UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

Unidad Académica de los Ciclos Profesional y de Posgrado del CCH

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

APLICACION DE METODOS MULTIVARIADOS EN LA DETERMINACION DE INDICES DE CORTE EN FRUTALES

T E S I S

Que para obtener el grado de:

Maestro en Estadistica e Investigación de Operaciones

Presenta el Actuario

Inocencio Rafael Madrid Rios

México, D. F.

Mayo 1989

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

INDICE

PROLOGO

El presente trabajo tiene como propósito colaborar en la solución del problema relacionado con el consumo de fruta con calidad óptima.

Cuando la fruta se comercializa es importante que ésta llegue con la calidad que el público consumidor la prefiere, para ello es fundamental determinar el estado de madurez en que debe cortarse el fruto de acuerdo al destino establecido, ya que dicho estado afecta determinantemente la calidad que alcanzará el fruto en el almacenamiento,previo al consumo.

Para determinar el índice de corte de un frutal se requiere de manera general, de un conjunto de características que informen del estado de madurez del fruto al momento de corte. Y en particular, del conocimiento de los valores asociados a dichas características, que posibilitan que el fruto al fin del almacenamiento tenga la calidad deseada por el consumidor a quien se destinará.

En estas circunstancias los métodos multivariados permiten que un efecto (estado de madurez alcanzado en almacén) pueda ser estudiado considerando la interrelacion múltiple de las variables que lo producen y que determinan el índice de corte, siendo evaluado así el efecto con modelos que realmente sean una mejor representación de la realidad. Los métodos mencionados permiten un análisis donde no se tenga un aislamiento de cada variable.

Así, el desconocimiento de herramientas técnicas del manejo de datos multivariados y paqueteria estadistica , no ha permitido que disciplinas como la de Fisiología y Tecnología Postcosecha de Frutas (FYTPDF), quien trata con la temática de la investigación presente, se haya desarollado. Esto ocurre también con muchas otras disciplinas científicas, por lo cual es conveniente incorporar en ellas a los métodos de análisis multivariado como herramientas de trabajo cotidiano; así su uso les dotará de una forma de pensar para abordar sus problemas , diseñar sus investigaciones y analizar la información obtenida.

Con el propósito de incorporar la tecnología multivariada en la Disciplina FYTPDF es que se plantea éste trabajo. Para ello se propone un procedimiento de análisis multivariado para determinar el índice de corte, basándose en un modelo conceptual o marco de referencia , cuyo contenido se refiere a la asimilación de la cultura científica de la disciplina donde se presenta el problema.

El procedimiento que se propone indica que cuando no hay expectativas para determinar el índice de corte se replantee cuanto antes el estudio, a diferencia de cuando las expectativas son altas , donde se recomienda realizar el análisis hasta validar el índice de corte, de ser posible.

La tesis se divide en tres capítulos , dos apéndices y un anexo. En el primer capitulo se describe el marco de referencia del estudio , se comentan los antecedentes , la problemática de la determinación del índice de corte en frutales , y se hace una propuesta de análisis para su determinación . En el segundo se

comenta el Análisis Discriminante y otras técnicas multivariadas desde un punto de vista aplicado. El último capítulo contiene la aplicación de la propuesta de análisis referido en el capítulo primero, considerando la técnicas comentadas en capítulo segundo , para la determinación del índice de corte para chicozapote, del estado de Veracruz, destinado para consumo en fresco.

En relación a los apéndices, el primero presenta la terminología y definiciones de la disciplina Fisiología y Tecnología Postcosecha de Frutas y las del análisis discriminante, que son referidos en el desarrollo del trabajo mediante un exponente; su consulta permite compenetrarse con el tema tratado.

En el apéndice segundo, se desarrollan los dos enfoques del análisis discriminante descritos en el capítulo dos, empleando notación algebraica y matricial. Enfatizando las relaciones de interés para comprender su naturaleza y poder con el uso de paqueteria de algebra de matrices (como el MATLAB) reproducir los resultados obtenidos en la aplicación.

Finalmente el anexo se refiere a cuadros con datos de las mediciones en estudio y salidas del proceso de cómputo con el paquete estadistico BMDP empleado.

C A P I T U L O 1

Introducción

1. Introducción.

La orientación de trabajos de investigación a la determinación de índices de corte (DIDC), en la disciplina Fisiología y Tecnología Postcosecha de frutas (FYTPDF), ha propiciado que exista una amplia demanda de metodología estadística en este tipo de estudios.

Cuando se habla de metodología estadística hay dos aspectos que mencionar: el primero, referente a la planeación del trabajo, donde se proponen lineamientos para la recolección de los datos. El segundo para la organización y el análisis de ellos, para producir las conclusiones del estudio.

El presente trabajo pretende proporcionar al investigador avocado a este tipo de estudios, un procedimiento de análisis de información expreso y en un menor grado, elementos que le sean útiles en el diseño de la recolección de datos.

El procedimiento de análisis se estructura a partir de una serie de razonamientos dentro de la disciplina FYTPDF , que conducen a formar el modelo conceptual para la DIDC. Con éste se pretende que el análisis se facilite y realmente este orientado a la verificación de la existencia del índice .

1.1.-Antecedentes y problemática.

Cuando se lleva a cabo una investigación en fruticultura, es común que se clasifique en cualquiera de las dos siguientes áreas de trabajo :

Precosecha.- Que se refiere a la etapa anterior a la cosecha y

tiene como finalidad la producción del fruto.

Postcosecha. -Que se inicia a partir del corte del fruto. En ésta área se consideran las acciones convenientes para que los frutos producidos lleguen en condiciones óptimas al consumidor.

Por otro lado, para tener un alto grado de desarrollo en fruticultura, es importante no tan sólo producir las cantidades necesarias tanto para consumo interno como para exportación, sino también garantizar que dicha producción llegue a su destino final con una calidad óptima requerida (estado de madurez comestible = EMCOM).

En la investigación frutícola de México, tradicionalmente se han destinado más recursos para incrementar la eficiencia de la producción (Precosecha), que para eficientar el manejo y conservación de las frutas después de cosechadas (Postcosecha). Sin embargo, debido a que en fechas recientes han sido reiterados los reportes acerca de que las pérdidas postcosecha son considerables, es que se ha reparado en la importancia que tiene el buen manejo de los frutos a partir del corte.

De lo mencionado se desprende, que es inaplazable el desarrollo de técnicas para: a) Aumentar la vida útil de la fruta , b) Conservar y aumentar su calidad y c) Reducir pérdidas.

Para dicho desarrollo se requiere entre otras cosas, de la realización de estudios para la determinación de :

i) Indices de corte, épocas y métodos para realizarla.

ii) Manejo y acondicionamiento apropiado de la fruta cortada

en su preparación para el mercado.

iii) Condiciones adecuadas para el almacenamiento, transporte y distribución.

Indice de Corte.

En postcosecha el indice de corte juega un papel muy importante, pues la pretención de conservar la calidad de cualquier fruta depende fundamentalmente del estado de madurez[1] en que se corte, pues éste afecta posteriormente su almacenamiento , transporte y distribución. Así por ejemplo, cuando las frutas son cosechadas precozmente, éstas maduran de forma irregular o bien en algunas ocasiones no maduran, lo cual propicia que su calidad sea baja. Una cosecha tardia reduce la vida útil del fruto y lo hace más susceptible al ataque por microorganismos y a los daños mecánicos, produciendo una pérdida de su valor en el mercado.

En la disciplina FYTPDF, el índice de cosecha se define como:

Una(s) variable(s), poco influenciada(s) por condiciones ambientales y posible de ser medida(s) en condiciones de campo, que identifica(n) el estado de madurez en que conviene cosechar una fruta para que alcance y mantenga la calidad requerida para un fin particular.

A partir de la definición pueden hacerse los siguientes comentarios:

-Lo ideal es que una sola variable sirva como indicador de corte, sin embargo muchas de las veces en la práctica no es asi , por lo

cual generalmente, es necesario tomar una combinación de variables, tomando en cuenta la experiencia del productor.

-El índice debe estar formado por variables que sean más influenciadas por la naturaleza propia de la variedad frutal[4] de la especie[5] en estudio, que por las condiciones ambientales en que crece. Esto posibilita que el IDC que se determine tenga un grado de aplicabilidad amplio. Sin embargo ésto conlleva a que para su validación, se requiera tener información de diversas zonas y ciclos productivos que sean representativos del medio ambiente donde se produce el fruto. Por lo tanto la validación del IDC en la práctica se refiere entre otras cosas a la persistencia del índice tanto en el espacio como en el tiempo.

-Que el índice pueda ser medido en condiciones de campo significa

1) Que sea fácil y rápido de identificar tanto por el productor y los cosechadores, como por el personal encargado de la selección en el centro de acopio.

2) Que sea una medida objetiva.

- Para una misma variedad de una misma especie, el valor del índice o bien el IDC no es único, hay tantos valores del índice como fines a que vaya a destinarse el fruto. Los fines de manera general pueden ser : procesamiento y consumo. En un caso, el IDC variará dependiendo del producto procesado que se quiera obtener (néctar, almibar, fruta deshidratada , etc). En otro caso, el índice será distinto si la fruta se va a almacenar o a distribuir a mercados locales, nacionales o de exportación. Evidentemente el destino determina el tiempo que se pretende conservar el fruto y

la calidad de consumo, y ambas características determinan el tipo de almacenamiento, transporte y distribución requeridos. Bajo este contexto es que se pretende que el IDC, permita identificar el estado de desarrollo en que deben cortarse los frutos, para alcanzar la calidad requerida para un fin particular.

Determinación del índice de corte en el área de estudio..

Tradicionalmente el análisis de datos para la determinación del IDC , se ha llevado a cabo como un estudio de causa a efecto . En él se contrastan, al momento de corte, las características de los grupos de frutos que han sido cortados en diversas épocas de corte en un mismo ciclo productivo. Y cuando el contraste se determina como significativo se concluye que las calidades alcanzadas en el almacenamiento son distintas. En otro caso las calidades alcanzadas se dicen son iguales.

Sin embargo dicho enfoque resulta conveniente, solo para frutos de tipo no Climatérico[6], mientras en los Climatéricos[6] tiene serias deficiencias para el cumplimiento de su propósito.

Para comprender lo anterior, es conveniente comentar que los frutos de tipo climatérico deben ser cortados a partir de su madurez fisiológica[2], ya que de no ser así, no maduran o lo harán con características muy pobres de color , sabor , aroma y textura.

Estos frutos se caracterizan porque después de ser cortados aún cuentan con material de reserva para continuar madurando, lo que propicia de manera natural que haya fuertes cambios en el estado de madurez durante el almacenamiento.

Los frutos no climatéricos, en cambio, no continúan su maduración después del corte, es por ello, que en este tipo de frutos solo se se pretende prolongar la vida de almacenamiento del fruto en el estado de madurez en que se corta.

Cuando el fruto es climatérico y el análisis es de causa a efecto, es común encontrar que después de contrastar las características al momento de corte en dos épocas , éstas denoten diferencias significativas estadísticamente. Dichas diferencias sugerirían en principio que los estados de madurez de corte (EMC) son distintos. Sin embargo, para que este resultado sea real, habría que verificar que los estados de madurez alcanzados en el almacenamiento (EMA_s) realmente son distintos y que al menos uno corresponde al estado de madurez preferido por los consumidores (EMCOM), o bien que los EMAs son iguales pero alcanzados en diferentes tiempos de almacenamiento.

En la verificación pudiera ocurrir que al determinar los EMAs de las épocas, éstos sean iguales y alcanzados en un mismo tiempo de almacenamiento, lo que indicaría entonces que las diferencias observadas entre épocas de corte son aleatorias y por lo tanto las características no determinan el IDC puesto que : dos causas distintas (valores del IDC) se espera produzcan efectos diferentes (EMAs distintos) y esta situación no corresponde al planteamiento expuesto.

Así pues, en frutos climatéricos con el enfoque de causa a efecto se tiene este riesgo , que se traduce en la realización de un gran esfuerzo y empleo de tiempo en el análisis de datos para al final

llegar a un resultado falso (contradictorio).

Para minimizar este riesgo en frutos climatéricos , es conveniente utilizar un análisis de datos, que permita a cada paso valorar las posibilidades de continuar con el análisis para determinar realmente el IDC . Y donde si las posibilidades son altas ,ello indique altas expectativas para determinar el IDC y por lo tanto valga la pena continuar. A diferencia de cuando las expectativas sean nulas, donde continuar sería infructuoso , por lo cual sería conveniente dar por terminado el análisis y empezar cuanto antes el replanteamiento del estudio.

Siendo deseable lo anterior y considerando además que los estudios en postcosecha han incursionado más exitosamente en la diferenciación de los estados de madurez del fruto en almacenamiento (aunque a un costo muy elevado, en relación a dinero y tiempo, pues se determina con variables químicas y fisiológicas), por lo que se cuenta entonces con información para determinar los EMAs, es que se propone que el análisis de datos se aborde como un estudio de efecto a causa, de modo que ello permita sólo emplear el tiempo necesario en el análisis.

1.2 Propuesta de análisis para la determinación del índice de corte en frutales.

El enfoque que se propone considera como punto fundamental, el establecimiento teórico de las posibles situaciones resultantes, que pudieran tenerse entre los estados de madurez alcanzados en el almacén y los estados de madurez de corte que los inducieron, después de realizar las mediciones de las variables involucradas

en el estudio. Estas situaciones para propósito del estudio se esquematizan y se denominan patrones de comportamiento que pueden ocurrir y, los cuales pueden validarse o no, en presencia del modelo conceptual del marco disciplinario.

El modelo conceptual que se considera se forma a partir de los siguientes "axiomas" extraídos de la disciplina FYTPDF :

A_1) A diferencias mayores entre tiempos de cosecha (tC) , mayor es la probabilidad de que exista una relación directa entre el tC y el estado de madurez de corte (EMC).

Az) En iguales condiciones de almacenamiento y para iguales estados de madurez alcanzados (EMA_s) :

Az.1) A mayor tiempo de almacenamiento (tA) para llegar al el estado de madurez alcanzado (EMA), menor deberá ser el estado de madurez de corte (EMC). La relación entre el tA y el EMC es inversa.

Az.z) Si los tiempos de almacenamiento (tA) son iguales , iguales serán los estados de madurez de corte (EMC) independientemente de la época o momento de corte (MC).

A3) En iguales condiciones de almacenamiento y para estados de madurez alcanzados (EMA_s) distintos , los estados de madurez de corte (EMC) serán distintos independientemente de los tiempos de almacenamiento (tA).

Vale la pena aclarar. la diferencia que existe entre el estado de

madurez comestible de un fruto (EMCOM) y el EMA. El EMCOM juega un papel muy importante en la conceptualización de la DIDC y se define como el estado en que el fruto posee características sensoriales que lo hacen ser preferido por una gran mayoría de consumidores del lugar de destino para su consumo . Dicho estado está asociado con ciertas características del fruto, tanto químicas como fisiológicas y se determina con evaluaciones sensoriales durante el almacenamiento.

