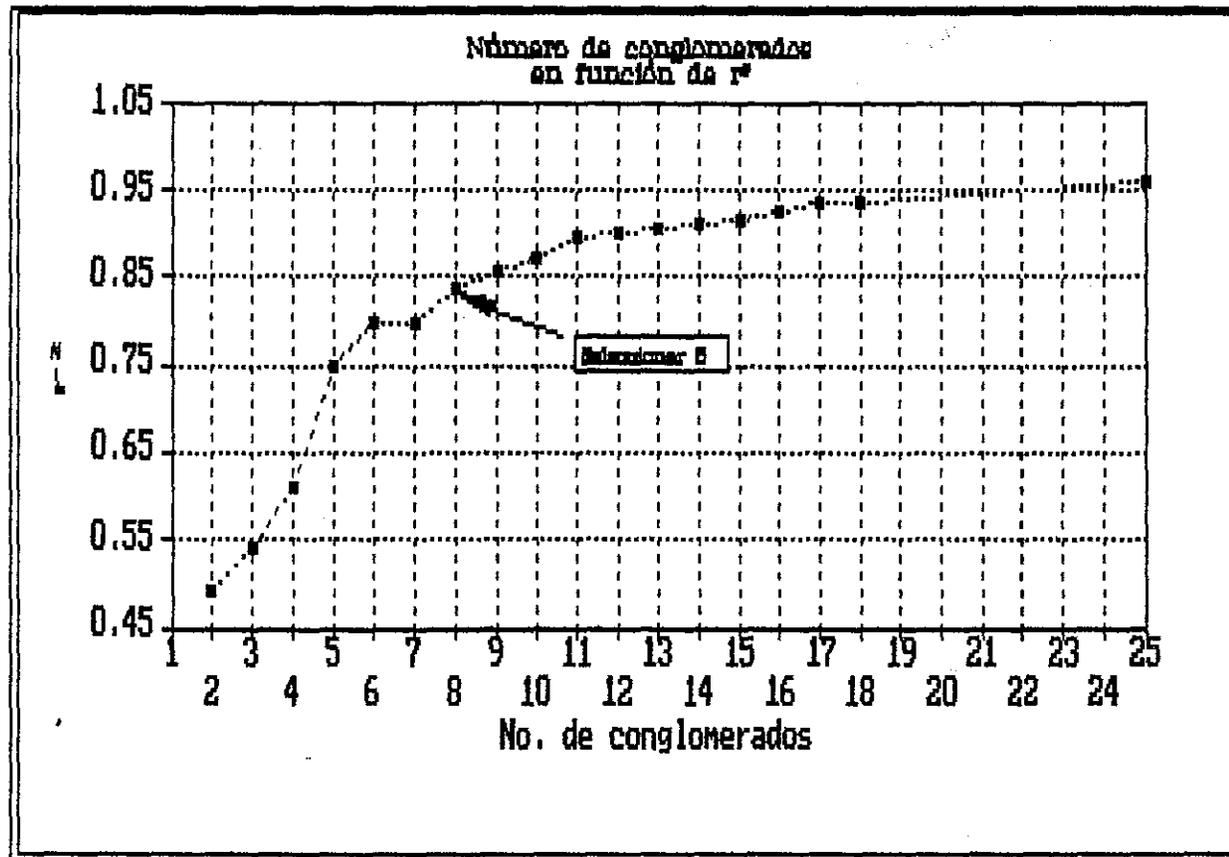


Figura 5.5-2 Número de conglomerados a formar en función de r^2 .

Los centroides de grupo tienen la finalidad de favorecer las comparaciones entre los grupos, así como de apoyar al discernimiento de las características de cada conglomerado. Estos valores y su interpretación se muestran en la tabla 17.

Tabla 17 Centroides por conglomerado e interpretación de los mismos.

Conglomerado	Centroide			Interpretación
	F1	F2	F3	
I	5.1	4.9	2.3	La mejor capacidad de producción de grano entre las variedades evaluadas; variedad intermedia-tardía y vigorosa. Plantas bien adaptadas bajo las condiciones del experimento
II	-2.3	-1.0	-2.0	Baja capacidad de producción de grano; plantas intermedias-precoces, poco vigorosas, con bajo rendimiento.
III	0.1	-0.6	0.7	Capacidad promedio de producción de grano ciclo vegetativo de duración ligeramente inferior al promedio, vigor ligeramente superior al promedio, plantas de doble propósito inferiores a las del grupo IV para producir grano.
IV	1.6	0.3	0.7	Capacidad buena para la producción de grano, variedades intermedias, ligeramente más vigorosas que el promedio. Aptas para la producción de grano y follaje.
V	2.9	3.7	1.4	Muy buena capacidad de producción de grano y follaje, intermedias, vigorosas, superiores a las plantas del grupo IV
VI	-1.2	2.9	0.4	Ligeramente ineficientes para producir granos, menos precoces que el promedio y vigor ligeramente superior al promedio. Uso forrajero, inferior a las plantas del grupo III.
VII	-3.9	-4.9	-2.4	Variedades no bien adaptadas bajo las condiciones del experimento, con pobre capacidad de producción de grano, precoces y poco vigorosas, no muy recomendables, inferiores al grupo II.

5.6 Análisis múltiple discriminante.

El análisis múltiple discriminante se efectuó sobre los siete grupos obtenidos mediante los análisis de factores, y de conglomerados. Fueron seleccionadas aleatoriamente 14 plantas en cada grupo, de manera que la prueba de Box refleje primordialmente la diferencia entre matrices de covarianza por grupos, o aún, fallas en el ajuste de los datos a la distribución normal multivariable. El método aplicado fue la minimización de Λ de Wilks, empleando distancias euclidianas al cuadrado, el cual es un procedimiento *stepwise*.

De las tres variables utilizadas para la discriminación, la primera en ser incluida en el análisis fue la temporalidad (F2), con , equivalente a una F de 40.284, con 6 grados de libertad en el numerador y 91 en el denominador, cuya significación es $\alpha < 0.0000$. La segunda variable discriminante fue la producción de grano (F1), con , equivalente a una F de 24.6331, con 12 grados de libertad en el numerador y 180 en el denominador. Finalmente, fue incorporada la

variable vigor (F3), con , equivalente a una F de 17.3225 con 18 grados de libertad en el numerador y 252 en el denominador, y significación $\alpha < 0.0000$. Esta situación es explicable porque los valores de factor resumen bien la variabilidad de todo el grupo, y fueron los tres factores que significativamente contribuyen a explicar la variación total del grupo.

Tabla 18 Elección de funciones discriminantes canónicas.

Función	Valor propio	% de la varianza	Correlación canónica	Posterior a la Λ de Wilks				
				Función	Λ	χ^2	G.L.	α
1	2.9316	69.96	0.8635	0	0.1027	209.423	18	0.0000
2	1.0505	25.07	0.7158	1	0.4036	83.472	10	0.0000
3	0.2083	4.97	0.4152	2	0.8276	17.408	4	0.0016

La tabla 18 indica los resultados del análisis que permiten seleccionar las funciones discriminantes apropiadas con base en la χ^2 , tomando en consideración la correlación canónica, el porcentaje de la variación explicado por cada función y los valores propios correspondientes. Para cada función, el valor propio es la razón entre las sumas de cuadrados entre grupos y la suma de cuadrados dentro de grupos. Así, el valor propio de la primera función indica una razón de 2.9316 entre las citadas sumas de cuadrados. Además, la correlación canónica es la raíz cuadrada de la relación entre la suma de cuadrados entre grupos y la suma de cuadrados totales; cuando se deja el cuadrado de este cociente, se indica la proporción de la variabilidad total "explicada" por las diferencias entre grupos. Al lado derecho de la tabla, se prueba la hipótesis de que las funciones prueban sólo diferencias de muestreo. El 0 del primer renglón es la prueba cuando ninguna de las tres funciones se ha removido del análisis, y es altamente significativa, indicando así que todas las funciones juntas se apoyan a discriminar bien los casos. El segundo renglón, es la prueba luego de haber removido del análisis la primera función discriminante, y como puede verificarse, las dos funciones restantes continúan explicando significativamente la variabilidad. El tercer renglón muestra que se han removido del análisis las dos primeras funciones discriminantes, y aún resulta significativa la proporción de la varianza explicada por la tercera función discriminante. De esto se desprende que las tres funciones resultan necesarias para buscar la discriminación de casos entre los grupos del análisis.

Las funciones discriminantes canónicas estandarizadas son:

$$Z1 = -0.54133 F1 + 1.33598 F2^* + 0.00794 F3$$

$$Z2 = 1.04750 F1^* - 0.98075 F2 + 0.75195 F3$$

$$Z3 = 0.97374 F1 + 0.17058 F2 - 1.08059 F3^*$$

Los asteriscos arriba indican la carga del coeficiente más importante para cada una de las variables en cada función discriminante. Las funciones discriminantes canónicas no estandarizadas se emplean para determinar los valores de la función discriminante en cada caso, y son:

$$Z1 = -0.2012150 F1 + 0.67598200 F2 + 0.005875218 F3 - 0.44868080$$

$$Z2 = 0.3893577 F1 - 0.49624350 F2 + 0.556629900 F3 + 0.15381460$$

$$Z3 = 0.3619403 F1 + 0.08631126 F2 - 0.799896900 F3 - 0.06491350$$

La variable temporalidad tiene mayor peso en la primera función; la variable producción de grano tiene mayor peso en la segunda función; la variable vigor, lo tiene en la tercera función, coincidiendo con el orden de selección de las variables. La correlación canónica de F2 con la primera función discriminante es de 0.93432; la de F1 con la segunda función discriminante es 0.72425, y la de F3 con la misma segunda función discriminante es 0.69831.

La primera función enfatiza las diferencias entre variedades buenas productoras de grano contra aquellas que predominantemente acumulan materia verde, y ello explicaría el 69.96% de la variación total; definitivamente se da más peso a la acumulación de materia verde ($r^2=0.7456$). En la segunda función, se contrastan la producción de grano y el vigor contra la temporalidad, lo cual explica el 25.07% de la variación total (95.03% acumulado, $r^2=0.5124$). En la tercera función, el contraste principal se da entre las plantas con buena producción de grano pero que no son tan vigorosas, explicando el 4.97% de la variación total ($r^2=0.1724$).

Para probar cuán buena capacidad de discriminación poseen las funciones discriminantes encontradas, se calcula la probabilidad *a priori* de pertenencia a un grupo determinado, equivalente a la probabilidad de discriminar correctamente por azar, y como hubo siete grupos, esa probabilidad es 0.14. Si la función discriminante es buena, la cantidad de casos correctamente clasificados debe superar ampliamente este 14%. Los resultados de la clasificación se muestran en la tabla 19, donde se observa que la proporción de casos correctamente clasificados fue de 69.39%, lo cual se considera satisfactorio.

En dicha tabla puede verse además que el grupo VII, el cual incluye a los híbridos, logró el mejor porcentaje de clasificación debido a la mayor homogeneidad entre estas plantas.

Tabla 19 Resultados de la clasificación mediante funciones discriminantes.

Grupo actual	Casos	Predicción de pertenencia al grupo						
		I	II	III	IV	V	VI	VII
I	14	11 78.6%	0 0.0%	0 0.0%	1 7.1%	1 7.1%	1 7.1%	0 0.0%
II	14	1 7.1%	12 85.7%	0 0.0%	0 0.0%	0 0.0%	1 7.1%	0 0.0%
III	14	0 0.0%	0 0.0%	9 64.3%	2 14.3%	2 14.3%	0 0.0%	1 7.1%
IV	14	2 14.3%	1 7.1%	5 35.7%	6 42.9%	0 0.0%	0 0.0%	0 0.0%
V	14	4 28.6%	0 0.0%	1 7.1%	1 7.1%	5 35.7%	3 21.4%	0 0.0%
VI	14	0 0.0%	1 7.1%	0 0.0%	0 0.0%	1 7.1%	12 85.7%	0 0.0%
VII	14	0 0.0%	1 7.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	13 92.9%

5.7 Análisis multivariable de la varianza.

El análisis multivariable de la varianza MANOVA se efectúa en SPSS mediante el ingreso de las variables en un orden determinado, para desarrollar el proceso llamado *stepdown* el cual consiste en ir relacionando en forma secuencial cada nueva variable incorporada al modelo con las que le precedieron, pero si el orden de ingreso de estas variables no tiene significación, el proceso *stepdown* no aporta nada específico. El MANOVA se desarrolló tomando 11 variables dependientes y ninguna covariable, en el orden de inclusión siguiente: peso de grano, altura de mazorca, altura de planta, diámetro de mazorca,

diámetro de tallo, longitud de mazorca, número de carreras, número de hojas, porcentaje de humedad, peso de mazorca y profundidad de grano. Como se mencionó en materiales y métodos, el diseño fue de bloques completos al azar con tres repeticiones y treinta tratamientos (variedades). Para facilitar el proceso de la información se redujo la matriz de datos obteniendo el valor promedio de cada variable por variedad en cada repetición, y trabajando entonces una matriz de orden 90 por 11 como base para obtener las matrices de correlaciones y de varianzas y covarianzas.

La tabla 20 muestra los valores propios de la matriz de correlaciones para las sumas de cuadrados entre individuos + residual. En esta tabla puede verse que hubo cuatro funciones con valor propio mayor a 1, las cuales explicarían en conjunto un 64.837% de la variación total.