De acuerdo a la definición del EMCOM, dado un lugar de destino, dicho estado es único . pues a pesar de que los consumidores pueden preferir distintos EMAs durante el almacenamiento, hay uno que es mayormente preferido. Es importante anotar que los estudios para DIDC se diseñan de modo que al menos alguno de los EMAs obtenidos sea el EMCOM.

Así entonces se considerarán diversos EMAs en un patrón y si se determina que son producidos por EMC distintos y alguno corresponde con el EMCOM, entonces hay posibilidades de identificar y caracterizar el EMC que induce el EMCOM del lugar de destino de consumo, con el potencial de conservación requerido. Esto, siempre y cuando las variables que se consideran para el índice puedan diferenciar los EMC.

Ahora bien, para la construcción de los patrones, considere que se tiene una especie frutal de tipo climatérico, de la cual se cortaron frutos en dos fechas de corte distintas donde se tenía seguridad de que los EMAs asociados estaban próximos al EMCOM del lugar de consumo, pues , como se ha comentado , el IDC debe

identificar el EMC en que conviene cosechar el fruto para que alcance el EMCOM del lugar de destino, en el tiempo y con el tipo de almacenamiento que se requieran. Es por esto que en la práctica se debe establecer el tiempo requerido en que el fruto debe mantener una vida útil de acuerdo a su destino , para a partir de ello establecer las condiciones de almacenamiento.

Con el propósito de conceptualizar posibles patrones de comportamiento, consideremos la siguiente notación:

tC(j)= tiempo transcurrido entre el marcaje de frutos amarrados[7] y el corte de los frutos de la cosecha j. Se representará con el simbolo - - - - - - ->, y se referirá como tiempo transcurrido para el corte.

EMCt(j)= estado de madurez de corte t, de los frutos de la cosecha j.

tA(j)= tiempo de almacenamiento de los frutos de la cosecha j. Se representara con el simbolo ---------->.

MC(j)= momento de corte de los frutos de la cosecha j.

EMAk(j)= estado de madurez comestible máximo alcanzado k, de los frutos de la cosecha j .

Ademas EMCt+1(j) > EMCt(j) y EMAk+1(j) > EMAk(j') , j≠j'

Es importante anotar que con el índice t se refiere el estado de madurez observado para una cosecha particular j en su momento de corte. Asimismo k al estado de madurez comestible máximo alcanzado al fin del almacenamiento para una cosecha particular j. El sumar un número a t o a k es con el propósito de estalecer un orden en

dichos estados de madurez.

El modelo se establece con el propósito de identificar si los patrones resultantes de la medición en la práctica, son consistentes o no con dicho modelo. Si hay consistencia entre un patron y el modelo conceptual, y además son altas las expectativas para determinarse el índice de corte, se recomienda llevar a cabo un análisis estadístico de la información.

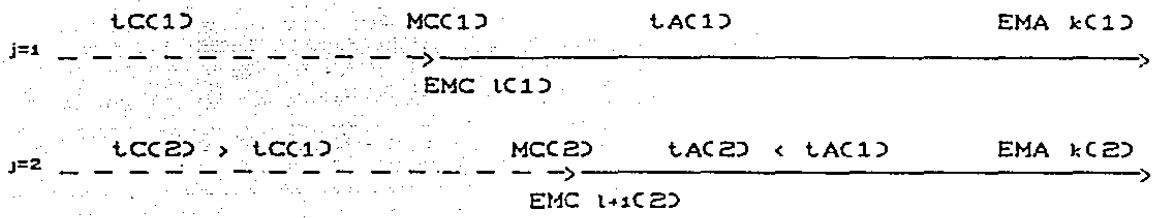
En el análisis de datos , el EMA se considerará como el efecto y el EMC como la causa, de esta manera se partirá del presente considerando los EMAs y se retrocederá al pasado tomando en cuenta los tA y los tC para proponer postulados acerca de los EMC, a partir de lo cual se identificaran los patrones donde existe mayor probabilidad de determinar el IDC (caracterizados por tener distintos EMC) y donde vale la pena hacer un analisis de datos.

A continuacion en el ESQUEMA 1 se presentan algunos patrones de comportamiento, de acuerdo a los efectos obtenidos , al considerar el marco disciplinario expuesto por los "axiomas".

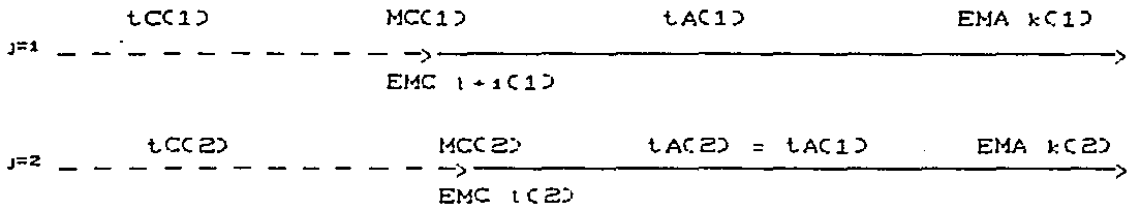
Diversos patrones se reportan en Lopez C. C [7].

ESQUEMA 1 ALGUNOS PATRONES DE COMPORTAMIENTO DE ACUERDO A SU CONSISTENCIA CON EL MODELO CONCEPTUAL, PARA LA DETERMINACION DEL INDICE DE CORTE.

Altas expectativas



Nulas Expectativas



El patrón con altas expectativas, corresponde a EMAs iguales alcanzados en tiempos distintos, por lo tanto de acuerdo al axioma Az.1 los EMC son diferentes. Por lo anterior puede establecerse que el intervalo de cosechas es suficientemente grande y entonces se cumple el axioma A1. En este patrón el modelo conceptual se valida.

En cuanto al patrón de nulas expectativas, se refiere a una

situación cuando el modelo conceptual no se valida , pues ante EMAs y tA iguales se esperaría correspondieran EMC iguales, esto de acuerdo al axioma Az.z. En este caso si se hiciese un análisis en los momentos de corte, este sería insustancial pues en caso de que existieran diferencias en las características entre los grupos de frutos cosechados , éstas realmente no estarían expresando el índice de corte.

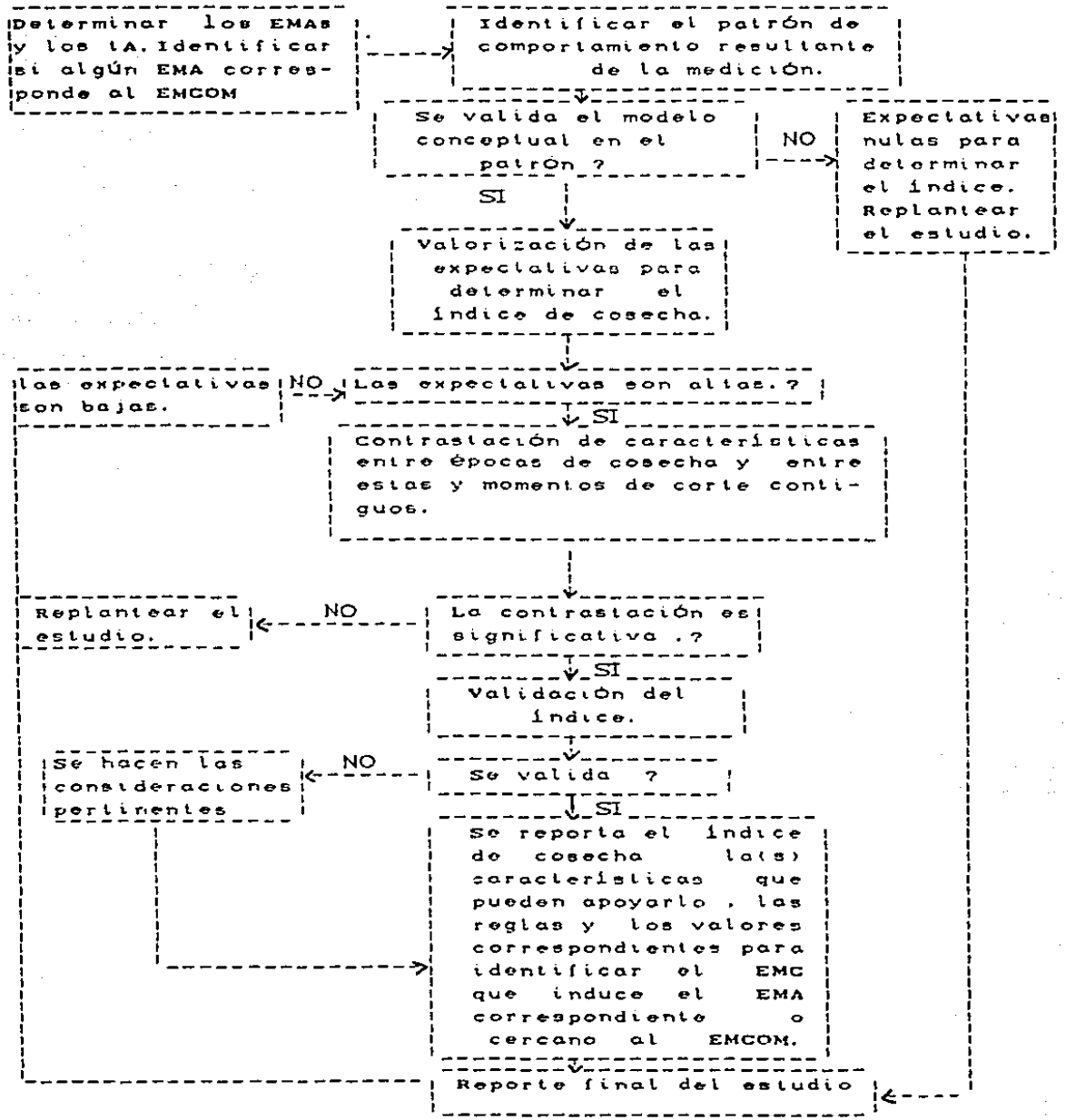
De esta manera al tomar en cuenta los patrones de comportamiento, junto con otras consideraciones (que se explican a continuación), se hace la propuesta del procedimiento de análisis de información a seguir para la determinación del índice de corte, en el DIAGRAMA 1 de la página 18.

En el diagrama se parte del presente comparando los EMAs y los tiempos de almacenamiento de las épocas de cosecha , para postular en el pasado si los EMC son iguales o no. Para ello se identifica el patrón de comportamiento y se verifica además si alguno de los EMAs corresponde al EMCOM.

Cuando en el patrón, el modelo conceptual no se valida, se hace necesario entonces un replanteamiento del estudio . La misma decisión se toma, como se puede ver, cuando el modelo se valida pero los EMC no pueden diferenciarse con las variables propuestas para formar el índice .

Por otro lado, si en el patrón resultante se valida el modelo y los EMAs son distintos , entonces hay posibilidad de determinar el índice realizando el análisis estadístico , al contrastar las características de tipo práctico en los momentos de corte. Si ésta

DIAGRAMA 1 PROCEDIMIENTO DE ANALISIS DE LA INFORMACION (DE EFECTO A CAUSA) PARA DETERMINAR EL INDICE DE CORTE EN FRUTALES.



De la contrastación puede resultar:

- Que no haya diferencias significativas, en este caso se realiza aún una contrastación con las variables de apoyo. Cuando la contrastación es significativa, dicha información sugerirá que el replanteamiento propuesto, se haga considerando variables prácticas relacionadas con las variables de apoyo que fueron discriminantes.
- Que la contrastación de mediciones de tipo práctico sea significativa, en cuyo caso se forma la regla para asignar o separar a los frutos de las diversas épocas de corte. Una vez formada la regla de diferenciación, es importante validarla tanto estadisticamente, como desde el punto de vista del área de estudio. Para la validación en el área de estudio,comparandola con las variables medidas en condiciones de laboratorio al momento de corte, que, como se ha dicho, han permitido la diferenciación de los estados de madurez en almacenamiento. Si se logra validar el indice se dice que se tiene un elemento de su existencia y se procede a reportarlo, indicando los valores correspondientes del indice del EMC que induce el EMA (=EMCOM) de acuerdo al mercado de consumo.

En el diagrama se sugiere considerar aparte de las dos fechas de corte distintas en estudio (con potencial para inducir el EMCOM) , fechas de corte contiguas (FDCC) a ellas (anteriores y posteriores) con el propósito de que si alguna de las fechas de corte potenciales induce un EMA que corresponde o esta cercano al EMCOM, se tenga la posibilidad de verificar que no haya traslape

en los valores de las características que determinan el IDC con las FDCC, pues, si este fuera el caso, la sensibilidad del índice dejaría mucho que desear.

Vale la pena recordar que la definitividad de un índice de cosecha se tiene cuando se valida en el tiempo y en el espacio, lo cual se logra solamente cuando se toman en cuenta nuevas cosechas en un ciclo productivo posterior y al menos otro huerto de la región, que envíe el fruto al mercado de consumo en estudio.

Así entonces el procedimiento que se propone indica que cuando no hay expectativas para DIDC se replantee el estudio, a diferencia de cuando las expectativas son altas, donde se recomienda de ser posible realizar el análisis hasta validar el índice de corte.

C a p í t u l o 2

Descripción del análisis discriminante y otras técnicas multivariadas.

2. Descripción del análisis discriminante y otras técnicas multivariadas.

2.1 Propósito del análisis discriminante.

El análisis discriminante (AD) es un método estadístico que permite derivar una regla que pueda ser usada óptimamente para diferenciar un conjunto de poblaciones conocidas (los EMC en nuestro caso) y además asignar nuevos frutos (observaciones) en alguna de estas poblaciones.

Para ello se parte del conocimiento inicial de pertenencia de frutos a los grupos por discriminar y de mediciones de características en los frutos que se consideran con potencial para diferenciar los grupos y permitan asignar nuevos frutos. De las características mencionadas se eligen o seleccionan aquellas que mejor discriminen los grupos[16].

En la literatura estadística se reporta que la regla de diferenciación puede construirse con dos enfoques:

- Las funciones de clasificación (FC), que se determinan basandose en los principios de la Teoría de las decisiones estadísticas. Y cuyo propósito es construir de una regla de clasificación óptima de acuerdo a un criterio predeterminado (el criterio más utilizado es aquel que hace mínimo el costo esperado de mala clasificación).
- Y La (s) función(es) discriminante(s) [FD] que se refiere a un enfoque que emplea la idea de investigar, si la representación de las diferencias que se tienen entre los individuos de los diversos grupos (variabilidad total entre grupos), al considerar q

variables discriminantes , es posible tenerla en una gran proporción en un nuevo espacio con pocas dimensiones (generalmente una o dos). Este enfoque de manera natural permite determinar los perfiles de los grupos y expresar gráficamente las diferencias entre los grupos.

En cualquiera de los dos enfoques, cuando las matrices de varianzas y covarianzas son iguales, la regla de diferenciación se expresa con funciones lineales, donde intervienen las variables seleccionadas para diferenciar los grupos. Véase apéndice 2.

Para continuar con la descripción del AD considere el siguiente problema: En fruticultura cuando se pretende comercializar una fruta , es fundamental determinar si el fruto que se pretende cosechar, alcanzará en determinado tiempo el estado de madurez comestible deseado, de acuerdo a su destino de consumo.

Suponga que se cortaron tres grupos de frutos de chicozapote en tres distintas épocas (C_1, C_2 , C_3) . Debido a la naturaleza destructiva de las mediciones por realizar y como la investigación era de efecto a causa , de cada grupo se consideraron 2 muestras. En una de ellas a los frutos se les midieron diversas características al momento de corte para determinar la(s) causa(s) y en la otra los frutos se median al inicio, en el transcurso y hasta el final del almacenamiento para determinar el efecto. Con base en las mediciones durante el almacenamiento , considere que los estados de madurez alcanzados por los grupos son respectivamente : semimaduro (EMA_1), maduro (EMA_2) y sobremaduro (EMA_3), obtenidos en diferentes tiempos de almacenamiento $tA_1 <$

$tA_2 < tA_3$. Lo mencionado anteriormente indica que los estados de madurez en que se cortaron los frutos son distintos , denotémoslos como EMC_1, EMC_2 y EMC_3. Como puede observarse entonces se tienen 3 EMC y se pretende determinar una regla que permita, predecir, si se cortara un fruto, a que época de corte corresponderá (EMC) y, como consecuencia, qué calidad comestible tendrá (EMA) y en que tiempo de almacenamiento la alcanzará.

Para establecer la regla de diferenciación de los frutos en los EMC, se requiere contar con frutos de cada una de las épocas de corte que se sabe alcanzarán determinadas calidades en los tiempos de almacenamiento mencionados , en los cuales se tengan mediciones de ciertas características al momento de corte, que de preferencia sean cuantitativas y prácticas. Y entonces con la aplicación del AD a esta información muestral, se pueden explorar las posibilidades de obtener un criterio cuantitativo de clasificación y discriminación con una medida asociada de efectividad y una indicación de la importancia relativa de las variables predictoras seleccionadas para diferenciar los EMC . Con el AD es posible obtener una forma de clasificar y separar los frutos , identificando las variables que influyen directamente en el estado de madurez que tiene el fruto en el momento que se corta. Intuitivamente, lo que se hace en AD, es comparar los frutos medidos considerando solo las variables predictoras selecciondas como discriminantes, para establecer que frutos se parecen o están más próximos en conjunto a los EMC y en base a ello proponer una regla de clasificación o discriminación según el procedimiento

empleado, para la asignación de los frutos a los EMC establecidos.

Internamente la regla de diferenciación valora en conjunto a las variables predictoras en cuanto a los errores de mala clasificación que producen. Dicha regla se resume en las FC o bien en las FD que están determinadas por las variables que influyen directamente en el EMC; en particular las FD constituyen nuevas variables índice. La regla de diferenciación formada a partir de las variables predictoras o discriminantes, se determina de tal manera que haya una alta probabilidad de clasificar a los frutos en el EMC del cual provienen. De esta manera cuando se quiere clasificar un nuevo fruto, se mide en el fruto prospecto para ser cortado, las variables que influyeron mayormente (solas y en conjunto) en el estado de madurez de corte y que sirvieron para formar la regla de diferenciación. Basándose en estas medidas y de acuerdo al criterio de clasificación o de discriminación establecido, se verifica si el fruto corresponde al estado de madurez de corte deseado; de ser así se corta y servirá para el fin propuesto. De ocurrir lo contrario se recomendará que no se corte el fruto, pues no alcanzará la madurez de consumo deseado de acuerdo a su destino en el tiempo especificado.

2.2 Procedimiento de análisis al aplicar el análisis discriminante.

2.2.1 Examen de datos.

Primeramente es necesario examinar los datos para detectar errores de codificación, de omisión y otros , ya que su presencia puede propiciar distorsión en el análisis y en las conclusiones

del estudio. Deben estudiarse también la presencia de datos discrepantes, tanto con criterios estadísticos como de la disciplina en estudio, de manera que su eliminación justificada o bien su tratamiento adecuado, permita aplicar un análisis de datos convenientemente. Así mismo es deseable tener presente como probar los supuestos del modelo como normalidad e igualdad de varianzas y covarianzas, y en el caso de que los supuestos no se cumplan , reflexionar sobre la conveniencia de aplicar o no correctivos , valorando para ello la dificultad de interpretación en el primer caso y el efecto en el resultado práctico en el segundo.

En este apartado es conveniente generar archivos de datos, con los cuales será posible producir cuadros de resultados. De los análisis estadísticos iniciales, son importantes aquellos que permitan determinar si hay características que pueden producir efecto de confusión en la discriminación de los frutos en los EMC inicialmente definidos. Cuando se encuentra que una característica actúa en este sentido, es necesario hacer algunas consideraciones en el análisis. Por ejemplo si en un estudio para determinar el índice de corte se tuvieran frutos de árboles de dos edades y existieran entre ellos diferencias sustanciales con las variables propuestas como discriminantes, para digamos 3 EMC, entonces en este caso como la edad de los árboles esta actuando como un factor de confusión seria conveniente determinar el índice de corte, para los 3 EMC por grupo de edad.

2.2.2 Selección de variables y el análisis discriminante.

La descripción que a continuación se comenta, se refiere a cuando se aplica el AD, haciendo uso de paquetes estadísticos como SPSS , SAS y BMDP entre otros. En particular se describen los procedimientos de la rutina 7M del paquete BMDP (Biomedical Package) versión 1987.

Para ello en la descripción primero se explicará como determinar las variables discriminantes (VDs) a partir de las variables potencialmente discrimiantes (VPDs) , y de manera simultanea se comenta como estudiar los atributos que las VDs tienen,al formar la regla de diferenciación de los EMC, para finalmente en la siguiente sección describir los mecanismos para validar una función de diferenciación.

Para determinar la regla de diferenciación de los EMC, primero se estudia de las VPDs cuales pueden considerarse como VDs, es decir, lo que se pretende es seleccionar aquellas variables que en su efecto individual y conjunto produzcan la mayor diferenciación posible entre los EMC. Un procedimiento usado para el fin anterior es la técnica de selección de variables a pasos (STEPWISE) que utiliza el BMDP y que consiste en incluir secuencialmente las VPDs que produzcan mayor diferenciación y excluir aquellas que no aporten una diferenciación sustancial.

En el procedimiento a pasos se procede como sigue:

a.- De las VPDs se determina cual produce una mayor diferenciación entre los EMC. Esta se considera como la primer VD.

b.- De las restantes VPDs se calcula cual de ellas produce una mayor diferenciación, al excluir el efecto de la primera VD. Esta se considera como la segunda VD.

c.- Al tener dos VDs se calcula la contribución parcial de cada una de las VD a la diferenciación total y esta se compara contra un porcentil preseleccionado de la distribución F, apropiado. Esta comparación proporciona una medida de contribución para cada VD en presencia de otras. Toda VD que no contribuya satisfactoriamente (F parcial < F preseleccionada) se declara VPD, es decir, sale de la regla de diferenciación. Este procedimiento garantiza que se maximice la estadística F o minimice la λ de Wilks, que sirven como criterios para probar la hipótesis de diferencia multivariada de medias entre los EMC con las VDs.

d.- El procedimiento a pasos sigue hasta que ya no se pueda incluir (ni excluir) variables a (de) la función de diferenciación.

En este procedimiento de selección de variables en cada paso se consideran todas las variables (VPD, VD), todas las VPDs pueden ser VD y las VDs puede ser VPD.

Puede decirse entonces, que en la selección a pasos una variable no se considera como discriminante o se omite de la regla para la discriminación, cuando entró por su efecto individual pero en una etapa posterior salió por no producir una diferenciación conveniente (pues incluso muchas de las veces reduce la diferenciación al estar presentes las demás variables) o bien porque nunca se incluyó, pues no produjo una diferenciación sustancial.

En el BMDP al hacer uso de la rutina 7M se obtienen la VDs con la selección a pasos y simultaneamente se reportan indicadores de lo conveniente de éstas variables para discriminar los grupos. Así entonces :

En el paso cero.- Se reportan los resultados de un análisis de varianza con un criterio de clasificación para cada variable y se incluye la variable que tenga la F univariada significativa máxima (que produce máxima diferenciación). También se reportan las funciones de clasificación (FC), una por cada EMC en estudio, y que sirven como funciones para asignar los frutos a los EMC, además se presentan estimaciones de las probabilidades a posteriori[19] para cada fruto de los EMC. En base a las FC o las estimaciones de las probabilidades a posteriori, un fruto se asigna al EMC en el cual se haya obtenido el valor mayor con cualquiera de éstos dos criterios. Se reporta además, una matriz de la estadística F multivariada, para probar la igualdad de medias para cada par de EMC . Esta F es proporcional a las estadísticas T^2 de Hotelling y la distancia de Mahalanobis al cuadrado, que sirven para el mismo propósito, y da una indicación de que pares de medias de los EMC son más próximas y cuales más apartadas. Esta estadística multivariada se denomina transformada de la λ de Wilks (la cual también se reporta) y puede compararse con la distribución F de Fisher. Las funciones de clasificación para separar pares de EMC que en la literatura se reportan (véase apéndice 2), puede obtenerse tomando la diferencia de los coeficientes de las funciones de clasificación respectivas.

Paso uno y subsecuentes.- Las F univariadas de las variables que potencialmente pueden incluirse, estarán condicionadas a las variables que ya están en la función, esto se lleva a cabo como un análisis de covarianza, donde las variables ya involucradas en la función se usan como covariables y las que potencialmente pueden entrar se utilizan como la dependiente, es decir, se generan F parciales. Las demás estadísticas mencionadas en el paso cero son recalculadas. En el último paso se reporta (n) las FDC's) con coeficientes estandarizados y no estandarizados.

Debe anotarse que el procedimiento a pasos es una técnica de selección de variables de tipo automático, donde:

a.- No necesariamente se tiene el mejor conjunto de VD's que diferencien los EMC.

b.- No se involucran los conocimientos que el investigador tiene sobre la naturaleza de las variables, pues el orden en que éstas entran como discriminantes usualmente no es de significancia práctica.

c.- No es posible incluir las VD's que puedan ser más económicas y prácticas en su obtención.

d.- Se inhibe la posibilidad de tener indices de corte alternativos, pues no es posible probar jerarquizaciones de las VPD's, para su evaluación como posibles grupos de VD's.

Como en la práctica de la determinación del índice de corte, muchas de las veces se quiere obtener una mejor diferenciación, garantizando la presencia de algunas variables, entonces es deseable proceder a jerarquizar las variables de interés y contar

fuera de ella, los frutos asignados erróneamente.

La matriz de clasificación mencionada tiene el inconveniente de que subestima la probabilidad de mala clasificación (sobreestima la probabilidad de clasificación correcta), ya que para formar la regla de clasificación y validarla se usa la misma muestra , por lo cual es de esperarse que la validación de la regla de diferenciación tengan deficiencias. Para subsanar este inconveniente, en la literatura estadística [Johnson A. R and Wichern W. D: 5] se propone contruir la matriz de clasificación obtenida por un procedimiento conocido como JACKKNIFE[22] y que consiste en dejar fuera del análisis uno a uno de los frutos muestra y con los demás estimar las funciones, para después clasificar el fruto omitido. El proceso se repite tantas veces como frutos muestra haya.

En el BMDP la matriz de clasificación JACKKNIFE puede imprimirse en cada paso de la seleccion de variables, además de las distancias al cuadrado de Mahalanobis y las probabilidades a posteriori estimadas para cada individuo , si en el párrafo DISCRIMINANT se especifica la instrucción JACKKNIFE.

Otra manera de proceder para la validación de las funciones de diferenciación[27,28] es considerar una submuestra de los EMC para formar las funciones ,y utilizar la otra parte de las muestras para elaborar la matriz de clasificación. Este procedimiento recibe el nombre de validación cruzada. Los frutos que se consideran para cada una de las dos submuestras se denominan respectivamente muestra para la clasificación y muestra para la

fuera de ella, los frutos asignados erróneamente.

La matriz de clasificación mencionada tiene el inconveniente de que subestima la probabilidad de mala clasificación (sobreestima la probabilidad de clasificación correcta), ya que para formar la regla de clasificación y validarla se usa la misma muestra , por lo cual es de esperarse que la validación de la regla de diferenciación tengan deficiencias. Para subsanar este inconveniente, en la literatura estadistica [Johnson A. R and Wichern W. D: 5] se propone contruir la matriz de clasificación obtenida por un procedimiento conocido como JACKKNIFE[22] y que consiste en dejar fuera del análisis uno a uno de los frutos muestra y con los demás estimar las funciones, para después clasificar el fruto omitido. El proceso se repite tantas veces como frutos muestra haya.

En el BMDP la matriz de clasificación JACKKNIFE puede imprimirse en cada paso de la selección de variables, además de las distancias al cuadrado de Mahalanobis y las probabilidades a posteriori estimadas para cada individuo , si en el párrafo DISCRIMINANT se especifica la instrucción JACKKNIFE.

Otra manera de proceder para la validación de las funciones de diferenciación[27,28] es considerar una submuestra de los EMC para formar las funciones .y utilizar la otra parte de las muestras para elaborar la matriz de clasificación. Este procedimiento recibe el nombre de validación cruzada. Los frutos que se consideran para cada una de las dos submuestras se denominan respectivamente muestra para la clasificación y muestra para la

validación ; estas pueden tomarse aleatoriamente dentro de cada EMC.

Las matrices de clasificación mencionadas, en base al porciento de clasificación errónea para cada EMC, nos proveen de un criterio de precisión relativo a la asignación de los frutos y nos da una idea de la bondad de la diferenciación lograda. Otros dos criterios que nos informan acerca de dicha bondad , se refieren al estudio de las funciones discriminantes , las cuales al evaluarse en los frutos muestra de cada EMC constituyen los llamados datos canónicos[24]. Con estos se procede a un análisis canónico para probar la independencia entre dos conjuntos de variables. El primer conjunto se forma con las variables predictoras y el segundo con variables indicadoras con valores cero y uno para indicar pertenencia a los EMC. Uno de los criterios para determinar la bondad es obteniendo el grado de dependencia de los dos conjuntos de variables canónicas mediante la correlación canónica[30], la cual indica que a mayor correlación mayor diferenciación de los EMC. Otro criterio de la bondad de la diferenciación lograda se refiere al porciento de diferenciación o variación de cada función discriminante[31] (variables canónicas del primer conjunto) en relación a la diferenciación o variación total obtenida con todas ellas.

2.2.4 Análisis de la regla de diferenciación.

para su reporte.

En base a las FD significativas, con coeficientes y variables

estandarizadas (véase apéndice 2), se pueden identificar finalmente las variables de mayor poder discriminante de acuerdo a la magnitud y signo de sus pesos discriminantes y su aportación a la diferenciación de los perfiles de los EMC, considerando sus valores de la estadística F parciales (a mayor valor mayor capacidad en la diferenciación) y la asociación lineal entre las variables discriminantes y las FD (contribución relativa de cada variable con respecto a la FD). Con esto se describen los perfiles de los EMC , considerando el rango de los valores de las variables discriminantes que deben ocurrir combinadamente,para clasificar los frutos en cada uno de los EMC.