Tabla 20 Valores propios de la matriz de correlaciones para sumas de cuadrados ENTRE INDIVIDUOS + RESIDUAL.

Componente	Valor propio	% de la varianza	% acumulado
1	3.249	29.532	29.532
2	1.640	14.912	44.444
3	1.169	10.630	55.074
4	1.074	9.763	64.837
5	0.898	8.165	73.002
6	0.870	7.908	80.910
7	0.752	6.839	87.749
8	0.411	3.732	91.481
9	0.376	3.419	94.901
10	0.352	3.197	98.098
11	0.209	1.902	100.000

El primer procedimiento del MANOVA consistió en obtener la matriz de correlaciones, extraer los componentes principales de esta matriz como primera aproximación al análisis de factores y aplicar posteriormente la rotación varimax. Una vez obtenidas las cargas de factor, se procedió al cálculo de las funciones canónicas, con las cuales se estimaron las nuevas medias.

La hipótesis nula es el planteamiento de igualdad de todos los vectores de medias, en tanto que la alternativa es que por lo menos un vector de medias difiere del resto.

Las matrices requeridas para efectuar la prueba se denominan **H** (hipotética para cualquier suma de cuadrados y productos cruzados) y **E** (matriz de errores). El MANOVA es una prueba que involucra a los valores propios de cualquiera de estas o de ambas matrices. Las matrices empleadas por SPSS se muestran en las tablas 21 y 22.

Tabla 21 Suma de cuadrados y productos cruzados de la matriz ENTRE INDIVIDUOS + RESIDUAL.

	PESGR	ALTMZ	ALTPL	DIAMAZ	DMTLL	LONGMAZ
PESGR	59918.765					
ALTMZ	-320.436	9115.350				
ALTPL	4960.465	4527.811	8999.516			
DIAMAZ	165.322	35.279	72.817	6.756		
DMTLL	57.400	6.901	10.753	0.037	3.403	
LONGMAZ	555.800	8.820	309.252	16.761	-2.767	153.047
NO_CARR	300.638	133.462	109.197	4.519	0.709	13.255
NO_HOJ	60.527	358.486	257.170	4.500	8.140	18.948
P_HUM	2.520	1.141	0.645	0.072	-0.032	0.265
PESOMAZ	32822.320	7154.491	12848.397	530.519	-39.561	2305.161
PROFG	9.839	16.460	15.549	1.250	0.281	2.860
	NO_CARR	NO_HOJ	P_HUM	PESOMAZ	PROFG	
NO_CARR	48.225					
NO_HOJ	14.480	175.591				
P_HUM	0.002	0.039	0.006			
PESOMAZ	339.240	125.470	10.073	93623.848		
PROFG	0.414	-1.169	0.030	117.415	1.324	

La matriz de la tabla 21 muestra la cuantía de la variabilidad en las variables dependientes sin haber particionado la matriz de varianzas y covarianzas en grupos significativos. La diagonal de esta matriz es (N-1) veces la varianza de las variables dependientes, y los elementos fuera de la diagonal son los productos cruzados de todas las variables agrupadas en pares $\sum(x-\bar{x})(y-\bar{y})$. Esta matriz representa los términos del error. Por otra parte, en la matriz de la tabla 22 puede observarse que los elementos situados dentro y fuera de la diagonal principal tienen un valor mayor que los de la matriz en la tabla 21, anticipando así un resultado prometedor en el MANOVA.

CUCBA



BIBLIOTECA CENTRAL

Tabla 22 Matriz ajustada hipotética de cuadrados y productos cruzados para tratamientos.

	PESOG	ALTMZ	ALTPL	DIAMAZ	DMTLL	LONGMAZ
PESOG	236366.034					
ALTMZ	54920.863	136215.411				
ALTPL	61817.366	143977.410	162108.549			
DIAMAZ	1909.154	-113.555	56.012	27.422		
DMTLL	353.468	385.810	403.288	2.295	3.221	
LONGMAZ	5167.338	4994.463	6095.478	24.340	22.804	510.809
NO_CARR	1130.783	153.220	718.491	34.606	6.100	-28.893
NO_HOJ	2842.174	2977.047	3772.091	19.265	18.619	242.330
P_HUM	29.803	29.463	36.756	0.160	0.093	2.058
PESOMAZ	224677.448	59831.096	76299.805	1983.779	437.513	6437.504
PROFG	476.075	125.279	171.399	5.483	0.510	9.825
	NO_CARR	NO_HOJ	P_HUM	PESOMAZ	PROFG	
NO_CARR	254.663					
NO_HOJ	81.235	247.587				
P_HUM	0.414	1.464	0.020			
PESOMAZ	1327.702	3899.683	38.877	243346.362		
PROFG	3.158	7.063	0.093	512.800	1.801	

Una vez que se han identificado las matrices básicas para el estudio, se presentan a continuación las pruebas multivariable de significación de los tratamientos en la tabla 23.

Tabla 23 Pruebas de significación multivariable¹⁶.

Nombre de la prueba	Valor	F aproximada	GL para la hipótesis	GL para el error	Significación de F
Pillai	6.02510	2.4220	319.00	638.00	0.000
Hotelling	55.03923	7.96806	319.00	508.00	0.000
Wilks	0.00000	3.92896	319.00	527.22	0.000
Roy	0.97100				
Nombre de la prueba	Efecto del tamaño	Noncent.	Poder		
Pillai	0.548	772.683	1.00		
Hotelling	0.833	2541.812	1.00		
Wilks	0.680	1122.757	1.00		

A través de los criterios de prueba de Pillai, Hotelling y Wilks mostrados en la tabla 23, se rechaza la hipótesis nula, concluyéndose que al menos un vector de medias difiere de los demás. La columna llamada efecto del tamaño da un valor de eta cuadrada (η^2) parcial entre cada variable y el factor estudiado. Para nuestro caso, es una medida de la capacidad de ajuste de todas las variables en tanto son explicadas por las variedades como factor. El efecto

¹⁶ $S=11$, $M=8\frac{1}{2}$, $N=23$. Estos son los parámetros de la distribución, donde S es el número de valores propios en la matriz de varianzas y covarianzas, mientras $M = (|g-u|-1)/2$, y $N = (N-r-u-1)/2$.

del factor muestra que el ajuste de la varianza de tratamientos entre la varianza total está en 54.8% en la prueba de Pillai, 83.3% en la de Hotelling y 68.0% en la de Wilks, coeficientes cuyos valores resultan muy adecuados para reforzar las conclusiones ya referidas.

Un MANOVA significativo, como ha sido el de este caso, permite efectuar ahora las pruebas de una variable, las cuales tuvieron 29 grados de libertad para tratamientos y 58 para el error. La ventaja de estas pruebas *a posteriori* del MANOVA es que hemos evitado los falsos positivos, es decir, tenemos una gran confianza en no transgredir las tasas para el error de tipo I declarando como significativas diferencias que son realmente aleatorias. Sin un MANOVA significativo, el nivel de significación de las pruebas sería realmente de $(1-\alpha)^p$, donde p es el número de pruebas independientes que se realicen, por lo cual si se pretendiera una confianza de 95% con 11 variables, se lograría apenas 56.88%.

En la tabla 24 se muestran las pruebas de una variable, las cuales tuvieron 29 grados de libertad para tratamientos y 58 para el error experimental. Las variables peso de grano, altura de mazorca, altura de planta, diámetro de mazorca, longitud de mazorca, número de carreras, número de hojas, porcentaje de humedad, peso de mazorcas y profundidad de grano tuvieron diferencias altamente significativas ($\alpha < 0.01$), mientras en diámetro de tallo $\alpha < 0.05$.

Tabla 24 Pruebas F de significación para una sola variable.

Variable	SC hipotética	SC del error	CM hipotético	CM del error	Valor de F	α	η^2 ¹⁷	Valor no centralizado de β	Poder de la prueba
PESOG	236366.03	59918.77	8150.55	1033.08	7.88955	0.000	0.7978	228.797	1.0000
ALTMZ	136215.41	9115.35	4697.08	157.16	29.88704	0.000	0.9373	866.724	1.0000
ALTPL	162108.55	8999.52	5589.95	155.16	36.02606	0.000	0.9474	1044.756	1.0000
DIAMAZ	27.42	6.76	0.95	0.12	8.11849	0.000	0.8023	235.436	1.0000
DMITLL	3.22	3.40	0.11	0.06	1.89281	0.000	0.4862	54.892	0.9759
LONGMAZ	510.81	153.05	17.61	2.64	6.67519	0.000	0.7695	193.580	1.0000
NO_CARR	254.66	48.23	8.78	0.83	10.56145	0.000	0.8408	306.282	1.0000
NO_HOJ	247.59	175.59	8.53	3.03	2.82004	0.000	0.5851	81.781	0.9992
P_HUM	0.02	0.01	6.8 E -4	1 E -4	7.00357	0.000	0.7779	203.104	1.0000
PESOMAZ	24346.36	93623.85	8391.25	1614.20	5.19838	0.000	0.7222	150.753	1.0000
PROFG	1.80	1.32	6.21 E -2	0.23	2.72160	0.001	0.5764	78.926	0.9988

¹⁷ La eta cuadrada es la proporción de la variación de la variable dependiente que puede ser explicada por la variación de la independiente --en este caso, las variedades--.

La mejor η^2 fue para la variable altura de planta, con valor de 0.9474, indicando con ello que 94.74% de la variación en la altura de planta se relaciona con las variedades. La η^2 más baja la tuvo la variable diámetro de tallo, cuyo valor fue 0.4862, valor cuyo significado es que 48.62% de la variación en diámetro de tallo se relaciona con las variedades. La tabla muestra también el valor de β no centralizado, mismo que es una función probabilística para determinar el poder de la prueba. La tabla muestra que dicho poder superó 95%, lo cual es muy satisfactorio para fines prácticos. Si hubiéramos tenido la intención de desarrollar un análisis *stepdown*, un criterio para la inclusión de las variables consistiría en su agregación secuencial al modelo en orden descendente de valor de F.

Para reducir las dimensiones de la matriz de varianzas y covarianzas, se emplea la información contenida en las tablas 25 y 26. Primero en la tabla 25, los valores propios y correlaciones canónicas de las raíces obtenidas después de la rotación varimax permiten determinar la significación de cada una de las funciones canónicas. La primera función explicaría 60.845% de la variación total, la segunda 12.585%, la tercera 11.305%, la cuarta 5.834% y la quinta 2.977% de la misma variación total, porcentajes que acumulan 93.546% de variación explicada en la matriz de varianzas y covarianzas.

Tabla 25 Valores propios y correlaciones canónicas para el efecto de los tratamientos.

Raíz número	Valor propio	Porcentaje	Porcentaje acumulado	Correlación canónica
1	33.489	60.845	60.845	0.985
2	6.927	12.585	73.430	0.935
3	6.222	11.305	84.736	0.928
4	3.211	5.834	90.569	0.873
5	1.638	2.977	93.546	0.788
6	1.485	2.698	96.245	0.773
7	0.767	1.394	97.638	0.659
8	0.636	1.155	98.794	0.623
9	0.400	0.726	99.520	0.534
10	0.187	0.341	99.860	0.397
11	0.077	0.140	100.000	0.267

La reducción de la dimensionalidad del modelo se efectúa mediante el análisis de la significación de la Λ de Wilks, medida por su aproximación a la F. En la tabla 26 se observa que las raíces 1 a 5 son significativas, por lo cual pueden emplearse las primeras cinco funciones canónicas, en lugar de las 11 variables.



Tabla 26 Análisis de reducción de dimensiones.

Raíces	Λ de Wilks	F	GL numerador	GL error	α
1 a 11	0.00000	3.92896	319	527.22	0.000
2 a 11	0.00012	2.78742	280	489.02	0.000
3 a 11	0.00097	2.30685	243	449.03	0.000
4 a 11	0.00700	1.78009	208	407.18	0.000
5 a 11	0.02948	1.41943	175	363.40	0.003
6 a 11	0.07777	1.20893	144	317.64	0.086
7 a 11	0.19328	0.93351	115	269.84	0.660
8 a 11	0.34155	0.78025	88	219.96	0.909
9 a 11	0.55872	0.57420	63	168	0.994
10 a 11	0.78200	0.37286	40	114	1.000
11	0.92858	0.23479	19	58	1.000

Un procedimiento para efectuar la prueba de separación de medias multivariable, consiste en calcular los valores estandarizados de las variables en la prueba, y luego multiplicar estos valores por las funciones canónicas estandarizadas, logrando así un valor único por variedad y quedando en posibilidad de aplicar cualquier prueba estándar de separación de medias. La tabla 27 muestra los coeficientes de las primeras cinco funciones discriminantes estandarizadas.

Tabla 27 Coeficientes de las funciones discriminantes estandarizadas.