En este punto es común presentar una gráfica de las FD[32] o variables canonicas (VC). Cuando se tiene sólo una VC significativa, se presenta un histograma. Para el caso de dos VC, estas se expresan en un plano discriminante donde cada fruto es un punto en el espacio y cada función discriminante o VC corresponde a una dimensión. Los puntos son proyectados sobre un plano , seleccionado de modo que los EMC se aparten lo más posible y presenten una buena impresión global de las diferencias entre los EMC . El eje de las abcisas es la dirección donde los EMC tienen una mayor separación (VC1) y la ordenada (VC2) es quien tiene la mayor separación de los EMC, en una dirección perpendicular (ortogonal) a la abcisa. Otra forma de representar gráficamente ,lo anterior es con el mapa territorial[33] para la clasificación de los frutos en los EMC. Aquí se delimitan las áreas asociadas a los EMC.

Es conveniente comentar que las ponderaciones discriminantes estan sujetas a las mismas críticas que los coeficientes de los modelos de regresión. Así,una variable discriminate con ponderación pequeña, puede interpretarse como irrelevante o bien que ha sido parcialmente excluída de la discriminación pues tiene un alto grado de asociación con las demás variables.

2.2.5 Clasificación de nuevos frutos a los EMC.

Esto puede llevarse a cabo con las FC o las FD :

En el primer caso se calculan los valores de las FC o se estima la probabilidad a posteriori en cada uno de los EMC. Con cualquiera de los dos criterios el fruto se clasifica en el EMC donde se obtenga el valor más grande.

En el segundo se consideran las FD relevantes, con coeficientes y variables no estandarizadas. Con los valores expresos de las variables consideradas como discriminantes, se evalúa el fruto con las FD y considerando su ubicación en el mapa territorial se asigna al EMC señalado.

C a p i t u l o 3

Aplicación : Determinación del índice de corte en chicozapote.

3.1 Descripción del estudio.

El estudio que se presenta se llevó a cabo en La Comisión Nacional de Fruticultura (CONAFRUT) a partir de 1985, y corresponde a la linea de investigación " Establecimiento de índices de cosecha mediante la observación de cambios físicos , químicos y fisiológicos durante las etapas de crecimiento desarrollo y maduración-senecencia[2,3] de los frutos ".

El chicozapote se desarrolla en un clima tropical y por su comportamiento fisiológico se clasifica como climatérico[6] . La investigación se realizó con el propósito de determinar el índice de cosecha que permitiera identificar el estado de madurez en que conviene cosechar un fruto para que alcance el estado de madurez comestible para la República Mexicana.

Los frutos considerados en el estudio provienen de 15 árboles (6 de 11 años y 9 de 17) representativos de una plantación comercial en Martinez de la Torre, Veracruz. De cada árbol se marcaron aproximadamente 100 frutos amarrados[7]. Estos fueron marcados de la parte media del árbol hacia arriba, ya que los de abajo generalmente no reciben suficiente luz y no alcanzan su desarrollo, añadiendo que la caída de los frutos en esta zona es muy frecuente.

De los frutos marcados se seleccionaron aleatoriamente muestras para cada una de las 9 evaluaciones realizadas y dos evaluaciones más correspondientes a 2 épocas de corte en estudio, primera y segunda cosecha : PC Y SC , con posibilidades de inducir el EMCOM.

En la muestra de frutos para cada evaluación estaba

representado,por lo menos con un fruto, cada uno de los árboles muestra y la selección de los frutos por árbol se realizó de manera aleatoria alrededor del árbol.

La primera evaluación se refiere al momento del marcaje de los frutos. En general las evaluaciones de 1-9 permitieron reportar los cambios físicos, químicos y fisiológicos durante el crecimiento, desarrollo y maduración de los frutos [López C. :7].

Los grupos de frutos relacionados con la determinación del índice de corte corresponden a la novena evaluación (E9) , la primer cosecha (PC) y la segunda cosecha (SC). La PC y la SC son épocas de cosecha donde se pensaba había muchas posibilidades de que se induciera el estado de madurez comestible al final del almacenamiento (EMCOM) y se refieren a la producción plena y tardía. La E9 es una época de corte contigua a la PC y a la SC.

Las muestras bajo estudio, en cada grupo, se subdividieron en dos , una para realizar mediciones al momento del corte y la otra durante el almacenamiento (en el transcurso y al final). Los cuadros CA y CB que a continuación se presentan corresponden a las características que se midieron en los momentos mencionados , en la PC y SC . En particularen el cuadro CA se señalan grupos de variables de acuerdo a su uso en al análisis de datos, que más adelante se comentará.

CA : Variables medidas en frutos de Chicozapote al momento de corte (árboles de 11 y 17 años)

USO DE LAS VARIABLES	V A R I A B L E S (unidades empleadas)	
Descriptivas	EV : Evaluación	EA : Edad del árbol
	NA : Número de árbol	NF : Número de fruto
Para determinar el índice de corte*	VO : Volumen (c.c)	DP : Diámetro polar (mm)
	DE : Diámetro ecuatorial (mm)	
	FS : Fuerza de separación o peso requerido para separar el fruto de la rama (gr).	
	LAE: Látex escurrido al cortar el fruto (ml/min).	
De apoyo al índice de corte* (físicas)	GB : Grados Brix (escala brixometro)	
	PFCC(PFSC): Peso del fruto con (sin) cáscara	
	NSC(PS) : Número (peso) de semillas	
De Laboratorio	AT : Azúcares totales+	
	FV : Fructosa verdadera+	
	GV : Glucosa verdadera+	
	ART : Azúcares reductores totales+	
	SV : Sacarosa verdadera = Azucares no reductores totales+	
	POL : Polifenoles**	

* Estas variables fueron medidas en condiciones de campo.
** Obtenido con el método de Folin-Denis[14], se reporta en %.
+ Obtenidos con el Método de Ting-V[14], se reporta en %.
Nota : Además se registraban los días de amarre al corte[7], los días grado[11], color externo de cascara y color de pulpa[14].

CB: Variables medidas en frutos de Chicozapote durante el almacenamiento (árboles de 11 y 17 años)

De laboratorio	Las mencionadas en el cuadro anterior pero de replicas*.
Sensoriales	Apariencia : Color, Olor, Sabor, Textura.
Fisiológicas	CO_2 desprendido, para construir la curva de respiración por fruto(también de replicas).

* se refiere a muestras que son tomadas de una mezcla que se hace con varios frutos.

En la E9 solo se tuvieron mediciones al momento de corte , véase cuadro CA.

Los tamaños de muestra para la E9, PC y SC al momento corte fueron de 19 frutos. Este tamaño se determinó de acuerdo a los recursos máximos con que se contaba para el trabajo de campo.

Determinación del patrón resultante de la medición.

Con el propósito de instrumentar el enfoque de análisis propuesto en el capítulo 1, primero se determinaron los EMAs para la PC y la SC. Para ello se consideró :

(1) Un análisis sensorial[12] al final del almacenamiento.- Con este análisis se obtuvo solo una diferencia en sabor donde el panel de acuerdo a la escala Hedónica[13] calificó a la SC como gusta mucho, a diferencia de la PC donde expresaron que gustaba moderadamente . En cuanto a el color, sabor y textura donde no hubo diferencias se estableció el siguiente perfil para la PC y para la SC : los frutos son firmes pero no duros con superficie granulosa y algunas partes lisas, la cáscara de color beige extremadamente delgada, desprendible y masticable, la pulpa es de color café con veteado rojizo, jugoso y dulce, el jugo escurre en el momento de la mordida, presenta una granulosidad fina similar a la pera que no molesta durante la masticación y pasa desapercibida. En conjunto se forma una pasta blanda de fácil deglusión.

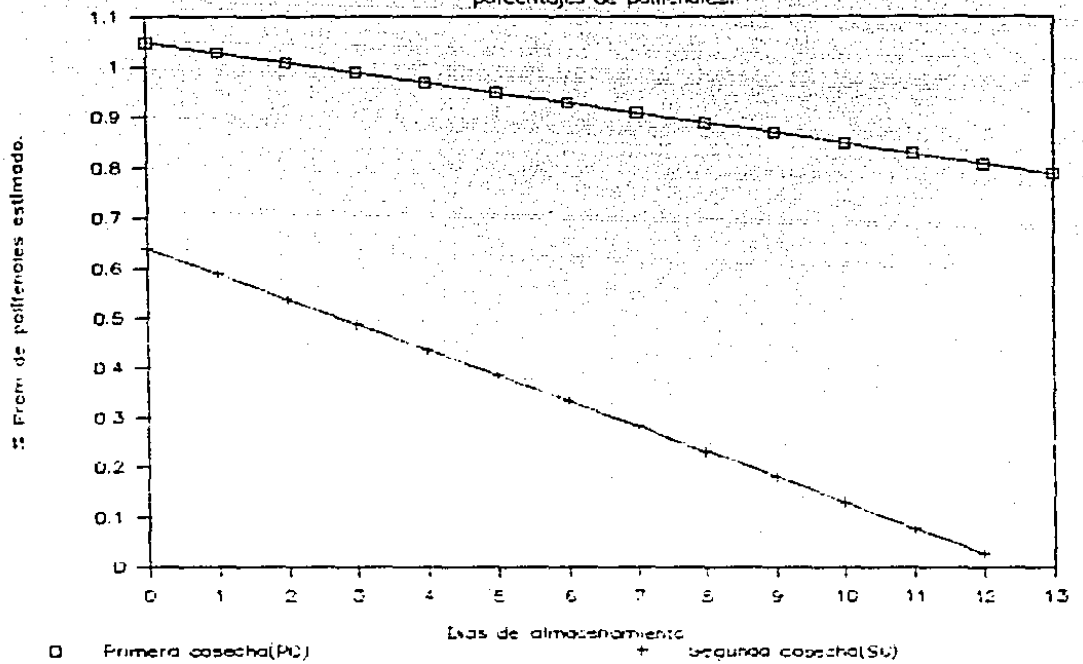
(2) Un análisis de tipo longitudinal durante el almacenamiento del fruto, con la variable porciento promedio de polifenoles[8].

Del análisis realizado, entre el tiempo de almacenamiento y el porciento promedio de polifenoles, se obtuvo una correlación lineal simple significativa alta y negativa, en las dos cosechas, como se esperaba de acuerdo a conocimientos del área cuando se considera que los EMAs son distintos. Al realizar la prueba de hipótesis para la igualdad de modelos de regresión simple se obtuvieron diferencias significativas, es decir, los modelos de regresión fueron distintos. En particular al establecer la diferencia de las relaciones de la PC con el tiempo y de éste con los polifenoles de la SC, se estableció que existía una diferencia de 0.408 y 0.761 % al inicio y al final del almacenamiento respectivamente, de la PC con respecto a la SC. Además los coeficientes de regresión de la PC y SC (- 0.02 y - 0.051) indicaban que la disminución por día de los polifenoles era más del doble en la SC con respecto a la PC. Lo comentado se refleja elocuentemente en la gráfica No 1.

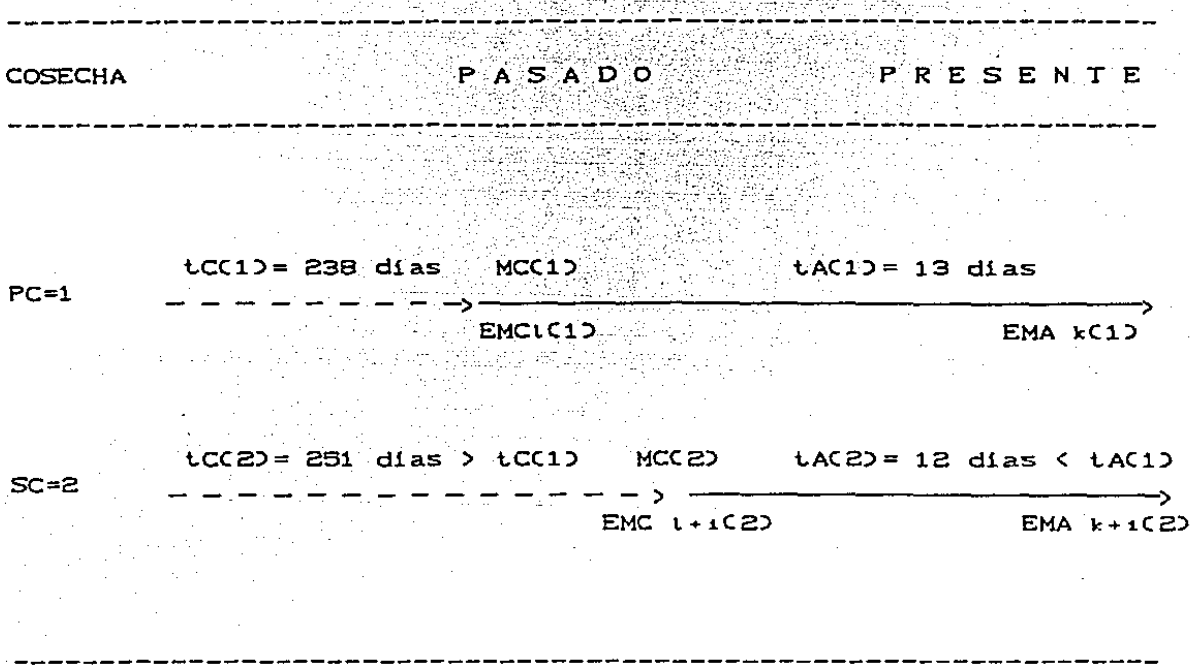
(3) Las curvas respiratorias[6] de la mayoría de los frutos al compararse entre cosechas denotaban EMAs diferentes.

Por los tres argumentos señalados es que se concluyó que el EMA de la SC era mayor que el de la PC y además correspondía al estado de madurez comestible (EMCOM). Por otro lado, como los días de almacenamiento de las cosechas fueron tA(1)=13 y tA(2)=12 fué, entonces que se propuso el patrón de comportamiento que se presenta en la página 43.

Grafica 1 Modelos ajustados para porcentajes de polifenoles.



PATRON DE LA SITUACION RESULTANTE DE LA MEDICION EN CHICOZAPOTE



Esta situación, como puede verificarse, corresponde con un patrón con altas expectativas para determinar el índice de corte, pues los EMC son distintos,y en particular el de la SC induce el EMCOM.

Por otro lado, con el propósito de observar si otro EMC contiguo al EMC de la SC no presenta traslape con él, se incluyó en el análisis discriminante el EMC asociado a la evaluación E9, quien como se comentó solo tenía información al momento del corte.

Considerando estos 3 EMC (E9,PC,SC) se llevó a cabo el análisis de datos, empezando por el diagnóstico o examen de datos, para determinar finalmente si las variables que se proponen para formar el índice de corte, permiten realmente diferenciar los EMC.