Variable	Número de función				
	1	2	3	4	5
PESGR	0.150	0.281	-0.689	0.177	-0.008
ALTMZ	0.575	0.393	-0.252	-0.412	-0.332
ALTPL	0.669	-0.436	0.075	-0.103	0.403
DIAMAZ	-0.209	-0.119	-0.751	-0.412	-0.172
DMTLL	0.223	-0.059	0.108	0.185	-0.493
LONGMAZ	0.547	0.394	0.257	0.459	-0.795
NO_CARR	-0.162	-0.925	0.120	-0.018	-0.281
NO_HOJ	-0.280	-0.012	-0.028	0.405	0.211
P_HUM	0.354	-0.410	0.242	0.669	0.407
PESOMAZ	-0.698	0.096	0.087	0.159	0.160
PROFG	-0.111	0.090	-0.202	-0.001	0.438

El análisis de las correlaciones entre las variables dependientes y las variables canónicas posterior a la rotación varimax se muestra en la tabla 28, y fue así: la primera función canónica tiene correlación de 0.939 con la variable longitud de mazorca, y esa fue la mayor correlación para la variable canónica, la segunda tiene una correlación de 0.975 con peso de grano, la tercera 0.949 con el porcentaje de humedad de grano, la cuarta, 0.886 con

diámetro de mazorca y la quinta 0.956 con profundidad de grano, siendo éstas las variables que mayores cambios aportan en la estructura multidimensional.

Tabla 28 Correlación entre las variables dependientes y canónicas posterior a la rotación varimax.

Variable	Función canónica				
	1	2	3	4	5
PESOGP	0.067	0.975	0.058	0.093	-0.008
ALTMZ	-0.030	-0.030	0.074	0.035	0.066
ALTPL	0.119	0.103	0.010	0.104	0.050
DIAMAZ	0.256	0.116	0.160	0.886	0.212
DMITLL	-0.066	0.072	-0.124	0.001	0.086
LONGMAZ	0.939	0.068	0.112	0.213	0.074
NO_CARR	0.059	0.080	-0.015	0.094	0.014
NO_HOJ	0.056	-0.007	0.027	0.048	-0.063
P_HUM	0.115	0.064	0.949	0.137	0.168
PESOMAZ	0.366	0.276	0.225	0.351	0.148
PROFG	0.076	-0.007	0.164	0.171	0.956

La separación de medias para variables canónicas estandarizadas se efectuó con la primera función aplicando la prueba de Bonferroni. La clasificación de grupos es diferente a la que se consiguió con el análisis de conglomerados porque en el MANOVA se partió de la matriz de varianzas y covarianzas y en el AC de la matriz de correlaciones para generar las funciones. En el MANOVA se trabajó con las variables estandarizadas, así que en lugar de cuadrado medio del error se empleó η^2 de la prueba de Wilks (efecto del tamaño), que de acuerdo con la tabla 23 tiene magnitud 0.680, con 527.22 grados de libertad para el error en la primera función discriminante. Los grupos se muestran en la tabla 29.



Tabla 29 Prueba de Bonferroni para separar promedios con base en la primera función discriminante canónica.

Variedad	Valor de la función F1	Grupo
20 Ancho Negro	1.714	a
10 Tuxpeño Caribe	1.547	ab
9 Blanco Cristalino	1.522	ab
21 Turicato Negro	1.236	abc
8 Blanco Subtropical	0.982	abcd
11 Tuxpeño	0.787	bcde
22 Turicato Negro	0.661	cdef
15 Tesistán	0.613	cdef
1 Trejos	0.327	defg
24 Tirándaro Michoacán Blanco	0.323	defg
23 Tirándaro Michoacán Amarillo	0.290	defg
4 Santa Ana Tepetitlán	0.249	defg
27 Purísima	0.221	defg
17 Cuquío Blanco	0.174	efg
16 Güino Negro	0.157	efg
25 Valle Altos Negro	0.109	efgh
28 Olote Colorado	0.070	efgh
18 Güino Blanco	0.056	efgh
26 Valles Altos Blanco	-0.120	fghi
12 Tequila	-0.255	ghij
3 Población CIMMYT	-0.277	ghij
13 La Mora	-0.371	ghijk
19 Gordo Blanco	-0.666	hijk
14 Ubalano	-0.668	hijk
2 Zapalote chico	-0.789	ijk
5 MGG1E	-1.000	jkl
29 D 877	-1.076	kl
7 Amarillo Oaxaca	-1.108	kl
6 Jaleño	-1.597	lm
30 D 880	-2.282	m

CUCBA



BIBLIOTECA CENTRAL

CUCBA



BIBLIOTECA CENTRAL

6 Conclusiones.

1. Las técnicas de análisis multivariable de datos nos han permitido encontrar dimensiones latentes en los datos, las cuales no son evidentes mediante técnicas estadísticas univariable.

2. El 83.6364% de las correlaciones entre variables fueron significativas, lo cual indica que los análisis univariable de varianza y otros tenderían a sobreestimar la tasa de error de tipo I, y por tanto, las técnicas multivariable son las indicadas para el análisis exploratorio de datos y determinar las diferencias entre las variedades.

3. El análisis de componentes principales, nos mostró que el número de dimensiones latentes en el conjunto de 11 variables puede reducirse a 3, enfatizando la varianza total de las variables y partiendo de la matriz de correlaciones entre las variables. Esas tres características pueden denominarse de acuerdo con las variables que mayor influencia presentan de acuerdo con las correlaciones entre componente principal y variable. El primer componente indica que todas las variables tienen influencia positiva, a excepción de número de carreras. Parece, por tanto, que las plantas con mayor crecimiento, tienden también a producir más grano, y eso explicaría el 39.37564% de la variación total; El segundo componente tiene correlaciones arriba de 0.5 con altura de planta y mazorca y diámetro de mazorca. Los signos diferentes entre las variables, indican que las plantas con tendencia a ser altas, tienen mazorcas más delgadas. La variable diámetro de mazorca tuvo signo negativo, así que altos valores en esta variable serían indicadores de plantas altas con mazorcas delgadas, y bajos valores, plantas más bajas con mazorcas de mayor diámetro. Esta situación explicaría el 17.19973% de la variación total. Finalmente, el último componente sería equivalente a la variable número de carreras: A mayor número de carreras, mayor carga en el componente tercero, y viceversa. Esta componente explica el 9.89718% de la variación total.

4. El análisis de factores es más sencillo de interpretar en este caso que el de componentes principales. Además, el análisis de factores se orienta a la covarianza a diferencia del de componentes principales, que lo hace a la varianza. Cada factor se obtuvo

mediante la rotación varimax, cuya finalidad es lograr que cada variable tenga mayor carga en un factor y casi cero en los restantes. En el primer factor, las variables con mayor carga fueron diámetro de mazorca, peso de mazorca, profundidad de grano y peso de grano. El factor más importante para discriminar las plantas de maíz será la producción de grano; de las cuatro variables incorporadas al análisis, la menos eficiente en ser explicada por el factor fue profundidad de grano. En el segundo factor, las variables con mayor valor fueron altura de planta, altura de mazorca, longitud de mazorca y porcentaje de humedad. Este factor es independiente del primero, así que agrupa las plantas principalmente por su tamaño, y ese aspecto está relacionado con su longevidad o precocidad. Las variables con más pobre contribución al factor fueron porcentaje de humedad y longitud de mazorca, por lo cual las plantas con poco valor en el factor 1 y alto en 2, serán buenas forrajeras. En el tercer factor, independiente de los otros dos, las variables más importantes fueron diámetro de tallo, número de hojas y número de carreras, todas con aproximadamente la misma comunalidad. Este factor se designó vigor. El análisis de factores, permitió separar las variedades en productoras de grano, de forraje, de doble propósito e inadaptadas. Adicionalmente, los valores de factor se graficaron para identificar conglomerados.

5. El análisis de conglomerados se efectuó con los valores de factor y permitió la conformación de 7 grupos afines. La determinación de siete conglomerados se determinó empleando dos pruebas, la r^2 y RMSSTD. Hubo buena coincidencia entre los siete grupos tentativos obtenidos mediante componentes principales, análisis de factores y análisis de conglomerados. La visualización del dendrograma permite identificar tanto los siete grupos como las variedades constituyentes de cada uno de los grupos, mientras los centroides de grupo se interpretan como el nivel de adaptación o la vocación productiva que las diferentes variedades mostraron durante este experimento. El análisis permitió identificar los grupos de variedades desde el punto de vista morfológico y productivo.

6. Mediante el análisis múltiple discriminante, se trató de clasificar y distinguir correctamente entre plantas individuales pertenecientes a los siete grupos de conglomerados obtenidos mediante el análisis correspondiente, empleando con este fin los valores de factor para las variedades. De cada uno de los grupos se sortearon aleatoriamente 14 plantas, por ser esta la cantidad de individuos del grupo 1 (constituido únicamente por la variedad Trejos), de

manera que se pudiera clasificar plantas en grupos equiprobables. Se determinaron tres funciones ortogonales, que son las necesarias para la discriminación en conformidad con la prueba de χ^2 . La prueba permitió identificar correctamente la pertenencia al grupo de 69.39% de las plantas. El examen de las fallas en la clasificación individual, es útil en la identificación de plantas fuera de tipo, el cual sería determinado mediante los centroides de grupo del análisis de conglomerados.

7. El análisis multivariable de varianza nos llevó a rechazar la hipótesis nula de que no hay diferencias entre los vectores de medias, con un 99% de confianza. Así, se pudo demostrar que las diferencias entre grupos son reales, aunque los grupos no coinciden con los formados mediante los análisis anteriores, principalmente porque el MANOVA se realizó a partir de la matriz de varianzas y covarianzas en lugar de la matriz de correlaciones empleada en los análisis anteriores.

8. Los tres criterios de prueba multivariable para el MANOVA, la prueba de Pillai, la de Hotelling y la de Wilks mostraron que hay diferencias significativas entre las variedades considerando los vectores de medias en lugar de valores individuales de una variable. Una vez definido que los vectores difieren, se siguió con las pruebas de F univariadas, para un diseño en bloques completos al azar con 3 repeticiones para los treinta tratamientos.

9. Todas las variables mostraron diferencias altamente significativas, a excepción de diámetro de tallo que sólo fue significativo al 95% de confianza.

10. El estadístico η^2 se emplea para medir la proporción de la variación de la variable dependiente en función de los cambios de la variable independiente (variedades) tomando en consideración al conjunto de variables como un todo, por lo que el enfoque de la prueba es holístico. Las variables con más alto puntaje para el citado estadístico fueron altura de planta y altura de mazorca, y las de más pobre puntaje la profundidad de grano y el diámetro de tallo.

11. Al efectuar la reducción de las dimensiones de la matriz, se admite que sólo cinco de las 11 dimensiones en la matriz de varianzas y covarianzas son significativas, por lo cual se puede lograr una buena aplicación de los criterios de prueba con sólo cinco variables

canónicas, las cuales se obtuvieron mediante el análisis de factores y la rotación varimax de la matriz de varianzas y covarianzas (a diferencia del análisis de factores y el de componentes principales, que se lograron a partir de la matriz de correlaciones).

12. Para la prueba de separación de medias, se eligió el criterio de Bonferroni, estandarizando los valores de las variables en la matriz de datos y aplicando luego a cada vector el producto por la función discriminante primera, que se enuncia a continuación:

$$\begin{aligned} Z_1 = & 0.150 \text{ ZPESOGR} + 0.575 \text{ ZALTMZ} + 0.669 \text{ ZALTPL} - 0.209 \text{ ZDIAMAZ} + \dots \\ & \dots + 0.223 \text{ ZDMITLL} + 0.547 \text{ ZLONGMAZ} - 0.162 \text{ ZNO_CARR} - 0.280 \text{ ZNO_HOJ} + \dots \\ & \dots + 0.354 \text{ ZP_HUM} - 0.698 \text{ ZPESOMAZ} - 0.111 \text{ ZPROFG} \end{aligned}$$

13. La primera función discriminante, enunciada arriba, correlaciona mejor con la variable longitud de la mazorca. Por lo que se refiere a esta variable, estimada tomando en cuenta a todas las demás, la tabla 29 mostró la conformación de los grupos, lo cual explicaría el 60.845% de la variación total.

7 Literatura Citada.

- 1- **ANDERSON T.W.** 1984. An introduction to multivariate statistical analysis. 2a. ed. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons. Estados Unidos de América. 675 págs.
- 2- **BALLESTER P.** 1995. *Cluster analysis*. [Http://www.eso.org/research/data-man/data-proc/systems/esomidias/doc/94NOV/vol2/node1.html](http://www.eso.org/research/data-man/data-proc/systems/esomidias/doc/94NOV/vol2/node1.html)
- 3- **BERNHARDT K.J., Allen J.C. y Helmers G.A.** 1996. Using *cluster* analysis to classify farms for conventional/alternative systems research. *Review of agricultural economics*. 18: 4, 599-611;
- 4- **BUOL S. W., Hole F. D. McCracken R. J.** 1981. Génesis y clasificación de suelos. (tr. inglés Agustín Contin). Ed. Trillas. México. 417 págs.
- 5- **CARROL J. D., Green P. E. y A. Chaturvedi.** 1997. *Mathematical tools for applied multivariate analysis*. 2a. ed. Academic Press. Estados Unidos de América. 376 págs.
- 6- **CETENAL** 1973. Carta topográfica "San Francisco Tesistán" F-13 D-55. Secretaría de la Presidencia. Aguascalientes, México.
- 7- _____ 1973. Carta edafológica "San Francisco Tesistán" F-13 D-55. Secretaría de la Presidencia. Aguascalientes, México.
- 8- **COOK D.** 1997. *Reducing Dimensionality. Principal Component Analysis*. Dept. of Statistics. Snedecor Hall. Ames, Iowa.
<http://www.public.iastate.edu/~dicook/stat501/97/lectures/2.6.html>
- 9- **CUADRAS C.M.** 1991. *Métodos de análisis multivariante*. 2a. ed. Colección "Estadística y análisis de datos". Promociones y publicaciones universitarias, S. A. Barcelona, España. 644 págs.
- 10- **DILLON W. R. y M. Goldstein.** 1984. *Multivariate Analysis. Methods and applications*. John Wiley & Sons. Estados Unidos de América. 587 pp.
- 11- **DOEYGARD S.** 1995. *Principal component analysis of the morphological characters in the genus Pachyrhizus DC.* Thesis.
<http://www.dsr.kvl.dk/~sodos/bachelor.hTM>.
- 12- **ERITJA R.** 1996. Wing biometry and statistical discriminant analysis as a technique to determine sex of a *Culex pipiens* (Diptera, Culicidae) gynandromorph. *Journal of Economic Entomology*, 89: 5, 1338-1341.