3.2 Diagnóstico o examen de datos..

El procesamiento de la información se hizo con el paquete estadistico de cómputo BMDP (Biomedical Package, versión 1987) y en cada análisis realizado se indica la rutina empleada. Inicialmente se procedió al filtrado de la información. Al revisar los datos al momento de corte se encontraron 3 frutos con problemas, en el caso 24 la fracción de cáscara (FC=(PFCC-PFSC)/ PFCC) era negativa , el 27 tenia información faltante en la glucosa verdadera y el 38 un valor de fuerza de separación muy pequeño en relación a su EMC. Para resolver este problema. despues de una discusión con los investigadores del área, se concluyó considerar como faltantes estos datos, los cuales fueron estimados al considerar un ajuste para cada una de las variables con una regresión multiple, considerando todas las variables cuantitativas en estudio. Esto se llevó a cabo utilizando la rutina AM del BMDP.

Con la estimación de los valores mencionados y todos los demás casos, se llevaron a cabo los análisis que a continuación se comentan. Los datos se presentan en el cuadro A del anexo .

En primer lugar se realizaron pruebas de hipótesis multivariadas para estudiar el efecto diferencial de las edades de los árboles dentro de cada uno de los EMC (E9, PC , SC), tanto para las mediciones denominadas de campo como para las de laboratorio al momento de corte , utilizando la rutina 3D . Los resultados se resumen en el cuadro C1.

C1. COMPARACION MULTIVARIADA EN MEDICIONES DE CAMPO Y DE LABORATORIO POR EPOCA DE CORTE DE ACUERDO A LA EDAD DE LOS ARBOLES

MEDICIONES	E9	PC	SC
DE CAMPO	0.9489	0.5691	0.1430
DE LABORATORIO	0.3400	3.0400	0.9253

NOTA: El criterio de prueba fué la estadística F multivariada y la diferencia considerada de árboles de 11 años menos los de 17. Nivel de significancia observado α=0.05.

El cuadro anterior muestra, que por época, no hay diferencias significativas al 5% entre la edad de los árboles, pues los valores reportados de la F indican que la discrepancia entre lo observado en la muestra y lo esperado al considerar que las poblaciones multivariadas son iguales, no es "grande". Por lo anterior se concluyó que la edad no constituye un factor de confusion para la discriminación de los EMC y que el resultado que se obtenga es aplicable al rango de edad de los chicozapotes en estudio.

Enseguida, se realizaron comparaciones de EMC dos a dos de manera multivariada, univariada (paramétrica y no paramétrica) para mediciones de laboratorio y de campo. Los resultados obtenidos del análisis se muestran en los cuadros C2 y C3 respectivamente; éstos indican en general diferencias multivariadas altamente significativas entre los EMC. Para las pruebas univariadas de laboratorio en el cuadro C2 se observa que la sacarosa verdadera

y los azucares totales[9] (AT) aumentan, siendo el aumento mayor de la E9 a la PC que de la PC a la SC. Esto es indicativo de que la PC y la SC han alcanzado su maduración Fisiológica[2] y muy posiblemente su madurez comestible, pues se reporta que los AT aumentan hasta que la madurez comestible se alcanza y dicho aumento cesa en el momento que la madurez comestible se alcanza. Esto es consistente con la disminución de las diferencias ya mencionadas. Asi también se reporta que los AT estan correlacionados positivamente con los azúcares no reductores , esto es, con la sacarosa verdadera.

C2. COMPARACION MULTIVARIADA, UNIVARIADA PARAMETRICA Y NO, ENTRE EPOCAS DE CORTE PARA MEDICIONES DE LABORATORIO.

Análisis	E9-PC		PC-SC		E9-SC	
Multivariado	11.29**		35.22**		26.76**	
Univariado	Par	Nopar	Par	Nopar	Par	Nopar
Glucosa verd	-1.06	0.37	0.92	0.73	1.68	0.16
Fructosa verd	-1.01	0.13	-2.71**	0.00	-3.37**	0.00
Gluc tot verd	-0.28	0.59	-1.50	0.15	-1.35	0.12
Sacarosa verd	-5.33**	0.00	-2.16**	0.03	-7.37**	0.00
Polifenoles	-4.05**	0.00	10.17**	0.00	3.80**	0.00
Azuc totales	-4.83**	0.00	-2.42**	0.02	-6.87**	0.00

NOTA: La estadística reportada para el caso paramétrico fué la t de Student y para el no paramétrico la de Mann-Whitney , los valores reportados para el último caso son niveles de significancia observados.

C3. COMPARACION MULTIVARIADA, UNIVARIADA PARAMETRICA Y NO, ENTRE EPOCAS DE CORTE PARA MEDICIONES DE CAMPO.

	E9-PC		PC-SC		E9-SC	
CASO MULT	7.83**		9.14**		56.27**	
CASO UNIVAR	PAR	NOPAR	PAR	NOPAR	PAR	NOPAR
Diámetro polar	-0.30	0.69	-0.88	0.40	-1.54	0.28
Diámetro ecuatorial	0.35	0.75	-0.77	0.57	-0.53	0.83
Volumen	-0.17	0.93	0.15	0.81	0.03	0.79
% del Fr con cascara	-0.35	0.96	0.04	0.87	-0.27	0.90
% del Fr sin cascara	-0.54	0.87	-0.31	0.80	-0.89	0.78
Grados brix	0.27	0.80	0.33	0.75	0.64	0.71
Fuerza de separación	5.32**	0.00	9.33**	0.00	16.95**	0.00
Látex escurrido	0.72	0.16	4.50**	0.00	11.92**	0.00
# de semillas x Fr	-0.43	0.43	-0.52	0.70	-0.87	0.37
Peso de semillas x Fr	-0.78	0.19	-0.50	0.29	-1.21	0.09
Fracción de cascara	0.63	0.37	10.62**	0.0	11.81**	0.00

NOTA: Las estadísticas empleadas para el caso paramétrico fué la t de Student y para el no paramétrico la de Mann-Whitney. En este último caso los valores reportados son niveles de significancia observados.

En polifenoles también los cambios son significativos, aumentando de E9 a PC y disminuyendo de PC a SC. Finalmente en fructosa verdadera (FV) solo hay un incremento significativo de la PC a la SC. Debe anotarse que la diferencia multivariada más acentuada de la PC a SC con respecto a la de E9 a la PC, es debida por un lado a la disminución tan alta de polifenoles en el primer caso en relación al incremento de E9 a PC y por otro a que la diferencia de FV fué significativa solo de la PC a la SC .

Por otro lado al observar las mediciones de campo en el cuadro C se tiene que la fuerza empleada para separar un fruto (FS), es la única con cambios altamente significativos entre los EMC , siendo notoria una disminución conforme aumenta el tiempo en el corte del fruto , siendo el cambio mayor al pasar de la PC a la SC que de la E9 a la PC .

En cuanto a látex escurrido (LAE) y la fracción de cáscara (FC) tienen una disminución significativa solo de la PC a la SC, esto concuerda con conocimientos que se tienen del área de estudio, pues se sabe que el látex junto con el agua que se encuentran en la cáscara se transfieren a la pulpa del fruto a medida que madurez comestible transcurre . De lo observado se deduce que la FS y el LAE. tienen una relación inversa con el tiempo.

3.3 Aplicación del Análisis Discriminante (AD) y otras técnicas. multivariadas.

A continuación del análisis de diagnóstico se aplicó el AD, donde se consideraron 2 fases :

F1.- Determinación de la función de diferenciación para obtener el índice de cosecha.

F2.- Validación de la función de diferenciación.

Para esto en la fase F1 se utilizó la rutina 7M, determinando los valores de entrada convenientes para realizar una selección de variables automáticamente con el procedimiento paso a paso (STEPWISE), considerando que se diferenciaran de la mejor manera la E9 de la PC y la PC de la SC tanto con variables originales para determinar el índice, como transformadas de ellas, resultando las funciones de clasificación (FC) y de discriminación (FD) muy deficientes. Fué por ello que finalmente se optó por seguir un procedimiento que influyera en la selección de las variables ; para ésto se determinaron los valores asociados a : LEVEL , FORCE y STEP considerando diversas jerarquizaciones de las variables y además que nuevamente se diferenciaran de la mejor manera E9 de la PC y la PC de la SC.

Por otro lado en la fase F2 se consideró tanto la validación "interna","externa" y la "externa-interna" para las FC o las FD. En la primera forma de validación, con la técnica JACKKNIFE y en la segunda denominada también validación cruzada, se consideraron submuestras al azar de diversos porcentajes de la muestra total de frutos, una para análisis y otra para la validación[27,28]. Finalmente la tercer forma de validación, que se refiere a una mezcla de la segunda con la primera, es decir, despues de realizar una validación cruzada se construye la matriz de clasificación JACKKNIFE. A continuación se reportan los resultados de las 2 fases

3.3.1) Funciones discriminantes en variables de campo y su validación estadística.

Los resultados del AD se presentan resumidos en los cuadros C4-C8. En el C4 se muestran los coeficientes asociados a las variables discriminantes que determinan las FDs para formar el índice de corte. Es conveniente aclarar que como las FDs se refieren a nuevas variables formadas como una combinación lineal de las variables discriminantes, es común también denominarlas variables canónicas (VCs). Por esto es que el índice de corte se referirá en lo que sigue con las FDs o las VCs indistintamente.

Por otro lado al evaluar cada una de las FDs o VCs en los frutos muestra observados, se obtienen los datos discriminantes o canónicos. Estos datos o valores canónicos se presentan en el cuadro C6 para los valores promedio por EMC que se reportan en el cuadro C5. En el cuadro C7 se reporta la relevancia de las FDs o VCs en la determinación del índice de corte y finalmente en el C8 se reportan los porcientos de clasificación correcta al considerar los frutos muestra en estudio con la FD propuesta.

C4 : VALORES DE LOS COEFICIENTES DE LAS VARIABLES DISCRIMINANTES QUE DETERMINAN EL INDICE DE CORTE

	COEFICIENTES ESTANDARIZADOS		COEFICIENTES NO ESTANDARIZADOS	
VARIABLES DISCRIMINANTES	FD1 (VC1)	FD2 (VC2)	FD1 (VC1)	FD2 (VC2)
FS	1.901651	-0.613282	0.00400	-0.00129
LAE	1.711834	-2.598318	7.49227	-11.37219
LF	-2.377378	2.858049	-0.00366	0.00440
TERMINO CONSTANTE			6.52972	3.48895

C5 : MEDIAS DE LAS VARIABLES DISCRIMINANTES DEL INDICE DE CORTE

VARIABLES CONSIDERADAS	E P O C A DE CORTE E9	PC	SC	PROMEDIO
FS	2790.00513	1749.47900	313.15790	1817.54736
LAE	0.73684	0.67368	0.28947	0.56667
LF	2034.00513	1300.54000	98.57895	1145.53894

C6. VALORES CANONICOS PROMEDIO POR EMC PARA EL INDICE

EMC	VC1	VC2
E9	2.71805	0.46311
PC	0.74293	-0.68070
SC	-3.46097	0.21059

C7. % DE DIFERENCIACION DE LAS VC PARA EL INDICE Y SU SU CORRELACION CON LOS EMC.

VC	%DIS	CORR. CAN.
VC1	0.96487	0.93548
VC2	0.03513	0.45087

C8. MATRIZ DE CLASIFICACION SIMPLE (Y JACKKNIFE) CON VARIABLES PARA EL INDICE DE CORTE

EMC	% CLASIFICACION CORRECTA	NO. FRUTOS CLASIFICADOS EN EL EMC E9	PC	SC
E9	94.7 (94.7)	18 (18)	1 (1)	0 (0)
PC	94.7 (84.2)	0 (2)	18 (16)	1 (1)
SC	100.0 (100.0)	0 (0)	0 (0)	19 (0)
TOTAL	96.5 (93.0)	18 (20)	19 (17)	20 (20)

En el cuadro C4 puede observarse que la FD1 está determinada por látex escurrido (LAE) , fuerza de separación (FS), en sentido positivo y LF (interacción LAE y FS) en sentido negativo. La contribución de la importancia de cada una de las variables en la FD1 se refiere a la magnitud de los coeficientes discriminantes estandarizados en valor absoluto. La FD1 que se refiere al índice de corte, indica que valores grandes positivos del IDC corresponderan a frutos con desviaciones del promedio mayores y positivas de LAE y FS y desviaciones pequeñas de LF como puede apreciarse en los cuadros C5 y C6; el comportamiento anterior es característico de los frutos de la E9 . Debe recordarse que las variables LAE y FS tienen una relación inversa con el tiempo en que se corta el fruto y entre ellas guardan una relación directa.

Por otro lado cuando el fruto corresponde a un EMC más avanzado puede notarse que el IDC será grande y negativo (ver C6), lo que corresponde a desviaciones del promedio mayores y negativas de LAE y FS , y desviaciones grandes de LF (ver C5) ;dicho comportamiento corresponde al perfil de la SC.

En relación al perfil de la PC este se refiere a diferencias menores y positivas de LAE Y FS y desviaciones pequeñas de LF , lo cual produce valores del IDC positivos y cercanos a cero.

Del análisis anterior se concluye que el IDC guarda una relación inversa con el tiempo en que se corte el fruto.es decir, el LAE y la FS serán menores a medida que que el tiempo en que se corte el fruto sea mayor. Esta conclusión es consistente con los datos reportados en los cuadros C5 y C6.

Por otro lado la FD2, ver cuadro C7, aporta solo el 4% a la discriminación total. Por lo cual se concluye que basta con la FD1 para establecer la discriminación de los EMC.

Así entonces la diferenciación que se tiene con la FD1 para los EMC, debido al comportamiento decreciente de las variables, propicia que sea mayor cuando al comparar dos épocas de corte, éstas sean más distantes.

Ahora bien, observando en detalle el cuadro C8 se puede decir que los pares de EMC mejor diferenciados cuando se utiliza la función de clasificación simple (y la Jackknife) son en orden decreciente los siguientes, de acuerdo al número de frutos mal clasificados:

1) E9 vs SC , ningun (ninguno).

2) PC vs SC , uno de PC clasificado en la SC (2 de PC en SC).

3) E9 vs PC , uno de E9 en la PC (1 de E9 en PC y 2 de PC en E9)

De lo anterior se obtiene que los porcentajes de clasificación correctos para SC, PC y E9 son 100 , 94.7 y 94.7 que corresponden a un 4.5% de mala clasificación en promedio por grupo (el orden fué el mismo y los %s fueron 100,84.2 y 94.7, en tanto el promedio por EMC es de 7%). Los resultados anteriores pueden apreciarse de una manera clara en la gráfica 2 de variables canónicas, mostrada en la siguiente página. Esta gráfica se construye con los datos canónicos para todos los frutos muestra. Podemos concluir, que la SC que se refiere al estado de madurez comestible (EMCOM) tiene pocas posibilidades de confundirse con E9 y PC.

Vale la pena comentar en este momento que en la literatura

estadística [Johnson:5] se reporta que antes de implementar una función de diferenciación se verifiquen los supuestos inherentes al análisis.

En particular en el enfoque de funciones de clasificación , se propone que se verifiquen en el siguiente orden:

i) Normalidad Multivariada.

ii) Igualdad de matriz de varianzas y covarianzas.

Pues la prueba del supuesto ii es afectado fuertemente por la no normalidad. Cuando el supuesto ii se cumple,se recomienda utilizar las FCL y cuando no es así las funciones de clasificación cuadráticas (FCC). Ciertos estudios han reportado que hay casos donde la presencia de no normalidad, al utilizar las FCL producen clasificaciones pobres aun cuando ii se cumple. Y se ha reportado además que el no cumplimiento de la normalidad de las poblaciones es más critico al usar FCC.

Por lo anterior, es importante saber que cuando los datos no siguen una distribución normal multivariada se han reportado dos opciones:

a) Los datos pueden transformarse de modo que se parezcan más a la normal y después probar el supuesto ii. De acuerdo al resultado anterior usar la FCL o la FCC.

b) Usar la FCL o la FCC sin considerar el parecido de distribuciones de las poblaciones y suponer que estamos trabajando razonablemente bien.

Es importe asentar que las FD a diferencia de las FC no dependen del parecido de las distribuciones de las poblaciones, sino solo

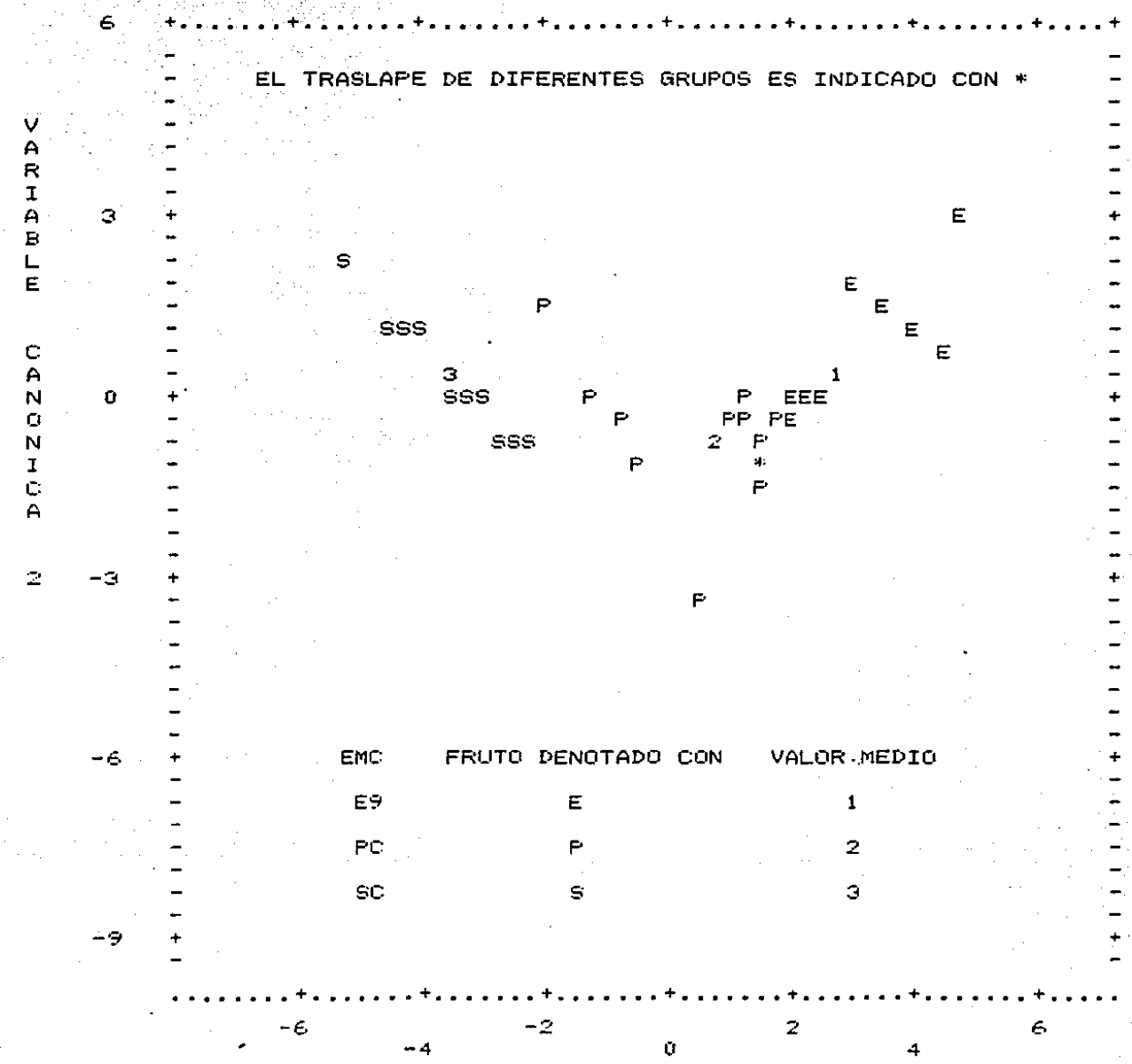
del requerimiento de la estructura de varianzas-covarianzas común.

Debe anotarse que cuando el supuesto ii se cumple y las muestras son razonablementes grandes, el enfoque de FC funciona bien , aunque el supuesto de normalidad de las poblaciones no se cumpla.

En el caso de chicozapote la normalidad de las poblaciones no se cumplio , esto se obtuvo al considerar la estadistica de Anderson Darling . Asimismo en la gráfica 2 se puede apreciar que las matrices de varianzas y covarianzas de los grupos distan mucho de ser comúnes. Sin embargo en muchas aplicaciones, el que no se cumplan los supuestos no invalida las concluciones obtenidas, ya que la validez de cualquiera de los enfoques para la discriminación, también depende fuertemente de que con los datos muestra, si los recursos de cómputo y los tamaños de muestra lo permiten, pueda verificarse la consistencia de la función utilizada, empleando el procedimiento JACKKNIFE y técnicas de validación cruzada, para observar finalmente el porcentaje de mala clasificación.

En el presente estudio se optó fundamentalmente por el análisis con FD, aunque el supuesto de varianzas-covarianzas no se cumplió, se procedio a la validación cruzada que a continuación se presenta y a la validación con criterios del área de estudio, como posteriormente se presentará.

Gráfica 2 Discriminación de los estados de madurez de corte (EMC), con el índice de corte.



Como hemos visto hasta este momento la validación interna (matriz Jackknife) de la regla de diferenciación resultó satisfactoria. En cuanto a la validación externa o cruzada se obtuvo al considerar diversos tamaños de muestra (25 ,50 y 75% : vea cuadro C9 y C10) y de ella se concluye :

1.- Que el porcentaje más bajo de clasificación correcta en promedio por EMC fué de 86.87 , en este caso asociado a la PC cuando la muestra fué de 25% , como puede observarse en el cuadro C10. Esto es consecuencia del traslape que tiene la PC con la E9.

Al comparar los cuadros C8 y C10 se observa que los porcentajes de clasificación correcta reportados con toda la muestra (C8), es a lo mucho un 10% mayor que con cualquier submuestra para la validación , ver C10.

2.- Al comparar los cuadros C7 y C10 puede notarse la gran similitud que tienen los % de variación total explicada por la primer variable canónica y el % de correlación canónica en la muestra total (C7) y los de las submuestras consideradas en la validación (C10) , destacando que tanto en las submuestras como en la muestra total basta solamente la primera variable canónica para realizar la discriminación de los EMC.

Así entonces puede decirse que la FD1 modela convenientemente la diferenciación de los EMC . Podemos concluir hasta este momento que estadísticamente es posible determinar el EMCOM (SC) con el índice de corte que se propone para el propósito de estudio, pues el EMCOM se confunde en una proporción muy baja con los otros EMC estudiados.

C9 Porcentajes y tamaños de muestra para la validación cruzada del índice de corte

Porcentaje de	frutos para Análisis			frutos para validación		
muestra	E9	PC	SC	E9	PC	SC
25	14	14	11	5	5	5
50	8	8	8	11	11	11
75	6	5	5	13	14	14

C10 Porcentajes de clasificación correcta para la validación cruzada del índice de corte

% de	% clasif para Análisis			% clasif para validación		
muestra	E9	PC	SC	E9	PC	SC
25	100.0	92.9	100.0	80.0	80.0	100.0
50	100.0	87.5	100.0	90.9	100.0	100.0
75	100.0	100.0	100.0	92.3	85.7	100.0

NOTA: Los % de variación total y coeficientes de correlación para la primera variable canónica en las muestras para validar del 25,50 y 75 % fueron respectivamente: (96.0, 94.0), (96.5, 93.0) y (98.0, 96.0).

3.3.2) Validación del índice de corte en el área de estudio.

Una vez realizada la discriminacion de los EMC con las variables de tipo práctico, se pasa a un segundo nivel de análisis donde

deben considerarse las otras variables que se midieron en condiciones de campo: las denominadas de apoyo y las de laboratorio , de éstas, las últimas son de interés inmediato pues en trabajos de investigación cotidiana en el área de FYTPDF han permitido realmente diferenciar los estados de madurez de los frutos durante el almacenamiento.

Cuando con las variables prácticas la discriminación es aceptable. parece razonable que inmediatamente se aplique el análisis discriminante a las variables de laboratorio para con ello poder validar el índice de corte obtenido , pues como ya se mencionó con las variables de laboratorio se espera que la discriminación sea también aceptable. De cumplirse lo anterior entonces se procedería a correlacionar las variables que determinan el índice de corte con las de laboratorio para con ello validar el índice, pues a mayor correlación mayor credibilidad se tendrá del índice de corte obtenido.

Una forma de validación más estricta, aunque más cara, es realizarla con variables al final del almacenamiento cuando se obtienen los EMA, es decir, realizar el análisis discriminante con estas variables para luego correlacionarlas con las variables que determinan el índice de corte y con ello medir la credibilidad en éste. Sin embargo en el caso de estudio no se tuvieron las muestras suficientes para abordar el procedimiento.

Cuando por otro lado la discriminación con las variables prácticas no es alta, valdría la pena echar mano de las restantes variables que fueron medidas tambien en condiciones de campo,para

explorar si añadiendo algunas de ellas pudiera tenerse una discriminación alta,aunque de ser así,el conjunto de variables no estaría determinando el índice de corte como ya se ha definido. Sin embargo, esto permitiría reflexionar acerca de la posibilidad de que en experiencias posteriores para la determinación del índice, se incluyan variables prácticas que esten muy relacionadas con estas variables añadidas. Por el uso que se propone dar a estas variables se decidió llamarles variables de apoyo para la determinación del índice de corte.

Para la validación del índice de corte en nuestro caso procedemos como a continuación se indica.

A.-Funciones discriminantes en variables de laboratorio y su validación estadistica.

Como puede observarse en el cuadro C11, la FD1 está determinada por POL,GV y SV en sentido positivo y con AT en sentido negativo; esto indica que los frutos que tengan valores grandes y positivos en la variable canónica serán aquellos que tengan valores mayores de estas 3 primeras variables y menores de la última . Este resultado es consistente al observar los resultados de los cuadros C12 y C13 , cuando se comparan la E9 contra la SC y la PC contra la SC . En cuanto a la FD2, está determinada por POL en sentido negativo y por GV en sentido positivo, veáse cuadro C11.

C11. VALORES DE LOS COEFICIENTES DE LAS VARIABLES DISCRIMINANTES MEDIDAS EN EL LABORATORIO

VARIABLES CONSIDERADAS	ESTANDARIZADAS VC1	ESTANDARIZADAS VC2	NO ESTANDARIZADAS VC1	NO ESTANDARIZADAS VC2
POL	5.030950	-3.405580	0.000000	-0.000000
GV	0.902680	0.990550	0.163000	0.178860
AT	0.831200	-0.119090	0.000000	0.000000
SV	0.60040	-0.204530	0.000000	0.000000
TERMINO CONSTANTE			0.271740	2.637760

C12. MEDIAS DE VARIABLES DISCRIMINANTES MEDIDAS EN LABORATORIO

VARIABLES CONSIDERADAS	E P O C A D E C O R T E E9	PC	SC	PROMEDIO
POL	0.78000	1.06000	0.60000	0.81330
GV	3.28000	3.11000	2.97000	3.12000
AT	10.41000	13.96000	16.10000	13.49000
SV	3.93000	7.38000	9.17000	6.83000

C13. VALORES CANONICOS PROMEDIO POR EMC PARA MEDICIONES EN EL LABORATORIO.

EMC	VC1	VC2
E9	0.85000	1.19000
PC	1.24000	-1.05000
SC	-2.09000	-0.14000

C14. % DE DIFERENCIACION DE LAS VC DE LABORATORIO Y SU CORRELACION CON LOS EMC.

VC	%DIS	CORR. CAN.
VC1	0.72000	0.84000
VC2	0.28000	0.69000

C15. MATRIZ DE CLASIFICACION SIMPLE (Y JACKKNIFE) CON VARIABLES MEDIDAS EN LABORATORIO

EMC	% CLASIFICACION CORRECTA	NO. FRUTOS CLASIFICADOS EN EL EMC		
		E9	PC	SC
E9	84.2 (73.7)	16 (14)	3 (4)	0 (1)
PC	89.5 (84.2)	2 (2)	17 (16)	0 (1)
SC	100.0 (94.7)	0 (1)	0 (0)	19 (18)
TOTAL	91.2 (84.2)	18 (17)	20 (20)	19 (20)

Para la FD2 los frutos que tengan valores grandes corresponden a valores pequeños de POL y valores grandes de GV. Este resultado es consistente y claro al observar la gráfica 3 de las VC1 y VC2, los datos de los cuadros C12 ,C13 y al comparar la E9 contra la PC.

Puede concluirse entonces que la VC1 permite diferenciar la E9 de la SC y la PC de la SC , en tanto para diferenciar la E9 de la PC es indispensable además la VC2. Este resultado se sustenta al observar el cuadro C14 que indica que la VC1 explica el 72% de la discriminación total y su correlación canónica corresponde a un 80% , mientras la explicación de la VC2 es de 28% , con una correlación de 70%.