- 13- **FITZPATRICK E. A.** 1984. Suelos, su formación, clasificación y distribución. (Tr. inglés Antonio Merino Ambrosio). Compañía Editorial Continental, S.A. México. 430 págs.
- 14- **FLURY B. y Riedwyl H.** 1988. Multivariate statistics. A practical approach. Chapman and Hall. University press, Cambridge. Gran Bretaña. 290 págs.
- 15- **FLYNN T.C.** 1996. Multivariate statistics: *Cluster analysis*. Universidad de Cornea. <http://trochim.human.cornell.edu/tutorial/flynn/cluster.htm>
- 16- **FORVEILLE L., Vercauteren J. y Rutledge D.N.** 1996. Multivariate statistical analysis of two dimensional NMR data to differentiate grapevine cultivars and clones. *Food Chemistry*. 57: 3, 441-450.
- 17- **HAIR Joseph F, Anderson Rolph E., Tatham Ronald L. y William C. Black.** 1992. Multivariate data analysis whit readings. 3a. ed. McMillan Publishing Co. Estados Unidos de América. 544 págs.
- 18- **JOHNSON R.A. y Wichern D. W.** 1992. Applied multivariate statistical analysis. 3a. edición. Prentice Hall, Inc. Estados Unidos de América. 642 págs.
- 19- **LAFITTE H. R., Edmeades G.O. y Taba S.** 1997. Adaptative strategies identified among tropical maize landraces for nitrogen-limited environments. *Field crop research*. 49: 2-3, 187-204.
- 20- **MALLARINO A.P., Hinz P.N., Oyarzabal E.S., Robert P.C. (ed.) Rust R.H. (ed) y Larson W.E.** 1996. Multivariate analysis as a tool for interpreting relationships between site variables and crop yields. Precision agriculture. Proceedings of the 3rd. International Conference, Minneapolis Minnesota, junio 23-26 de 1996. Depto. de Agronomía. Universidad Estatal de Iowa. Ames, Iowa, Estados Unidos de América. Págs. 151-158.
- 21- **MANLY B. F. J.** 1994. Multivariate statistical methods. A primer. 2a. Ed. Chapman and Hall, Gran Bretaña. 215 págs.
- 22- **MONTGOMERY D.C. y Peck E.A.** 1992. Introduction to linear regression analysis. 2a. ed. John Wiley & Sons. Estados Unidos de América. 527 págs.
- 23- **MORRISON D. F.** 1990. Multivariate statistical methods. 3rd. ed. McGraw Hill International Editions. Singapur. 495 págs.
- 24- **NORUSIS, MARIJA J. S/F.** SPSS/PC+ Advanced Statistics V2.0, for the IBM PC/XT/AT and PS/2. SPSS Inc. Chicago Illinois, Estados Unidos de América. Pág. B-110.

- 25- **PEREZ García A. y Penilla González C.** 1988. Los suelos de la República Mexicana, según la taxonomía de suelos USDA. Tesis profesional. Facultad de agricultura. Universidad de Guadalajara. Guadalajara, Jalisco, México. Pág. 85.
- 26- **PRESIDENCIA DE LA REPÚBLICA.** 1994. Sexto informe de gobierno. Estados Unidos Mexicanos. Anexo estadístico.
<http://www.inegi.gob.mx/homeing/estadistica/economia/biosa/bio-4.html>.
- 27- **RADLOFF S.E.** 1996. Multivariate analysis of selected honeybee populations in África. Tesis. Rhodes University, República de Sudáfrica. 282 págs.
- 28- **RIBEIRO A. B. y Nielsen A. A.** 1997. An application of discriminant analysis to pattern recognition of selected contaminated soil features in thin sections. *Geoderma*. 76: 3-4, 253-262;
- 29- **SAGAR.** 1996. Anuario estadístico de la producción agrícola de los Estados Unidos Mexicanos. Centro de Estadística Agropecuaria.
<http://www.inegi.gob.mx/homeing/estadistica/economia/biosa/bio-4.html>.
- 30- **SHARMA Subhash.** 1996. Applied multivariate techniques. John Wiley & Sons, Inc. Estados Unidos de América. 493 págs.
- 31- **SHAW P.E., Moshonas M.G., Buslig B.S., Barros Santiago M. y Wimer Wilbur W.** 1997. Principal component analysis to classify commercial orange juices based on relative amounts of volatile juice constituents. TEKTRAN. Depto. de agricultura de los Estados Unidos de América. Biblioteca Nacional de Agricultura. Centro de información y transferencia de tecnología.
<http://www.nalusda.gov/ttic/tektran/data/000008/19/0000081949.html>.
- 32- **SILVA Machorro Carlos.** 1978. Unidades del suelo interpretadas para su uso en ingeniería civil y aprovechadas por el campesino en usos agropecuarios. CECSA. México. 63 págs.
- 33- **SMITH L.** 1997. MANOVA: testing multiple outcome measures. School of Psychology. Curtin University.
<http://psych.curtin.edu.au/studinfo/honours/outvars.htm>
- 34- **USTIMENKO G.V.** 1980. El cultivo de plantas tropicales y subtropicales. (Tr. ruso, Ramiro Rincón Z. y Francisco Vargas S.). Editorial MIR. Moscú, URSS. 430 págs.
- 35- **VILLALPANDO F. y García Enriqueta.** 1993. Agroclimatología del estado de Jalisco. Universidad de Guadalajara. Centro Universitario de Ciencias Biológicas y Agropecuarias. Lab. Bosque la Primavera. 80 págs.

CUCBA



BIBLIOTECA CENTRAL

- 36- **WULDER, M.** 1995. A practical guide to use of selected multivariate statistics. Waterloo Laboratory for Earth Observations. University of Waterloo. Waterloo, Ontario, Canadá.
<http://glenn.uwaterloo.ca/~mwulder/gg616>.
- 37- **ZHENG-Y.H. y Wang-H.** 1995. Analyses of the feasibility of using genetic distances to predict the heterosis in maize (*Z. mays* L.). (sic). Acta Agriculturae Borealei Sinica. 10: Supplement, 15-21;

BASE DE DATOS DEFINITIVA

DATO	TRAT	PL	REP	ALTPL	ALTMZ	DMITLL	NO_HOJ	LONGMAZ	DIAMAZ	PESOMAZ	PROFG	NO_CARR	PESOGRA	P_HUM
1	1	1	1	320	215	2.70	17	18.25	4.785	419.3	1.275	14	319.3	18.6%
2	1	1	2	255	184	1.76	15	10.1	4.20	74.7	0.93	14	60.1	14.4%
3	1	2	1	280	200	2.50	15	17.15	4.82	383.1	1.565	14	293.0	19.0%
4	1	2	2	320	212	2.54	16	21.4	5.12	330.2	1.21	14	254.8	17.3%
5	1	2	3	290	190	1.73	10	19.2	4.71	220.8	1.07	14	273.9	17.0%
6	1	3	1	270	190	2.80	15	13.5	5.04	331.3	1.46	15	240.9	16.3%
7	1	3	2	300	215	2.90	15	20.1	4.79	238.6	1.12	16	159.8	18.3%
8	1	3	3	260	170	1.63	10	19.3	4.59	206.8	1.32	10	200.4	16.4%
9	1	4	1	290	220	2.10	15	20.5	4.63	249.6	1.72	14	197.9	14.8%
10	1	4	2	305	212	1.98	15	16.05	4.65	388.5	1.01	13	346.0	17.1%
11	1	4	3	260	170	1.64	16	21.0	5.08	289.5	1.33	12	272.0	19.0%
12	1	5	1	305	194	2.30	15	19.0	4.69	240.1	1.43	14	188.5	16.1%
13	1	5	2	292	198	1.98	17	14.15	4.46	293.8	1.29	13	182.6	19.3%
14	1	5	3	230	200	1.75	14	15.3	3.99	69.9	0.84	14	185.6	15.4%
15	2	1	1	265	165	1.57	12	12.0	3.29	57.4	0.89	8	49.7	14.7%
16	2	1	2	220	140	1.46	7	14.9	3.42	68.3	0.81	12	50.2	11.5%
17	2	1	3	200	160	1.57	8	14.0	3.03	58.5	0.88	10	46.1	13.6%
18	2	2	1	208	142	1.40	12	15.0	2.68	37.7	0.88	12	30.4	15.3%
19	2	2	2	230	130	1.39	8	14.1	3.21	66.2	0.83	10	54.2	12.5%
20	2	2	3	210	120	1.32	7	10.5	3.69	48.9	1.00	12	37.7	14.8%
21	2	3	1	206	149	1.27	11	17.0	3.28	84.9	1.10	10	72.7	15.4%
22	2	3	2	220	140	1.46	8	14.2	2.985	85.4	0.76	8	64.3	12.5%
23	2	3	3	240	145	1.49	9	15.4	3.38	72.1	1.10	10	62.5	13.1%
24	2	4	1	235	125	1.48	10	10.0	3.30	68.0	1.18	12	59.6	14.5%
25	2	4	2	180	110	1.24	9	13.7	3.19	57.2	0.69	10	43.1	12.0%
26	2	4	3	120	120	1.33	8	15.6	3.60	75.4	1.19	10	66.3	13.2%
27	2	5	1	208	149	1.26	13	13.5	3.59	66.3	1.15	10	58.5	14.9%
28	2	5	2	270	170	1.63	9	13.05	2.86	49.2	0.68	10	33.5	10.0%
29	2	5	3	230	150	1.52	8	13.5	1.63	75.9	1.10	12	59.4	13.6%
30	3	1	1	195	100	2.10	13	8.25	3.04	74.8	0.92	13	103.8	12.1%
31	3	1	2	171	93	1.65	11	10.4	4.15	57.0	0.68	16	45.5	13.5%
32	3	1	3	200	100	1.16	14	15.9	5.11	162.7	0.97	14	132.1	14.8%
33	3	2	1	190	95	2.30	14	14.5	4.06	84.0	1.98	14	55.9	14.3%
34	3	2	2	173	92	1.76	14	13.0	4.49	107.7	0.81	16	81.0	13.5%
35	3	2	3	170	90	1.07	11	11.8	2.31	51.4	0.76	14	38.4	13.6%
36	3	3	1	180	97	2.10	14	14.0	3.67	46.1	1.36	16	67.3	13.7%
37	3	3	2	171	93	2.10	13	13.1	4.72	94.6	0.71	16	70.3	14.4%
38	3	3	3	140	90	1.14	13	12.4	3.21	61.7	1.03	14	51.8	13.0%
39	3	4	1	175	58	1.80	11	11.5	3.94	31.2	1.39	12	37.5	13.0%
40	3	4	3	190	80	1.03	11	15.0	4.27	112.3	1.00	14	37.5	14.2%
41	3	5	1	193	105	2.30	11	14.5	3.94	100	1.39	15	103.0	13.7%
42	3	5	3	170	90	1.13	13	17.5	3.21	52.8	0.73	12	103.0	14.0%
43	4	1	1	195	110	2.30	13	18.5	4.59	190.0	1.49	10	151.4	13.5%
44	4	1	3	168	104	1.65	13	12.9	4.85	242.5	1.295	14	198.1	13.6%
45	4	2	1	260	165	2.30	14	14.25	4.065	210.6	1.36	13	167.5	13.6%
46	4	2	2	200	102	1.88	10	16.5	4.54	162.9	0.84	12	126.9	14.3%
47	4	2	3	182	115	1.65	11	11.85	4.05	137.1	1.08	13	104.8	12.6%
48	4	3	1	265	170	2.30	14	14.25	4.495	226.6	1.58	14	163.3	14.7%
49	4	3	2	177	99	1.43	12	14.1	4.41	116.5	1.21	14	88.0	14.4%
50	4	3	3	235	169	1.54	13	14.5	4.555	230.6	1.12	14	117.5	14.4%
51	4	4	1	200	114	2.40	12	18.0	4.67	161.4	1.05	16	120.3	13.3%
52	4	4	3	240	150	2.90	15	16.25	4.28	255.3	0.9	12	184.5	13.1%
53	4	5	1	190	128	2.10	14	14.75	3.565	156.3	1.195	15	53.9	13.1%

DATO	TRAT	PL	REP	ALPL	ALFMZ	DMITLL	NO_HOJ	LONGMAZ	DIAMAZ	PESOMAZ	PROFG	NO_CARR	PESOGF	P_HUM
54	4	5	2	184	110	1.87	11	12.3	4.41	78.2	1.21	14	67.2	14.2%
55	4	5	3	235	147	1.98	16	19.0	4.81	153.5	1.39	14	80.3	13.0%
56	5	1	1	210	137	1.12	11	13.5	5.28	197.0	1.40	12	163.0	17.0%
57	5	1	2	245	174	1.98	13	11.2	4.01	162.8	1.21	14	128.5	14.4%
58	5	1	3	290	220	1.84	11	22.2	5.08	317.6	1.95	12	248.9	17.6%
59	5	2	1	295	191	1.97	16	16.65	5.235	463.9	1.375	14	316.8	15.7%
60	5	2	3	260	170	1.63	12	14.0	4.56	111.5	1.58	12	85.9	13.9%
61	5	3	2	229	152	1.97	10	10.3	4.98	95.5	0.69	14	636.5	14.3%
62	5	3	3	160	160	1.58	10	17.1	5.76	323.8	1.28	12	262.1	19.1%
63	5	4	1	230	150	1.17	13	14.6	4.76	149.3	1.60	14	130.2	14.6%
64	5	4	3	250	180	1.70	10	15.0	5.26	210.3	1.67	12	166.6	15.8%
65	5	5	1	270	180	1.81	13	19.3	5.07	208.4	1.08	12	173.5	14.0%
66	5	5	3	250	240	1.89	8	18.5	4.75	192.3	1.65	14	157.7	15.5%
67	6	1	1	191	110	2.30	14	16.8	5.98	349.0	1.24	16	266.1	20.5%
68	6	1	2	210	110	1.24	9	16.9	4.91	228.8	1.05	14	177.9	15.2%
69	6	1	3	190	90	1.03	12	14.0	4.41	128.3	1.35	12	99.8	14.8%
70	6	2	1	213	80	1.82	15	16.6	5.13	230.0	1.49	16	186.6	14.6%
71	6	2	2	170	90	1.07	11	15.1	4.19	206.5	0.97	16	155.6	20.2%
72	6	3	1	213	113	2.12	14	17.4	5.13	249.5	1.73	16	192.5	18.3%
73	6	3	2	180	80	0.98	8	15.7	4.83	240.8	1.23	14	187.3	17.6%
74	6	3	3	240	90	1.09	13	15.0	4.88	171.4	1.33	14	124.9	15.6%
75	6	4	1	211	85	2.09	13	11.1	4.73	127.4	1.45	14	101.7	15.3%
76	6	4	2	190	80	0.99	10	9.3	4.29	71.5	0.89	14	54.0	14.3%
77	6	4	3	200	100	1.14	7	13.5	4.48	131.0	1.19	12	100.2	15.9%
78	6	5	1	205	85	1.45	13	12.3	3.60	41.3	1.05	8	25.0	13.9%
79	6	5	2	210	90	1.08	13	17.4	4.50	181.9	1.03	14	150.8	16.0%
80	6	5	3	170	90	1.08	10	9.5	2.57	21.4	0.78	16	16.8	17.3%
81	7	1	1	200	85	1.99	15	16.3	5.07	195.9	1.33	14	155.0	16.9%
82	7	1	2	190	70	0.89	8	15.1	5.07	180.1	1.13	14	146.3	17.5%
83	7	1	3	165	88	1.54	13	15.0	4.87	147.1	1.34	14	150.7	14.3%
84	7	2	1	190	95	1.89	12	15.3	4.93	192.9	1.41	14	134.2	19.0%
85	7	2	2	200	100	1.16	10	16.8	4.90	246.4	0.93	16	187.1	19.4%
86	7	2	3	162	70	1.76	12	15.6	4.98	87.8	0.99	12	160.7	19.4%
87	7	3	1	190	90	1.80	12	17.7	4.85	198.1	1.43	14	142.5	19.2%
88	7	3	2	170	70	0.91	10	17.2	4.39	151.4	1.02	14	121.7	15.6%
89	7	3	3	190	100	1.87	15	17.2	4.90	178.9	1.20	14	132.1	19.0%
90	7	4	1	190	80	1.91	14	16.8	4.24	112.8	1.28	14	66.3	19.1%
91	7	4	2	170	70	0.87	9	12.8	3.99	66.4	1.03	14	49.5	15.6%
92	7	4	3	200	118	1.98	13	17.6	4.98	205.3	1.98	12	57.9	14.8%
93	7	5	1	185	110	1.83	15	17.5	4.63	112.6	2.09	14	59.5	19.5%
94	7	5	2	160	80	0.98	10	15.8	4.49	154.8	1.23	12	114.1	18.5%
95	7	5	3	195	108	1.76	13	18.4	5.05	95.8	1.20	14	86.8	17.5%
96	8	1	1	285	208	1.90	16	17.5	1.99	77.7	0.98	12	47.4	18.9%
97	8	1	2	300	210	1.81	12	22.1	3.58	110.4	0.78	14	61.5	16.5%
98	8	1	3	280	200	1.76	8	19.3	3.89	116.4	0.76	14	70.5	18.9%
99	8	2	1	260	168	1.24	16	16.0	3.67	96.3	1.26	12	71.9	15.6%
100	8	2	2	270	185	1.71	13	23.4	4.12	213.4	0.89	12	149.0	19.0%
101	8	2	3	260	180	1.67	9	12.3	3.09	40.0	0.85	12	19.3	16.6%
102	8	3	1	280	185	1.95	16	19.0	4.23	140.3	1.35	14	90.3	17.6%
103	8	3	2	300	180	1.68	15	10	2.2	44.3	0.795	10	32.4	19.0%
104	8	3	3	310	220	1.84	12	17.7	2.06	103	0.91	11	56.9	16.1%
105	8	4	1	290	205	1.41	16	20.3	3.55	136.6	0.85	12	92.7	19.3%
106	8	4	2	270	173	1.65	13	16.5	3.185	136.2	0.805	13	98.4	15.0%
107	8	4	3	290	180	1.69	10	15.5	2.59	39.9	0.97	16	10.7	15.4%

DATO	TRAT	PL	REP	ALPL	ALTMZ	DMITLL	NO_HOJ	LONGMAZ	DIAMAZ	PESOMAZ	PROFG	NO_CARR	PESOGP	P_HUM
108	8	5	1	350	230	2.21	14	20.6	3.95	115.1	0.96	14	49.4	19.0%
109	8	5	2	290	215	1.82	11	16	2.72	66.1	0.7	12	40.2	15.5%
110	8	5	3	290	210	1.81	13	15.5	4.22	127.8	1.29	14	80.2	16.4%
111	9	1	1	290	205	1.78	14	19	2.015	58.3	0.865	13	28.8	17.0%
112	9	1	2	300	200	1.76	15	12.8	2.98	36.0	1.29	16	22.0	13.7%
113	9	1	3	320	200	1.98	16	16.1	3.02	38.7	0.91	15	25.4	17.2%
114	9	2	1	280	191	1.70	14	22.2	3.75	178.1	1.05	16	91.7	15.5%
115	9	2	2	290	210	1.81	12	13.6	3.01	36.0	1.29	20	3.0	16.6%
116	9	2	3	305	198	1.87	17	18.7	3.11	66.0	1.00	14	47.4	18.4%
117	9	3	1	325	225	1.76	16	18.0	3.68	115.9	1.14	16	67.7	17.5%
118	9	3	2	270	180	1.68	10	12.2	3.39	25.0	0.73	14	32.9	16.3%
119	9	3	3	275	187	1.43	15	16.2	3.21	65.8	1.00	14	50.3	12.9%
120	9	4	1	280	156	1.90	17	16.3	2.89	55.2	0.90	14	33.0	18.0%
121	9	4	2	300	220	1.78	12	19.1	3.21	53.1	0.97	12	46.4	18.5%
122	9	4	3	320	222	1.65	15	22.0	2.93	39.0	1.08	16	39.7	18.0%
123	9	5	1	310	205	1.96	16	16	3.26	144.9	1.15	13	103.7	15.3%
124	9	5	2	300	200	1.75	12	17.3	2.99	78.5	0.93	12	31.1	18.4%
125	9	5	3	350	257	1.76	16	13.9	2.24	15.8	0.93	14	67.4	15.3%
126	10	1	1	270	193	1.50	15	13.5	3.05	62.6	0.95	12	51.6	12.6%
127	10	1	2	300	200	1.77	11	15.75	3.405	148.3	1.06	13	109.8	14.5%
128	10	1	3	270	200	1.79	10	13.8	2.61	34.1	0.71	12	80.7	15.7%
129	10	2	1	290	190	1.83	17	14.9	3.00	46.0	0.89	14	29.7	13.5%
130	10	2	2	280	200	1.78	8	17.4	1.57	5.5	0.97	12	0.7	14.5%
131	10	2	3	260	190	1.75	10	15.6	3.04	144.2	0.685	12	15.2	16.0%
132	10	3	1	270	180	2.08	17	12.3	2.27	7.6	0.50	14	1.0	14.7%
133	10	3	2	270	180	1.66	8	10.45	3.49	110.5	1.14	13	86.7	13.6%
134	10	4	1	280	200	2.71	14	19.9	3.51	99.8	0.87	12	59.4	18.4%
135	10	4	2	270	190	1.71	8	13.6	3.44	71.9	0.73	16	53.5	15.2%
136	10	4	3	250	170	1.63	6	18.4	3.07	52.3	0.83	10	56.5	18.0%
137	10	5	1	290	205	2.71	18	14.5	2.60	17.3	0.82	14	1.3	15.7%
138	10	5	2	300	220	1.84	10	15.0	2.39	21.7	0.63	12	9.4	15.2%
139	10	5	3	250	170	1.66	10	13.9	2.81	55.1	0.69	16	5.4	15.3%
140	11	1	1	280	152	2.20	15	20.2	5.59	380.0	2.14	16	286.0	19.2%
141	11	1	3	270	190	1.65	15	15.7	3.625	161	0.79	13	128.5	13.0%
142	11	2	1	310	187	2.40	16	16.5	4.565	283.2	1.485	13	233.0	15.8%
143	11	2	2	265	182	1.61	15	19.0	4.37	73.3	0.84	14	15.6	17.5%
144	11	2	3	329	181	1.98	17	13.15	3.84	107	0.88	10	40.1	17.0%
145	11	3	1	260	170	1.60	15	18.7	4.68	178.0	1.38	16	111.0	19.0%
146	11	3	2	240	150	1.37	14	16.2	4.87	136.5	0.97	14	94.5	13.8%
147	11	3	3	283	205	1.76	14	20.1	4.95	260.0	1.08	12	188.7	18.0%
148	11	4	1	223	118	2.00	15	12.9	3.38	71.1	1.29	14	64.3	13.5%
149	11	4	2	300	175	1.43	14	17.0	3.38	143.8	1.04	16	88.7	17.0%
150	11	4	3	245	160	1.65	14	13.9	5.08	147.4	1.19	18	116.6	13.0%
151	11	5	1	330	194	2.00	17	19.5	4.48	189.9	2.02	14	156.8	17.4%
152	11	5	3	320	232	1.55	17	18.4	4.13	70.6	0.89	14	31.1	15.5%
153	12	1	1	220	125	2.20	13	16.0	2.39	22.6	1.28	10	14.4	14.0%
154	12	1	2	220	110	1.24	7	17.1	3.78	72.8	0.90	8	55.4	13.1%
155	12	1	3	210	119	1.65	12	16.7	3.84	75.0	1.38	10	58.6	13.2%
156	12	2	1	220	130	2.50	10	16.0	4.07	87.0	0.86	10	67.0	13.6%
157	12	2	2	210	130	1.40	9	13.9	4.41	111.8	0.83	8	84.8	13.2%
158	12	2	3	256	158	1.98	15	21.0	4.32	163.9	1.42	8	135.8	13.0%
159	12	3	1	230	142	1.37	11	17.0	3.32	78.2	1.39	6	63.8	14.0%
160	12	3	3	203	129	1.77	12	15.5	3.325	129.2	0.805	8	104.8	12.9%
161	12	4	1	204	112	1.33	13	13.7	3.38	62.1	0.90	10	50.4	14.1%