Ahora bien, del análisis del cuadro C15, se concluye que los pares de grupos mejor diferenciados, de acuerdo con el número de frutos mal clasificados al considerar la matriz de clasificación simple (y la jackknife) son respectivamente :

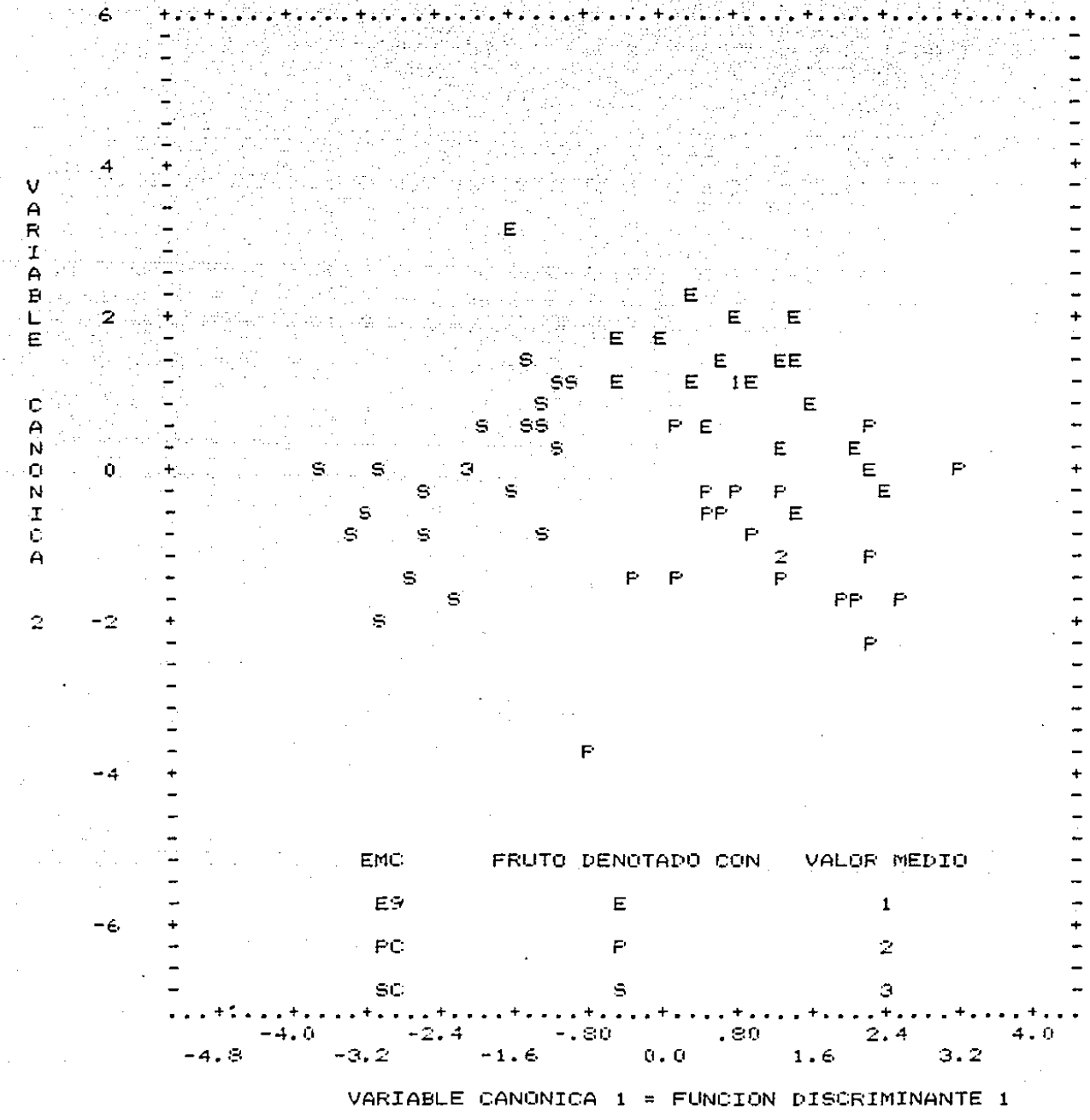
1.-PC VS SC ninguno (1 en la SC).

2.-E9 VS SC ninguno (1 en la E9 y 1 en la SC).

3.-E9 VS PC dos en E9 y tres en la PC (2 en E9 y 4 en la PC).

Lo anterior indica que los grupos mejor clasificados fueron respectivamente SC,PC y E9 con 100,89.5 y 84.2 % de clasificación correcta; en el peor de los casos en la E9 hay un 15% de mala clasificación y en promedio el porcentaje de mala clasificación por grupo es del 9% (en este caso los % de clasificación correcta fueron de 94.7, 84.2 y 73.7 y el peor grupo clasificado fué E9 con 25% ; además en promedio el % de mala clasificación por grupo fué de 15). Las diferencias pueden observarse en la gráfica 3.

Gráfica 3 Discriminación de los estados de madurez de corte, con las variables de laboratorio.



VARIABLE CANONICA 1 = FUNCION DISCRIMINANTE 1

Nuevamente podemos concluir en base a las variables de laboratorio que es poco probable que la SC se confunda con los E9 y la PC. Otra vez al considerar muestras de 25, 50 y 75 % (el número de frutos por EMC fué similar al cuadro 9 del índice de corte) para llevar a cabo la validación cruzada en estas variables, se obtuvo el cuadro C16.

C16 Porcentajes de clasificación correcta para la validación cruzada de las variables de laboratorio.

Porciento de	% clasif para Análisis			% clasif para validación		
muestra	E9	PC	SC	E9	PC	SC
25	87.5	100.0	100.0	90.0	72.7	100.0
50	87.5	75.0	100.0	81.8	90.9	100.0
75	83.3	80.0	100.0	84.6	85.7	100.0

Del análisis del cuadro C16 se tiene que:

1.- El porcentaje de clasificación correcta más bajo en promedio por EMC fué de 87.57 .asociado a la PC cuando la muestra fué de 25% . Ello es la resultante del traslape que tiene con la E9. Al comparar los cuadros C16 y C15 se tiene que los porcentajes de clasificacón correcta de toda la muestra es a lo mucho 10% mayor que cuando se toma dicho porcentaje de muestra para la validación.

Por otro lado la relevancia de cada una de las variables canónicas en la discriminación para la validación cruzada se presentan en el cuadro C17.

C17. PORCENTAJES DE DIFERENCIACION DE LAS VC DE LABORATORIO Y SU CORRELACION CON LOS EMC, PARA LA VALIDACION CRUZADA.

	25 % de muestra		50% de muestra		75% de muestra	
VC	%DIS	CORR. CAN.	%DIS	CORR. CAN.	%DIS	CORR. CAN
VC1	71.0	83.0	84.0	86.0	80.0	88.0
VC2	29.0	68.0	16.0	60.0	10.0	52.0

En C17 puede observarse la gran influencia que tienen las dos variables canónicas para la diferenciación de los grupos.

Así pues, dado que las variables del índice y las de laboratorio discriminaron convenientemente los EMC, se procede a continuación a la validación del índice de corte de acuerdo a el área de estudio , considerando un análisis de correlación canónica entre las variables que constituyen el índice y las de laboratorio.

B.- Correlación canónica entre las variables del índice de corte y las de laboratorio.

Al llevar a cabo la correlación canónica entre el grupo de variables que constituyen el índice de corte (FS,LE y LF) y las de laboratorio (GV,SV,POL,At) con la rutina 6M y se obtuvo un resúmen del análisis en el cuadro C18.

C18. Correlación canónica entre variables del índice y de laboratorio.

Estados de madurez	Correlación	Nivel de significancia
E9	0.58	0.49
PC	0.72	0.12
SC	0.57	0.19
E9 + PC	0.50	0.14
PC + SC	0.96	0.00
E9 + SC	0.83	0.00
E9 + PC + SC	0.74	0.00

NOTA : La correlación se refiere al primer par de variables canónicas. El criterio de prueba para la significancia se refiere a la prueba de Bartlett.

Del cuadro C18 pudiera pensarse que la relación obtenida es satisfactoria,pues a pesar de que dentro de cada estado de madurez la relación no fué significativa, ésta si es considerablemente alta con un nivel de significancia relativamente bajo, cuando se consideran al menos 2 EMC , y cuando se consideran los tres la relación entre las variables es evidente.

La relación entre los grupos de variables con los tres EMC, indica que tanto las variables que forman el índice de cosecha como las variables químicas medidas en el laboratorio, expresan realmente las diferencias entre los EMC, sin embargo la forma de diferenciación no es dependiente como lo indica la ausencia de correlacion entre los dos grupos de variables cuando se considera por EMC. Puede concluirse entonces en base a los resultados mencionados que el índice de corte obtenido, realmente permite determinar el estado de madurez comestible (SC) que se pretendía obtener.

3.4 Reporte del índice de corte

En este momento es de importancia comentar que en el estudio, al momento de corte,se hicieron otro tipo de mediciones en los frutos tales como color de la cáscara y de la pulpa[14] y los días grado[11]. Siendo el color de pulpa similar en las tres épocas de estudio, a diferencia del color de la cáscara que en E9 se asemejaba a un color café de cuero viejo y para la PC y la SC el color era beige. Esta característica fué medida con la carta 7.5 YR de la escala de colores de Munsell que se refiere a matices de amarillo y rojo con cercanía al primero. En cuanto a los días

grado fueron distintos en las tres épocas.

Asi entonces, en base a los resultados obtenidos del análisis de datos, sabemos que los frutos que alcanzan el EMCOM se refieren a la SC y los intervalos de confianza para los promedios poblacionales de FS, LAE y LF se presentan en el cuadro C19.

C19.-Intervalos de confianza para el peso promedio requerido para desprender el fruto de la rama (FS) y la cantidad promedio de látex escurrido. Para la SC (EMC2) que alcanza el EMCOM.

Características	Intervalo de Confianza (α=0.05)
FS	$\mu_{fs} \in$ [267.1 , 359.2]
LAE	$\mu_{lae} \in$ [0.2 , 0.3]

por todo lo mencionado , las reglas para identificar el EMCOM con el indice de corte son :

R1.- Que el color de la cáscara en el fruto sea beige, ello garantiza que el fruto sea de la PC o de la SC.

R2.- Si el fruto se desprende aplicando un peso de 359.2 gramos y la cantidad de LE se encuentra entre 0.2 y 0.3 hay grandes posibilidades de que el fruto sea de la SC y alcance entonces el EMCOM.

Cuando los frutos se asignan al EMCOM, se tiene implicado que ellos tienen aproximadamente 251 días del amarre al momento en que se cortaron y, además tienen acumulados 2569.4 días grado[11].

3.5 Conclusiones

1.- La propuesta de análisis de datos para la determinación del índice de corte en frutales :

a.- Es aplicable tanto a frutos de tipo climatérico como no climatérico, cuando se tienen una o más características en estudio para formarlo.

b.- Induce al investigador a que desde la planeación y recolección de información se apegue realmente a determinar el índice de corte, orientando el costo y el tiempo de investigación.

c.- Da la posibilidad de validar el índice obtenido.

2.- Una limitación del índice de corte para chicozapote, es que no se considero una nueva experiencia en el tiempo ni en el espacio.

3.- En cuanto a las opción de transformar los datos para cumplir el supuesto de normalidad, es conveniente meditar que tanto limita la elaboración de los perfiles de los EMC y paralelamente proceder a sensibilizar a los investigadores de área de FYTPDF para su aceptación.

4.- Considerar en la determinación del índice, el Análisis Discriminante para cuando las variables son cualitativas, o bien cuando son cuantitativas y cualitativas, pues seguramente en otras especies frutales se tendran variables de esta naturaleza.

5.- El EMC correspondiente a los frutos que alcanzan el estado de madurez comestible para el destino determinado en el estudio, bajo las condiciones y tiempo de almacenamiento definidos, es la SC. Los frutos del EMCOM se identifican con el índice de corte determinado en este estudio.

APENDICE

1. Definicion de términos en:

   Fisiologia y Tecnologia Postcosecha de frutas. Y

   Análisis Discriminante.

Definición de términos en Fisiología y Tecnología Postcosecha de frutas

(1) Estado de madurez de corte .- Estado de desarrollo de un fruto en el cual, éste posee las características adecuadas para ser cosechado y destinado a un fin particular.

(2) Maduración .- Existen dos etapas de maduración : Maduración fisiológica , también llamada sazonamiento, es el periodo durante el cual el fruto alcanza las características para evolucionar separado de la la planta que le dió origen. Durante este periodo el fruto alcanza plenitud de crecimiento. Maduración de consumo, periodo durante el cual el fruto , fisiológicamente maduro, evoluciona hasta alcanzar su estado de madurez comestible.

(3) Senecencia .- Periodo donde se presenta una pérdida progresiva de la organizacion celular de un fruto. Es la etapa de envejecimiento del fruto, pues conduce a la muerte de sus tejidos.

(4) Variedad .- Es un grupo de frutos dentro de la misma especie que se distingue de otros grupos dentro de la misma especie por su forma o función; desde el punto de vista genético las variedades se pueden intercruzar libremente.

(5) Especie .- Es un tipo particular de fruto que retiene sus diferencias en otros tipos de igual naturaleza por un periodo de muchas generaciones sucesivas, por ejemplo : la manzana y el melón son especies diferentes que no pueden cruzarse porque se presentan entre ellos barreras naturales.

(6) Frutos de tipo climatérico .- En base al comportamiento que presenta la actividad respiratoria de los frutos, estos se han

clasificado en climatéricos y no climatéricos , los primeros, son aquellos que presentan una elevación en el patrón que describe el comportamiento de sus actividades respiratorias. A esa elevación se le conoce como pico climatérico y marca el paso del fruto a la fase de senecencia o envejecimiento .

La actividad respiratoria es el ritmo o velocidad de respiración de un fruto y puede determinarse midiendo la cantidad de oxígeno consumido por el fruto , o bien , la cantidad de CO_2 desprendido por unidades de peso y tiempo (mg/KgxH). Al comportamiento que presenta la actividad respiratoria de un fruto se le llama patrón respiratorio. A continuación se muestra gráficamente el patrón respiratorio de los frutos climatéricos y no climátericos. En el eje horizontal se grafica el tiempo transcurrido desde el amarre del fruto hasta su muerte y en el vertical la cantidad de CO_2 desprendido :

P a t r o n e s d e R e s p i r a c i ó n

Frutos climatéricos. Fruto no climatéricos.

---

1 = mínimo climatérico, 2 = elevación climatérica, 3 = pico climatérico, 4 = postclimatérico.

(7) Amarre de un fruto .- Se refiere al momento en que el fruto tiene ya una alta probabilidad de permanecencia en el árbol para continuar su crecimiento y llegar a su desarrollo pleno . A este momento le antecede una gran caída o aborto de frutos.

(8) Polifenoles .- Son compuestos fenólicos. El fenol ordinario , ácido fénico o carbólico, es un derivado monohidroxilado (un solo grupo oxhidrilo OH) del benceno. Los polifenoles son compuestos que contienen varios grupos oxhidrilos OH ligados directamente al núcleo bencénico.

(9) Azúcares totales , reductores y no reductores .- Los azúcares reductores totales son la suma de glucosa y fructosa en el fruto. Se llaman así porque tienen la capacidad de reducir o atacar agentes oxidantes en el fruto. La sacarosa del fruto constituye los azúcares no reductores totales. Los azúcares totales del fruto son la suma de los dos tipos de azúcares mencionados.

(10) Látex .- Es un líquido de composición muy compleja del que se obtienen substancias como el caucho y la goma de mascar.

(11) Días Grado .- Es una medida de la influencia que tiene la temperatura ambiente en el crecimiento del fruto . Se define como la cantidad de grados que utiliza un fruto para alcanzar su tamaño final. el cálculo se efectua día a día a partir del amarre del fruto , obteniendo la diferencia entre la temperatura ambiente y una temperatura base que es la mínima suficiente para que exista crecimiento en el fruto. Cuando esta diferencia es positiva se reporta como el número de días grado de esa fecha, de lo contrario, dicho número se reporta como cero. Los días grado se van acumulando

con los del día anterior.