DATO	TRAT	PL	REP	ALTPL	ALTMZ	DMTLL	NO_HOJ	LONGMAZ	DIAMAZ	PESOMAZ	PROFG	NO_CARR	PESOGP	P_HUM
162	12	4	2	210	140	1.46	8	11.2	4.38	67.4	0.66	10	56.2	12.7%
163	12	4	3	260	136	1.98	13	19.8	4.11	135.2	1.16	10	53.3	13.3%
164	12	5	1	215	152	1.57	11	17.0	3.58	83.1	1.26	8	63.9	13.4%
165	12	5	2	230	110	1.24	8	18.1	3.68	87.5	0.83	8	77.8	12.4%
166	12	5	3	207	113	2.90	13	16.5	4.23	119.0	1.36	10	70.9	13.3%
167	13	1	1	163	93	1.60	12	9.8	3.10	28.4	0.78	14	21.3	15.0%
168	13	1	2	191	118	1.76	15	12.1	4.21	108.2	0.97	14	90.0	13.2%
169	13	1	3	180	110	1.32	14	14.9	2.98	40.1	0.71	14	36.7	14.6%
170	13	2	1	220	105	1.80	15	15.0	4.47	121.6	1.17	16	93.9	13.7%
171	13	2	2	178	109	1.43	13	9.4	3.62	42.8	0.79	14	70.9	12.8%
172	13	2	3	160	100	1.12	11	12.6	4.03	77.2	1.20	14	65.1	15.2%
173	13	3	1	172	89	2.10	13	11.5	4.75	117.9	1.28	14	90.5	14.7%
174	13	3	2	195	95	1.43	13	14.1	4.02	90.0	1.12	14	30.8	13.0%
175	13	3	3	190	90	1.20	13	14.7	5.11	140.8	1.13	14	110.1	15.5%
176	13	4	1	210	113	2.10	14	11.75	2.805	58	0.76	14	38.5	15.0%
177	13	4	2	190	100	1.88	12	10.9	3.73	50.6	0.84	14	38.9	13.6%
178	13	4	3	170	90	1.03	12	16.7	4.17	127.6	1.39	14	104.6	14.4%
179	13	5	1	218	123	2.00	14	12.0	4.43	68.1	0.92	14	52.9	14.5%
180	13	5	3	140	80	0.98	11	12.5	5.56	139.9	1.41	14	111.3	15.9%
181	14	1	1	290	150	1.95	14	16.2	5.33	262.0	1.49	14	220.0	14.7%
182	14	1	2	250	130	1.39	10	14.2	4.51	177.6	1.27	14	156.0	12.0%
183	14	1	3	232	144	1.32	12	17.8	3.79	131.9	1.38	8	114.6	13.0%
184	14	2	1	250	154	1.19	15	11.5	3.75	59.1	1.05	12	46.5	13.4%
185	14	2	2	230	110	1.24	9	11.9	3.47	100.0	1.00	16	82.7	13.0%
186	14	2	3	212	112	1.65	12	6.45	1.88	8.6	0.57	14	1.4	13.0%
187	14	3	1	253	156	1.63	16	18.4	4.79	227.2	1.80	12	186.6	14.8%
188	14	3	2	190	100	1.16	7	5.2	3.43	20.4	1.10	12	12.3	17.0%
189	14	3	3	228	143	1.87	12	16.25	4.41	252	1.2	14	221.6	12.6%
190	14	4	1	213	139	1.20	12	13.5	4.50	124.2	1.55	12	106.2	13.4%
191	14	4	2	230	145	1.49	9	16.2	4.63	193.9	1.02	12	171.1	14.5%
192	14	4	3	210	119	1.43	13	13.7	4.47	102.1	1.28	12	88.4	12.6%
193	14	5	1	215	163	1.65	12	11.0	5.47	99.3	1.81	18	86.2	13.8%
194	14	5	2	250	170	1.63	9	14.7	4.37	127.6	0.97	14	108.4	13.0%
195	14	5	3	225	130	1.55	14	17.7	4.45	177.6	1.09	12	153.1	15.0%
196	15	1	1	179	95	2.20	10	13.5	4.50	108.5	1.29	14	82.6	12.5%
197	15	1	2	203	100	2.90	14	17.9	5.09	266.7	0.91	16	191.5	15.2%
198	15	1	3	155	125	1.43	10	11.1	3.89	50.4	0.89	14	31.5	12.5%
199	15	2	1	160	70	2.40	12	15.0	4.75	189.9	1.57	14	156.8	13.6%
200	15	2	2	180	107	1.54	15	13.8	5.09	190.4	1.10	22	146.8	14.5%
201	15	2	3	191	122	1.54	13	10.4	4.45	68.4	0.96	14	48.9	12.5%
202	15	3	1	193	108	2.10	14	15.3	4.97	186.3	1.40	18	153.8	14.3%
203	15	3	2	159	73	1.65	11	12.7	3.21	49.0	0.79	14	43.3	15.1%
204	15	3	3	147	59	1.66	10	9.0	3.42	24.4	0.98	14	16.9	12.5%
205	15	4	1	190	105	2.00	13	16.0	4.75	153.7	1.19	20	121.0	13.9%
206	15	4	2	180	92	1.54	11	13.2	4.42	112.1	0.89	16	83.3	12.5%
207	15	4	3	181	101	1.43	14	11.6	4.35	63.7	0.89	18	42.9	12.0%
208	15	5	1	165	79	2.00	13	12.5	4.59	88.6	1.29	16	71.9	14.0%
209	15	5	2	182	101	1.65	14	16.6	4.79	206.5	1.00	14	156.2	15.8%
210	15	5	3	219	123	1.87	13	14.9	4.38	102.1	0.85	18	75.7	13.0%
211	16	1	1	260	160	1.78	13	18.9	5.24	268.6	1.55	14	223.3	14.8%
212	16	1	2	265	168	1.76	16	14.9	4.91	181.5	1.38	12	150.0	14.3%
213	16	1	3	160	150	1.54	11	14.0	3.75	82.9	1.28	14	62.1	14.1%
214	16	2	1	225	135	1.51	8	15.6	5.04	180.8	1.04	14	150.7	14.3%
215	16	2	2	242	123	1.87	11	10.5	4.81	102.4	1.13	12	77.0	13.5%

DATO	TRAT	PL	REP	ALTPL	ALTMZ	DMITLL	NO_HOJ	LONGMAZ	DIAMAZ	PESOMAZ	PROFG	NO_CARR	PESOGF	P_HUM
216	16	2	3	230	130	1.39	8	11.5	4.55	125.2	1.00	14	93.4	15.0%
217	16	3	1	260	148	1.61	13	17.6	5.07	230.6	1.59	14	190.6	14.5%
218	16	3	2	200	161	1.98	10	18.4	4.48	149.9	1.08	12	116.8	13.0%
219	16	3	3	260	130	1.38	8	13.0	4.94	220.0	1.48	16	171.4	16.9%
220	16	4	1	285	162	2.66	14	13.85	4.045	188.2	1.27	14	128.8	17.0%
221	16	4	3	220	130	1.41	7	16.5	3.90	69.5	1.44	12	53.9	15.0%
222	16	5	1	270	176	1.62	12	16.9	4.68	160.3	1.37	14	138.1	13.1%
223	16	5	2	241	129	1.76	12	13.8	5.01	183.5	0.96	12	150.1	14.6%
224	16	5	3	250	180	1.67	14	13.0	4.35	103.6	1.65	12	58.3	16.2%
225	17	1	3	272	180	1.66	17	18.6	4.37	192.4	1.49	10	151.8	14.5%
226	17	2	2	285	182	2.21	16	26.5	5.07	380.3	1.59	10	260.4	17.5%
227	17	2	3	320	180	1.98	15	19.2	5.03	257.4	1.25	14	193.5	19.0%
228	17	3	1	180	130	1.70	11	3.5	1.75	2.4	0.93	2	2.1	17.4%
229	17	3	2	235	135	1.32	12	11.5	4.14	89.3	0.80	10	67.9	14.6%
230	17	3	3	249	160	1.87	15	23.2	3.55	61.4	1.14	10	11.8	14.5%
231	17	4	1	280	193	2.40	14	17.2	4.25	99.3	1.25	12	165.8	13.3%
232	17	4	2	242	180	1.85	11	22.9	4.88	231.0	1.09	12	158.3	19.1%
233	17	5	1	230	190	2.20	12	20.0	4.64	207.3	1.28	12	78.2	13.2%
234	17	5	2	275	166	2.04	14	18.8	4.19	124.4	1.20	10	83.9	12.8%
235	18	1	1	230	137	1.90	12	23.8	4.99	193.9	1.49	12	63.1	13.0%
236	18	1	2	250	140	1.47	9	17.1	5.22	231.7	1.13	14	187.7	14.1%
237	18	1	3	253	168	1.54	14	13.5	1.32	95.8	0.98	14	125.4	14.0%
238	18	2	1	280	190	2.30	16	17.4	5.14	229.4	1.69	16	131.7	16.1%
239	18	2	2	270	170	1.63	14	18.9	5.66	335.5	1.12	12	247.6	18.5%
240	18	2	3	213	142	1.77	15	16.2	4.21	109.8	1.24	14	189.7	14.1%
241	18	3	1	270	157	2.00	14	17.5	4.63	157.9	1.68	12	182.5	13.2%
242	18	3	2	240	140	1.46	10	12.5	4.515	234.5	1.005	15	181.1	18.5%
243	18	3	3	158	147	1.55	13	17.9	4.22	121.6	0.73	14	181.8	14.2%
244	18	4	1	210	165	2.20	11	10.0	3.99	60.1	1.30	12	125.8	13.7%
245	18	4	2	300	230	1.86	10	17.5	4.57	400.3	1.195	11	314.4	19.0%
246	18	5	2	230	160	1.55	9	13.9	4.59	122.5	1.00	18	94.7	14.3%
247	18	5	3	228	149	1.87	12	14.9	4.61	91.2	1.21	16	47.4	16.4%
248	19	1	1	215	135	1.70	13	10	1.255	25.3	0.785	15	6.9	13.9%
249	19	1	2	215	105	1.67	13	13.65	4.355	126.1	0.755	14	43.0	17.4%
250	19	1	3	230	140	1.46	11	10.25	4.26	160	1.11	12	116.1	17.7%
251	19	2	1	220	135	1.60	13	13.1	2.215	54.7	1	14	25.6	13.8%
252	19	2	2	233	134	1.64	16	12.35	3.105	127.6	0.885	14	76.4	15.3%
253	19	2	3	210	120	1.30	10	9.2	4.015	96.2	1.065	16	71.8	15.8%
254	19	3	1	215	120	1.87	14	12.5	4.755	286.3	1.52	16	201.7	18.0%
255	19	3	2	228	131	1.70	15	14.6	5.19	157.6	1.28	14	101.5	18.4%
256	19	3	3	250	140	1.43	10	11.85	4.665	209.9	1.415	16	165.2	14.6%
257	19	4	1	237	137	2.12	18	11	4.48	173.3	1.24	12	128.9	14.7%
258	19	4	2	232	138	1.51	17	10.5	4.025	112.6	1.075	14	77.7	17.6%
259	19	4	3	250	140	1.46	10	11.5	3.92	86.8	1.125	16	60.7	15.8%
260	19	5	1	227	123	2.01	16	10.5	4.66	194.9	1.41	14	145.6	14.3%
261	19	5	2	215	124	1.44	13	11.4	3.59	43.9	0.80	14	26.8	14.2%
262	19	5	3	190	110	1.24	7	11.7	3.06	100.0	1.57	14	75.0	14.3%
263	20	1	1	290	180	1.46	12	10.3	4.37	97.7	1.30	16	84.5	12.8%
264	20	1	2	225	187	1.55	12	8.5	4.87	79.0	1.18	18	67.5	13.5%
265	20	1	3	220	150	1.52	8	11.2	4.59	75.5	1.08	18	61.2	13.0%
266	20	2	1	230	170	1.92	11	11.0	4.40	102.5	1.20	16	92.5	12.3%
267	20	2	2	275	215	1.98	14	7.7	4.77	79.7	1.31	16	68.0	15.8%
268	20	2	3	270	150	1.50	10	14.9	4.64	112.8	1.26	16	84.0	14.5%
269	20	3	1	260	120	2.47	11	6.1	4.07	29.1	1.13	16	23.4	12.0%

DATO	TRAT	PL	REP	ALTPL	ALTMZ	DMTLL	NO_HOJ	LONGMAZ	DIAMAZ	PESOMAZ	PROFG	NO_CARR	PESOG	P_HUM
270	20	3	3	250	150	1.54	6	15.4	3.64	33.8	0.75	16	12.8	13.0%
271	20	4	2	247	176	1.76	12	7.3	3.19	17.4	1.13	16	11.7	15.8%
272	20	4	3	300	180	1.71	7	14.7	4.59	105.9	1.28	16	76.1	15.0%
273	20	5	3	230	110	1.20	5	10.9	4.00	58.6	0.85	12	47.4	13.0%
274	21	1	1	210	150	2.00	12	14.9	3.46	61.7	1.03	18	48.0	12.0%
275	21	1	2	200	130	1.39	10	8.1	3.33	13.6	1.11	12	8.2	18.5%
276	21	1	3	186	115	1.11	9	6.2	4.24	24.4	1.31	14	15.9	13.5%
277	21	2	1	235	160	1.53	12	12.4	5.03	49.0	1.28	18	26.9	13.5%
278	21	2	3	260	179	1.76	13	16.8	4.90	95.2	0.89	12	74.1	13.0%
279	21	3	2	250	175	1.66	11	11.8	4.52	109.3	0.63	18	95.4	15.0%
280	21	3	3	213	134	1.65	12	13.4	4.98	84.4	1.04	12	60.7	12.9%
281	21	4	1	242	180	1.28	12	12.9	4.07	81.0	1.15	18	63.9	11.9%
282	21	4	2	240	130	1.39	7	4.9	3.19	18.7	0.67	10	17.3	13.5%
283	21	4	3	222	130	2.87	9	13.4	4.32	62.4	0.95	16	42.7	12.9%
284	21	5	2	300	190	1.73	10	21.1	5.26	323.9	1.43	16	258.4	16.5%
285	21	5	3	229	143	1.76	13	10.6	4.55	101.5	1.47	16	85.5	13.0%
286	22	1	1	230	156	2.05	11	11.5	3.76	43.9	1.03	16	28.9	13.3%
287	22	1	2	230	115	1.28	13	16.1	4.59	161.8	1.20	12	132.0	13.5%
288	22	1	3	240	130	1.39	10	14.0	5.33	211.6	1.61	16	80.5	18.0%
289	22	2	1	290	139	2.80	13	16.0	4.48	97.4	1.49	16	65.2	14.7%
290	22	2	2	260	140	1.46	9	13.05	4.81	224.8	0.84	13	176.2	14.0%
291	22	2	3	250	180	1.66	8	16.1	5.23	228.3	1.58	14	120.7	15.5%
292	22	3	1	240	139	2.40	13	11.5	5.00	82.0	1.09	16	56.6	14.7%
293	22	3	2	300	180	1.68	12	11.5	5.24	235.8	1.175	12	181.3	16.5%
294	22	3	3	240	190	1.73	9	13.5	5.21	169.7	1.17	14	119.0	15.7%
295	22	4	1	285	162	2.40	14	5.5	2.66	10.3	1.29	10	121.4	15.8%
296	22	4	2	270	190	1.73	12	18.9	5.78	293.4	1.13	16	227.5	18.5%
297	22	4	3	250	170	1.66	14	15.8	5.83	291.6	1.40	14	174.5	20.0%
298	22	5	1	280	174	2.20	15	18.5	4.57	176.3	1.49	14	178.7	12.5%
299	22	5	3	280	190	1.77	14	21.4	5.81	399.1	1.58	16	294.3	19.1%
300	23	1	1	285	210	1.50	14	16.5	4.29	127.2	1.95	12	98.2	19.5%
301	23	1	2	320	240	1.89	16	21.5	4.80	290.8	1.53	12	225.2	19.0%
302	23	1	3	300	200	1.78	9	15.75	3.125	153.5	1.27	11	120.1	15.2%
303	23	2	1	280	188	1.60	16	15.0	4.39	130.7	1.45	12	84.8	20.0%
304	23	2	2	250	190	1.73	12	16.8	4.75	128.7	1.00	14	103.1	14.0%
305	23	2	3	330	100	1.17	7	21.0	3.38	23.4	1.19	12	14.9	14.9%
306	23	3	1	310	185	1.61	14	14.5	5.04	193.3	1.84	14	145.7	19.5%
307	23	3	2	280	150	1.52	14	17.1	5.47	265.8	1.34	18	208.5	19.0%
308	23	3	3	270	190	1.72	13	15.3	4.49	257.7	1.63	12	137.5	14.8%
309	23	4	1	285	190	1.59	16	8.5	2.79	14.9	1.28	10	6.5	19.5%
310	23	4	2	310	220	1.84	11	18.8	4.67	187.5	1.34	14	141.7	16.5%
311	23	4	3	280	200	1.77	6	13.7	4.90	162.7	1.89	16	212.2	15.7%
312	23	5	1	335	235	1.81	16	14.9	4.99	211.1	1.97	14	163.4	18.0%
313	23	5	2	280	200	1.77	15	15.9	3.66	73.6	0.93	16	34.0	19.0%
314	23	5	3	310	220	1.84	14	15.2	4.00	126.4	1.28	12	87.8	17.3%
315	24	1	1	310	230	1.69	16	16.7	3.87	96.2	1.69	10	71.2	18.2%
316	24	1	2	240	190	1.73	10	14.4	4.71	134.6	0.91	12	99.5	18.4%
317	24	1	3	275	189	1.65	15	14.6	3.92	86.8	1.30	12	74.0	14.0%
318	24	2	1	310	230	1.59	16	16.5	3.43	69.9	1.25	12	40.8	18.6%
319	24	2	2	270	190	1.75	10	14.1	3.41	70.0	0.83	10	63.5	14.7%
320	24	2	3	236	167	1.76	14	7.0	2.29	20.8	1.28	10	16.0	14.0%
321	24	3	1	270	160	1.45	15	17.0	4.73	180.9	1.44	16	145.6	15.8%
322	24	3	2	280	150	1.52	9	17.7	3.155	134.4	0.98	11	99.1	18.4%
323	24	3	3	280	198	2.34	15	13.3	3.55	61.5	1.23	12	50.4	16.5%

DATO	TRAT	PL	REP	ALPL	ALTMZ	DMTLL	NO_HOJ	LONGMAZ	DIAMAZ	PESOMAZ	PROFG	NO_CARR	PESOGN	P_HUM
324	24	4	1	305	210	1.84	16	15.75	4.01	170.7	1.37	12	131.9	13.7%
325	24	4	2	260	230	1.86	11	12.2	2.715	63.5	0.865	8	49.8	15.4%
326	24	4	3	260	198	1.54	15	10.9	3.41	20.0	0.79	12	13.2	12.5%
327	24	5	1	314	220	1.58	16	19.0	4.95	221.8	1.68	12	172.3	16.0%
328	24	5	2	260	160	1.58	10	19.9	4.11	136.4	1.12	12	118.1	15.9%
329	24	5	3	260	188	1.87	16	17.8	4.13	132.0	1.37	12	112.7	13.5%
330	25	1	1	230	128	2.30	13	14.5	4.49	25.8	1.43	14	9.0	12.7%
331	25	1	2	195	104	1.76	13	11.0	4.09	72.6	0.71	14	59.1	14.3%
332	25	1	3	208	102	1.76	13	16.6	4.23	80.8	0.89	16	50.2	13.0%
333	25	2	1	185	65	2.30	12	15.5	4.07	85.5	1.53	12	103.8	14.9%
334	25	2	2	215	115	1.55	14	14.3	4.50	138.6	1.02	14	105.5	13.8%
335	25	2	3	212	119	1.32	15	12.5	4.28	71.7	0.88	14	52.7	13.4%
336	25	3	1	210	111	1.80	13	14.7	3.78	49.6	0.82	14	71.4	13.2%
337	25	3	2	167	104	1.54	11	11.2	3.81	59.0	0.72	18	44.3	13.8%
338	25	3	3	195	91	1.44	11	15.7	3.22	39.5	0.49	12	21.9	13.0%
339	25	4	1	205	123	2.00	13	11.0	3.89	57.8	1.26	14	35.5	14.1%
340	25	4	2	200	103	1.87	13	12.5	4.39	56.9	0.79	16	39.4	13.4%
341	25	4	3	164	59	1.21	11	10.1	4.09	52.8	0.78	14	38.2	13.0%
342	25	5	1	172	105	2.00	13	13.5	4.64	86.5	1.38	16	45.2	13.5%
343	25	5	2	202	104	1.87	14	12.75	3.655	146.9	0.71	14	91.5	13.5%
344	25	5	3	201	102	1.54	16	16.1	4.19	146.5	1.08	14	105.8	15.4%
345	26	1	1	210	110	2.50	13	17.0	4.96	161.6	1.75	14	123.0	14.6%
346	26	1	2	170	80	0.97	8	11.8	4.29	90.9	0.83	14	74.0	12.6%
347	26	1	3	170	110	1.28	14	16.4	4.32	147.2	1.06	14	98.5	15.2%
348	26	2	1	180	92	2.00	11	12.0	4.57	82.3	1.55	14	53.9	15.6%
349	26	2	2	210	110	1.21	8	13.0	4.55	102.8	0.67	14	74.2	13.8%
350	26	2	3	170	110	1.24	10	11.4	4.68	191.4	1.03	14	64.1	14.2%
351	26	3	1	179	65	2.30	13	10.5	4.18	63.7	1.38	14	50.7	14.8%
352	26	3	2	190	110	1.22	8	11.5	3.51	61.2	0.81	12	50.0	13.2%
353	26	3	3	190	100	1.16	12	16.4	4.61	113.7	1.21	12	50.4	13.2%
354	26	4	1	190	120	2.00	13	14.5	3.63	52.4	0.65	12	28.5	13.6%
355	26	4	2	180	190	1.73	10	10.8	4.01	45.1	0.98	14	32.1	13.0%
356	26	4	3	280	190	1.74	8	10.1	3.85	63.7	1.12	12	30.3	14.1%
357	26	5	2	170	100	1.16	10	9.7	4.00	55.6	1.00	12	43.4	14.3%
358	26	5	3	270	250	1.91	10	11.5	3.32	30.0	0.82	12	21.7	16.1%
359	27	1	1	270	180	1.88	14	13.5	4.015	138.9	1.095	17	109.0	13.3%
360	27	1	2	260	142	1.15	14	14.9	5.32	192.0	0.82	22	151.0	14.0%
361	27	1	3	270	160	1.57	7	15.5	4.93	204.0	1.56	14	154.1	17.5%
362	27	2	1	260	164	2.15	14	15.75	5.035	218	1.44	13	300.0	17.9%
363	27	2	2	238	130	1.22	11	12.9	4.59	125.0	1.16	12	103.2	14.8%
364	27	2	3	280	120	1.32	9	16.0	4.59	102.9	1.48	14	72.7	13.8%
365	27	3	1	250	160	1.97	13	14	3.3	89.7	1.14	15	30.2	15.4%
366	27	3	2	240	129	1.43	13	14.4	4.93	119.9	1.03	14	96.7	15.7%
367	27	3	3	230	130	1.38	10	13.5	4.82	140.8	1.37	16	113.1	13.6%
368	27	4	1	240	174	1.98	14	13.5	3.83	119.3	1.165	14	91.1	13.1%
369	27	4	2	276	159	1.99	12	19.3	4.99	180.2	1.37	12	111.2	19.3%
370	27	4	3	230	150	1.53	10	12.5	5.13	135.5	1.15	16	106.5	17.8%
371	27	5	1	255	150	2.07	14	16.5	5.25	204.5	1.08	16	168.1	15.1%
372	27	5	2	232	132	1.65	13	14.0	4.29	127.9	1.01	16	100.5	13.5%
373	27	5	3	220	140	1.46	11	13.9	5.09	139.2	1.18	16	104.7	16.5%
374	28	1	1	177	85	1.45	8	12.3	4.14	97.0	1.50	10	77.8	12.0%
375	28	1	2	104	48	1.10	8	7.4	3.01	23.4	0.55	10	17.1	13.0%
376	28	1	3	180	110	1.20	6	13.2	3.95	98.5	0.73	14	79.3	11.0%
377	28	2	1	155	87	1.00	9	12.9	3.44	69.9	1.05	10	60.2	11.6%

DATO	TRAT	PL	REP	ALPL	ALTMZ	DMITLL	NO_HOJ	LONGMAZ	DIAMAZ	PESOMAZ	PROFG	NO_CARR	PESOGRA	P_HUM
378	28	2	2	186	102	1.54	10	11.1	4.44	56.0	0.71	12	54.6	12.6%
379	28	2	3	190	80	1.07	8	9.6	4.32	84.5	0.85	12	70.0	11.0%
380	28	3	1	187	104	1.38	8	9.75	3.285	146.6	1.69	11	195.8	12.6%
381	28	3	2	123	90	1.11	10	7.4	3.33	34.3	0.89	10	27.3	13.0%
382	28	3	3	150	50	0.74	8	9.4	4.23	63.9	1.00	14	50.6	12.5%
383	28	4	1	200	120	1.77	10	15.1	4.35	109.7	1.37	14	94.5	11.8%
384	28	4	2	166	97	1.22	10	12.9	4.16	82.1	1.08	12	66.0	12.5%
385	28	4	3	170	70	0.92	7	9.1	3.45	56.4	0.84	10	46.5	12.0%
386	28	5	1	165	105	1.66	11	12.2	3.66	63.6	0.94	12	53.4	12.5%
387	28	5	2	191	98	1.54	10	13.1	4.48	87.5	1.23	10	71.1	12.4%
388	28	5	3	200	100	1.16	7	11.6	4.23	116.4	0.99	12	94.0	12.0%
389	29	1	1	161	80	1.71	9	10.5	2.68	19.9	0.55	12	16.5	12.6%
390	29	1	3	148	95	1.32	7	10.1	4.12	58.9	1.08	16	47.3	12.0%
391	29	2	1	228	130	1.50	14	16.5	3.77	105.6	0.75	12	84.8	12.0%
392	29	2	2	170	100	1.15	8	13.9	3.69	87.8	0.91	12	75.7	12.9%
393	29	2	3	195	94	1.76	10	14.7	3.56	77.1	0.62	12	59.3	12.0%
394	29	3	1	178	107	1.50	11	15.9	3.79	99.2	1.24	16	83.8	12.3%
395	29	3	2	180	90	1.07	6	7.4	2.28	22.8	0.62	10	17.3	12.5%
396	29	3	3	147	110	1.11	9	13.1	3.84	61.0	1.06	12	47.9	12.5%
397	29	4	1	191	105	1.70	10	16.3	4.39	101.0	1.13	12	78.4	12.2%
398	29	4	2	150	60	0.79	5	7.2	3.41	16.3	0.91	14	8.8	14.0%
399	29	4	3	173	106	1.54	10	11.1	3.90	49.8	0.75	12	31.6	12.0%
400	29	5	1	153	67	1.58	7	14.7	3.63	61.2	1.12	12	50.1	12.3%
401	29	5	2	200	110	1.26	7	11.7	3.21	46.8	1.10	10	39.1	12.2%
402	29	5	3	169	80	1.66	9	11.6	4.49	76.7	1.09	14	56.5	12.0%
403	30	1	1	141	94	1.70	10	16.0	5.47	217.5	1.56	18	8.7	14.5%
404	30	1	2	115	73	1.21	7	5.5	2.49	14.5	0.82	8	12.3	13.2%
405	30	1	3	133	92	1.54	9	7.8	3.02	25.7	0.81	9	10.5	13.7%
406	30	2	1	140	92	1.80	10	6.5	3.33	24.2	1.15	12	21.1	12.9%
407	30	2	2	106	45	1.44	7	3.5	2.70	4.8	0.88	8	2.7	13.6%
408	30	2	3	132	71	1.54	8	3.9	2.70	4.0	0.93	10	11.9	14.0%
409	30	3	1	162	115	1.40	13	7.6	1.60	40.8	1.17	12	31.1	12.3%
410	30	3	2	115	75	1.32	7	7.3	3.29	25.7	0.95	10	22.0	13.2%
411	30	3	3	122	55	1.32	7	7.3	2.78	6.6	0.64	10	26.6	14.5%
412	30	4	1	156	80	1.50	9	5.5	3.01	11.2	1.15	12	9.0	13.7%
413	30	4	2	116	66	1.00	8	6.8	3.52	20.4	1.11	10	17.0	13.0%
414	30	4	3	157	86	1.55	10	7.8	3.45	31.2	1.09	12	13.0	14.0%
415	30	5	1	150	85	1.60	9	4.9	2.89	13.8	1.08	10	11.8	13.3%
416	30	5	2	138	64	1.10	8	7.1	2.71	16.4	0.71	8	13.7	13.2%
417	30	5	3	149	82	1.43	9	7.5	3.43	23.1	1.02	10	18.6	13.0%

CUCBA



BIBLIOTECA CENTRAL

VECTORES DE MEDIAS DE LAS 30 VARIETADES ESTUDIADAS.

	MEDIAS 1	MEDIAS 2	MEDIAS 3	MEDIAS 4	MEDIAS 5	MEDIAS 6	MEDIAS 7	MEDIAS 8	MEDIAS 9	MEDIAS 10
ALTPL	284.07143	216.13333	178.30769	210.07692	244.45455	199.5	183.8	288.33333	301	275
ALTMZ	197.85714	141	91	129.46154	177.63636	92.357143	88.933333	196.6	203.73333	192
DMTLL	2.165	1.426	1.6646154	2.0230769	1.6963636	1.3914286	1.5426667	1.7433333	1.7713333	1.8871429
NO_HOJ	13.928571	9.2666667	12.538462	12.923077	11.545455	11.571429	12.066667	12.933333	14.466667	11.571429
LONGMAZ	17.5	13.763333	13.219231	15.165385	15.668182	14.328571	16.34	17.446667	16.893333	14.928571
DIAMAZ	4.6825	3.1423333	3.8558846	3.7284615	4.9788182	4.545	4.756	3.2696667	3.0456667	2.8753571
PESOMAZ	266.87143	64.76	79.715385	178.57692	221.12727	169.91429	155.08667	104.23333	67.086667	62.635714
PROFG	1.255	0.9493333	1.0561538	1.0230769	1.4077273	1.1985714	1.306	0.9433333	1.0156667	0.8196429
NO_CARR	13.642857	10.4	14.307692	13.461538	12.909091	14	13.733333	12.8	14.6	13
PESQGR	226.76071	52.546667	71.315385	124.9	224.51818	131.37143	117.62	64.72	46.03	40.057143
P_HUM	0.1707143	0.1344	0.1367692	0.1367692	0.1562727	0.1639286	0.1768667	0.1725333	0.1657333	0.1520714
	MEDIAS 11	MEDIAS 12	MEDIAS 13	MEDIAS 14	MEDIAS 15	MEDIAS 16	MEDIAS 17	MEDIAS 18	MEDIAS 19	MEDIAS 20
ALTPL	281.15385	221.07143	184.07143	231.86667	178.93333	240.57143	256.8	240.92308	224.46667	254.27273
ALTMZ	176.61538	129	101.07143	137.66667	97.333333	148.71429	169.6	161.15385	128.8	162.54545
DMTLL	1.7846154	1.7564286	1.5535714	1.49	1.8606667	1.71	1.923	1.7769231	1.61	1.6918182
NO_HOJ	15.230769	11.071429	13	11.733333	12.466667	11.214286	13.7	12.230769	13.066667	9.8181818
LONGMAZ	17.019231	16.392857	12.710714	13.646667	13.566667	14.889286	18.14	16.238462	11.606667	10.727273
DIAMAZ	4.38	3.7725	4.0703571	4.2166667	4.4433333	4.6267857	3.768	4.4357692	3.8366667	4.2845455
PESOMAZ	169.36923	92.485714	86.514286	137.56667	124.04667	160.5	164.52	183.4	130.34667	72
PROFG	1.2303846	1.0739286	1.0335714	1.2386667	1.0666667	1.3014286	1.082	1.2130769	1.1356667	1.1336364
NO_CARR	14.153846	8.8571429	14.142857	13.066667	16.133333	13.285714	10.2	13.846154	14.333333	16
PESQGR	119.60769	68.360714	68.25	117.00667	94.873333	126.03571	117.37	159.44615	88.193333	57.190909
P_HUM	0.1613077	0.133	0.1436429	0.1372	0.1359333	0.1473571	0.1559	0.1531538	0.1572	0.137
	MEDIAS 21	MEDIAS 22	MEDIAS 23	MEDIAS 24	MEDIAS 25	MEDIAS 26	MEDIAS 27	MEDIAS 28	MEDIAS 29	MEDIAS 30
ALTPL	232.25	260.35714	295	275.33333	197.4	197.07143	250.06667	169.6	174.3	135.46667
ALTMZ	151.33333	161.07143	194.53333	194	102.33333	119.89286	148	89.733333	95.285714	78.333333
DMTLL	1.6775	1.8721429	1.676	1.7166667	1.7506667	1.6014286	1.65	1.2573333	1.4035714	1.43
NO_HOJ	10.833333	11.928571	12.866667	13.6	13	10.571429	11.933333	8.6666667	8.7142857	8.7333333
LONGMAZ	12.208333	14.525	16.03	15.123333	13.463333	12.614286	14.676667	11.136667	12.478571	7
DIAMAZ	4.3208333	4.8785714	4.3163333	3.7593333	4.0883333	4.1771429	4.674	3.8983333	3.6257143	3.0926667
PESOMAZ	85.425	187.57143	156.48667	106.63333	78.033333	90.114286	149.18667	79.32	63.15	31.933333
PROFG	1.08	1.2910714	1.4593333	1.207	0.966	1.0814286	1.2033333	1.028	0.9235714	1.0046667
NO_CARR	15	14.214286	13.266667	11.533333	14.4	13.142857	15.133333	11.533333	12.571429	10.6
PESQGR	66.416667	139.76071	118.90667	83.873333	58.233333	56.764286	120.80667	70.546667	49.792857	15.33
P_HUM	0.1385	0.1598571	0.1746	0.1570667	0.1366667	0.1416429	0.1542	0.1216667	0.1239286	0.1347333



BIBLIOTECA CENTRAL

MATRIZ DE DISTANCIAS CUADRADAS DE MAHALANOBIS.

VARIEDADES	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
V1	0									
V2	7.5918218	0								
V3	10.180114	4.8514702	0							
V4	4.0418964	6.4238337	3.9117056	0						
V5	3.6550548	6.8369205	9.9530567	9.1477465	0					
V6	9.6126269	6.5820845	3.942135	7.2387799	9.4844411	0				
V7	12.966884	7.3139817	4.5335296	9.0760225	11.378335	1.9376457	0			
V8	7.4183547	2.0121061	7.9972304	6.9079088	10.356169	8.8322532	10.653229	0		
V9	10.328434	3.5955125	7.614468	8.1335034	12.321319	11.254972	14.39739	1.6878717	0	
V10	7.8414168	2.7169794	8.4976597	6.4426971	11.361366	12.513848	15.162818	1.6449948	2.1486676	0
V11	5.0963169	3.6234898	4.950744	4.3436768	7.0705823	5.7087232	9.4216135	3.6876351	3.8360376	5.4737203
V12	10.52636	3.2492172	7.1177849	8.0930339	11.041863	9.7873651	8.9748303	6.7374747	9.8673346	8.0240918
V13	9.6023436	4.7085477	0.6887877	4.7749413	9.0907107	3.3652379	4.2313824	7.1403784	7.3865584	8.4555836
V14	5.9183319	3.01905	3.7358443	2.7178645	4.850734	4.2658478	7.9016536	5.8195339	5.6918477	6.7059746
V15	9.022299	7.151126	1.3877691	2.7178645	10.052922	5.9132521	7.2310028	10.335437	10.432672	9.3608479
V16	4.5003088	2.2609994	4.5528599	4.5108454	3.8273749	4.8487989	7.4915218	5.4947045	7.0320393	5.8087048
V17	4.8555131	3.4380811	7.1243059	4.40656	8.235168	8.4681577	8.6391993	3.4240699	6.6363672	5.8281089
V18	2.383478	2.8295569	4.0906508	2.9829887	2.2807309	5.1382502	7.1286057	4.6581247	6.0270937	5.330159
V19	6.9251794	5.5388788	3.7751392	4.6802424	9.0349601	3.2315218	6.9941619	5.9477317	6.1640905	7.1052933
V20	11.787804	6.943785	10.092034	11.080372	10.725175	13.127057	18.922653	10.184523	9.0240037	6.6307691
V21	8.1024424	3.6485265	5.3256139	6.6256756	7.0673863	8.7066905	12.213177	6.5826187	6.4910891	4.6371283
V22	3.4856028	3.9468165	6.321542	4.5102798	4.8174834	5.4397074	9.3483117	6.1187833	2.9645477	3.9468165
V23	6.3001219	2.6860341	8.3746896	6.5635306	5.3300158	6.7403787	9.9425263	3.8836287	4.7681865	5.7378244
V24	6.3884407	2.269205	7.5485248	7.7553886	6.050687	9.6270232	11.54416	2.960024	3.5816101	4.1135535
V25	9.915963	4.7212438	0.8014956	3.9273754	11.054543	5.0762799	8.6592645	7.3301374	7.4495804	7.2558368
V26	7.788496	2.2140786	2.774179	4.7535707	7.2514962	4.7533335	5.6510172	5.3174501	7.1935578	5.381581
V27	5.5294289	2.9838887	3.8388878	4.9563128	5.2123671	3.7896267	7.7806372	5.2554157	5.8806243	5.5977485
V28	8.9475812	2.3992333	2.5124786	8.9475812	2.3992333	2.5124786	6.55334	6.3940918	8.1648481	7.6121553
V29	9.5251462	1.9368137	1.9406202	5.1385925	6.4249663	4.3820019	5.3283417	5.5834985	6.6429615	5.4863126
V30	12.850349	6.2216956	5.4316618	6.9017301	11.830528	6.8108798	6.0182772	9.8865694	9.8865694	10.103097
	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20
V11	0									
V12	6.2262415	0								
V13	3.7915754	7.3578743	0							
V14	1.4898023	6.2508448	3.2651622	0						
V15	6.2524898	9.8095555	2.0203131	5.0204105	0					
V16	2.4783506	4.7436044	4.1364474	0.8210988	4.7707001	0				
V17	3.2544813	2.4866784	6.6081245	5.6020744	9.0473092	4.4771365	0			
V18	2.55131	6.4866046	3.8840443	1.8087492	4.0729406	1.2420571	3.829821	0		
V19	2.9185047	10.688793	2.7215946	2.8062808	4.8773497	3.7851939	7.403554	4.1058936	0	
V20	8.0998733	13.861785	9.5053888	5.0905907	6.7542364	4.6834483	14.581506	7.2318383	7.4061212	0
V21	4.4656432	8.6173372	4.4834489	2.6501866	4.4556268	2.1088984	8.7367953	3.4321391	4.749977	1.3314343
V22	5.3469683	1.8842841	5.5978899	0.704308	5.41799	2.0839355	5.41799	2.0639355	3.1898121	4.3176476
V23	2.8635489	7.5183187	7.201648	2.1079716	10.845547	2.3022387	5.4494834	3.4272548	4.0831574	6.9970554
V24	2.2756499	5.2156867	6.010559	3.0654997	10.050896	3.2194161	3.2021433	3.8008243	5.3548745	8.048214
V25	3.9180725	5.7706839	0.9197032	3.5993853	1.5645414	4.0479978	6.2491079	4.4954533	4.0280502	8.2935434
V26	3.9584346	4.9311909	1.9305148	2.7248056	2.9750589	1.6441145	5.3277991	2.6266365	3.4777267	5.306264
V27	2.0800585	7.1052258	3.1799784	0.6479857	4.040568	0.8005297	6.2732874	1.6901579	2.5890676	3.6726401
V28	4.2912121	4.9111556	2.1732694	1.6073298	3.8261205	1.3858099	6.6381153	2.9926141	3.9997395	5.7532711
V29	5.0291087	5.0342161	2.3818811	2.9240641	2.7344834	2.3099879	6.5262155	3.1587851	4.8077467	5.90537
V30	10.122518	10.244671	4.6582409	7.8604745	7.0880474	6.7069259	10.229747	7.8180835	4.620352	10.87283
	V21	V22	V23	V24	V25	V26	V27	V28	V29	V30
V21	0									
V22	2.4015872	0								
V23	5.0751579	3.0429416	0							
V24	4.623892	4.2519854	1.7501112	0						
V25	3.8923706	5.0885844	6.6094913	6.7914412	0					
V26	1.7358089	2.353815	4.8062479	4.2302672	2.0396571	0				
V27	1.5811598	1.1300944	3.021016	3.1289967	3.1289967	1.9102874	0			
V28	2.5880022	2.8263605	4.4145509	5.0912316	2.4180791	1.0028284	1.3836073	0		
V29	2.6181174	3.9411473	5.9581283	6.1472886	1.9065129	1.0522854	2.0579158	0.8358428	0	
V30	7.1927007	7.0010601	8.1291211	8.280091	6.5688029	3.4084161	7.5148766	4.7766229	5.3552745	0