(12) Evaluaciones sensoriales .- Son una serie de evaluaciones que se valen de los sentidos de la vista, el olfato, el gusto, el tacto y algunas veces el oído para juzgar la calidad de un fruto . El resultado obtenido de estas evaluaciones debe ser representativo.

(13) Escala Hedónica .- Es una escala utilizada en los análisis sensoriales. Va del uno al nueve de la siguiente manera :

1 = disgusta extremadamente.	2 = disgusta mucho.
3 = disgusta moderadamente.	4 = disgusta ligeramente.
5 = ni gusta ni disgusta.	6 = gusta ligeramente.
7 = gusta moderadamente.	8 = gusta mucho.

9 = gusta extremadamente.

(14) Color externo de la cáscara y de la pulpa del fruto.- Para su medición se utilizan las escalas de colores de Munsell y Pantone. Estas escalas se refieren a un conjunto de tablas que contienen una clasificación de colores.

(14') Método de Foling-Denis.- Es un método colorimétrico usado para medir el contenido de polifenoles en un fruto. Los métodos colorimétricos son métodos de análisis químicos para determinar cuantitativamente sustancias químicas por comparación de la intensidad del color producido por ellas con un reactivo determinado y la intensidad de color producido con el mismo reactivo con una cantidad conocida de la misma sustancia.

(14'') Método de Ting V.- Es un método colorimétrico en el cual , apartir de ciertos reactivos y mediante calibración es posible medir el contenido de azúcares en un fruto.

(14'''). - Acido absicico.- Es un ácido contenido en la capa de absición o capa de desprendimiento del fruto.

Definición de términos en Análisis Discriminante.

Glosario :

FC .- Función de clasificación.
FD .- Función discriminante (= VC .- variable canónica).
VPD .- Variable potencialmente discriminante.
VD_2 .- Variable discriminante o predictora.
DM^2 .- Distancia de Mahalanobis al cuadrado.

Definiciones :

(15) Dato Multivariado.-Es un conjunto de características de un individuo que se expresa en un vector. También se cita usualmente como observación multivariada.

(16) Variables potencialmente discriminantes (VPD).- Se refiere a las variables con las cuales se pretende realizar la diferenciación entre grupos. De éstas se seleccionan las variables que permitirán la clasificación o discriminación de los individuos a los grupos; a las variables seleccionadas se les denomina variables predictoras o discriminantes (VD).

(17) Procedimiento de selección de variables a pasos (Stepwise).- Es un procedimiento empleado para seleccionar las VD a partir de las VPD (vease capítulo 2).

(18) Probabilidad a priori de pertenencia.- Es parte de la información con que se cuenta al llevar a cabo un análisis discriminante, adicional a la información muestral, y se refiere a a la probabilidad de pertenencia de un individuo a cada uno de los grupos en estudio.

(19) Probabilidad a posteriori de pertenencia.- Es la probabilidad que tiene un individuo de pertenecer a un grupo particular , tomando en cuenta su información muestral y las probabilidades a

priori de pertenencia para cada grupo.

(20) Función de Clasificación (FC).- Se determina una por grupo y los valores obtenidos cuando se evalua la función en un dato multivariado son usados para calcular las probabilidades a posteriori de pertenencia a los grupos. A partir de esto, un individuo se asigna al grupo en el cual se haya obtenido el máximo valor de las FC o equivalentemente la mayor probabilidad a posteriori estimada.

(21) Matriz de clasificación.- También referida como matriz de asignación o de predicción. Contiene conteos que indican la habilidad predictiva de las VD. Cada individuo de cada uno de los grupos predichos es clasificado en uno de esos grupos de acuerdo a su FC o FD. Los números de la diagonal se refieren al número de individuos clasificados correctamente y los de fuera los incorrectos. En los márgenes de dicha matriz se reportan el número y porciento de individuos bien clasificados , tanto por grupo como en su totalidad.

(22) Matriz de clasificación Jackknife.- Similar a la matriz de clasificación , solo que cada individuo es clasificado en la matriz , de acuerdo a la FC que es obtenida cuando es omitido el individuo que se pretende clasificar. Véase muestra de análisis y muestra para la validación (28).

(23) Función Discriminante (FD).- Es una función lineal de las variables discriminantes, también llamada variable canónica (VC). Para la diferenciación de los grupos puede tenerse más de una VC, sin embargo en la práctica se prefiere una o a lo mucho dos

para separar los grupos en estudio tanto como sea posible y poder elaborar sus perfiles.

(24) Dato canónico.- Es el valor obtenido en un individuo a partir de una particular FD (o VC) y se denomina dato canónico univariado (DC). El número de datos de este tipo por individuo, se refiere al número de FD consideradas necesarias para la discriminación y forma un dato multivariado canónico por individuo. Los datos canónicos son estandarizados, cuando se refieren a observaciones estandarizadas. En este caso se dice que la FD está estandarizada. La estandarización de los datos es conveniente, cuando las VD son medidas en distintas unidades.

(25) Distancia de Mahalanobis al cuadrado (DM^2).- Estadística propuesta para medir la distancia entre dos poblaciones multivariadas con el mismo número de variables, para lo cual considera el vector de medias de ellas (μ_g) y la matriz de varianzas-covarianzas iguales ($\Sigma_g=\Sigma$ para toda g) en la expresión:

$$DM^2(g,g') = (\mu_g-\mu_{g'})' \Sigma^{-1}(\mu_g-\mu_{g'})$$

Cuando las poblaciones son normales multivariadas, la DM^2 se estima con las muestras de las poblaciones y sirve para probar la igualdad de medias de cada par de grupos, de manera multivariada.

También con ella, en el enfoque de FC se estima la distancia que tiene un individuo ($\underline{x}o$) con cada uno de los grupos, como sigue:

$$\hat{DM}^2(\underline{x}o,g) = \hat{DM}^2g(\underline{x}o) = (\underline{x}o-\bar{\underline{X}}g.)' \; S_p^{-1} \; (\underline{x}o-\bar{\underline{X}}g.) \qquad g=1,..,m$$

y entonces un individuo se clasifica en el grupo con el cual diste menos. Para el enfoque de FD se utiliza en forma similar a la comentada para las FC, solo que considerando para cada individuo

sus datos canónicos .

(26) Centroide .-Se refiere a la media de un conjunto de datos cuando tenemos una variable. Cuando se habla de observaciones multivariadas, el centroide se refiere al vector que contiene como coordenadas a cada uno de los promedios de las variables. Así en el AD es común referirse al centroide para un grupo o para toda la muestra, tanto para datos multivariados originales como canónicos. En general este concepto se utiliza cuando se calcula la DM^2, pues dicha distancia para un grupo particular se refiere a la distancia que existe entre la información multivariada de un individuo y el centroide del grupo.

(27) Validación de una regla de discriminación.- Vease 29.

(28) Muestra de análisis y muestra para la validación.- Cuando se pretende validar una regla de discriminación, se emplea el procedimiento de validación cruzada. Esto consiste en dividir la muestra en estudio por grupo en dos partes (generalmente de manera aleatoria) , una parte para determinar la regla y la otra para validarla. La primera muestra es referida como muestra para el análisis y la segunda para la validación.

(29) Peso o coeficiente discriminante de una VD .- Es el peso o coeficiente asociado a las VD en la FD (o VC); su tamaño está determinado por la estructura de las VD. Los pesos discriminantes grandes se refieren generalmente a las VD que tienen gran poder en la discriminación , lo cual permite establecer los perfiles de los grupos.

(30) Poder Discriminante de una FD o Correlación Canónica.-Mide la

correlación lineal simple entre una combinacion lineal de las VD y otra de las variables indicadoras de pertenencia a cada grupo. A mayor poder o correlación mejor discriminación de la FDC o VC).

(31) Porciento de diferenciación o explicación de la variación de una FD.- Medida asociada al poder discriminante de una FD, se refiere al porciento de discriminación o variación que una particular FD aporta con respecto a la discriminación o variación total producida por las VDs.

(32) Plano discriminante principal.- Es el espacio generado por las dos primeras variables canónicas y en el se representan los individuos formando diversas nubes de puntos. En éste plano un nuevo individuo relacionado con los grupos en estudio es asignado al grupo cuya nube de puntos sea más próxima.

(33) Mapa Territorial.- Este mapa se forma en el enfoque de FD. Es una gráfica donde estan delimitadas las regiones para asignar "nuevos" individuos a los grupos estudiados. Cuando se tienen las dos primeras VC (FD) se refiere al plano discriminante principal.

A P E N D I C E

2. Análisis Discriminante.

2. Análisis Discriminante

Imaginemos una muestra de n frutos, de una especie frutal , que se sabe provienen de m distintos estados de madurez de corte (EMCg : g=1,2,...,m ; $\sum_g n_g = n$, con ng=número de frutos del EMCg). Y donde cada fruto tiene asociado un conjunto de q mediciones $\underline{x}=(x_1\ x_2\ \ldots\ x_q)'$; si el fruto pertenece al g-ésimo EMC entonces $\underline{x}$ se distribuye según una función de densidad $f_g(\underline{x})$.

Bajo el contexto anterior, los problemas que se plantean en el Análisis Discriminante (AD) son:

A.- Construir una regla que diferencie tanto como sea posible los m EMC. Y

B.- Con la regla asignar " nuevos frutos " (ajenos a la construcción de la regla) en alguno de los m EMC.

Para la solución de estos problemas considere la siguiente:

2.1 Notación

Representemos las q mediciones del fruto i en el EMCg como un vector columna $\underline{X}_{gi}=(X_{gi1}\ X_{gi2}\ \ldots\ X_{giq})'$, donde X_{gij} es la medición de la característica j en el fruto i del EMCg (g=1,2,...,m ; i=1,2,...,ng ; j=1,2,...,q). De aquí en adelante a $\underline{X}_{gi}$ le llamaremos simplemente vector y a su transpuesto ($\underline{X}_{gi}'$) vector renglón. Los vectores para una muestra de ng frutos del EMCg son $\underline{X}_{g1}$, $\underline{X}_{g2}$, ..., $\underline{X}_{gng}$ y las variables asociadas a las mediciones de cada vector están correlacionadas en mayor o menor grado.

Considerando lo anterior es que a los datos muestra podemos denotarlos en forma matricial como :

$$\underset{n\times q}{\underline{X}'} = \begin{bmatrix} X_{111} & X_{112} & \dots & X_{11q} \\ X_{121} & X_{122} & \dots & X_{12q} \\ \vdots & \vdots & & \vdots \\ X_{1n_1 1} & X_{1n_1 2} & \dots & X_{1n_1 q} \\ \vdots & \vdots & & \vdots \\ X_{m11} & X_{m12} & \dots & X_{m1q} \\ X_{m21} & X_{m22} & \dots & X_{m2q} \\ \vdots & \vdots & & \vdots \\ X_{mn_m 1} & X_{mn_m 2} & \dots & X_{mn_m q} \end{bmatrix} = \begin{bmatrix} \underline{X}_1' \\ \vdots \\ \underline{X}_m' \end{bmatrix} = \begin{bmatrix} \underline{X}_{11}' \\ \underline{X}_{12}' \\ \vdots \\ \underline{X}_{1n_1}' \\ \vdots \\ \underline{X}_{m1}' \\ \underline{X}_{m2}' \\ \vdots \\ \underline{X}_{mn_m}' \end{bmatrix} = \begin{matrix} \\ EMC_1 \\ \\ \\ \\ EMC_m \\ \\ \end{matrix} \quad \dots (1)$$

Donde $\underline{X}'$= matriz de suministro de datos para el procesamiento, de dimensiones n x q, con :

$\underline{X}_g'$ = matriz de dimensiones n_g x q para el EMC_g y donde :

$$\underline{X}_g' = \begin{bmatrix} \underline{X}_{g1}' \\ \underline{X}_{g2}' \\ \vdots \\ \underline{X}_{gng}' \end{bmatrix}$$

Por otro lado, cuando se reportan las expresiones para realizar las estimaciones de los parámetros se considera la matriz transpuesta de suministro de datos $\underline{X}_{q\times n}$ y se definen además:

$\overline{\underline{X}}_{g.}' = (\overline{X}_{g.1}\ \overline{X}_{g.2}\ \dots \overline{X}_{g.q})$.

como el vector de promedios de las q características en frutos del EMC_g ($\overline{X}_{g.j}$=promedio de

la característica j en el EMC_g). Y

$\bar{\underline{X}}_{..} = (\bar{X}_{..1}\ \bar{X}_{..2}\ \ldots\ \bar{X}_{..q})'$

como el vector de promedios de las q características ($\bar{X}_{..j}$=promedio de la característica j considerando todas las observaciones muestra).

De lo comentado puede decirse que $\underline{X}'_{nxq}$ contiene como vectores la información de cada variable para todos los frutos. Y en $\underline{X}_{qxn}$ cada vector contiene toda la información de un fruto.

2.2 Enfoques para el Análisis Discriminante

Para la solución de los problemas que se plantean en el AD, dos enfoques, ambos debidos a Fisher , son reportados en la literatura estadística:

1.- Un enfoque de Funciones de Clasificación para derivar reglas que puedan usarse óptimamente en la asignación de nuevos frutos a los EMC predichos.

2.- Un enfoque que emplea la idea de maximizar la razón de suma de cuadrados, entre grupos / dentro de grupos, y que produce las variables canónicas también , denominadas funciones discriminantes, las cuales permiten separar los distintos EMC y describir la diferencia entre ellos, de manera gráfica.

En este apéndice despues de presentar los dos enfoques, se mostrará que bajo ciertas condiciones una función que separa diversas poblaciones puede servir como una regla de asignación para ellas y viceversa.

2.2.1 Funciones de Clasificación.

Para este enfoque se comentará en detalle el caso de clasificación o asignación para dos poblaciones, la generalización a más de dos poblaciones solo se referirá a los criterios para clasificar los frutos a los EMC.

Cuando tenemos dos EMC, se pretende determinar dos regiones , $\mathcal{R}_1$ y $\mathcal{R}_2$, tales que si un fruto cae en $\mathcal{R}_1$ lo asignamos al EMC_1 y si cae en $\mathcal{R}_2$ lo asignamos al EMC_2. Esta regla de clasificación no está exenta de error, pues muchas de las veces no hay una clara distincion entre los frutos de los dos EMC con las q variables en estudio, y entonces es posible que un fruto del EMC_1 sea asignado erróneamente en el EMC_2 y viceversa.

Una buena regla de clasificacion es aquella que produce pocas asignaciones erróneas, tomando en cuenta :

a.- Que los frutos de un EMC pueden tener mayor probabilidad a priori de ocurrencia[10].

b.- Que la asignación errónea de un fruto a uno de los EMC es de consecuencias mayores (costo de clasificación errónea).

Una regla de clasificación óptima debe tomar en cuenta los puntos anteriores y además el conocimiento que se tenga de la familia a la que pertenecen las funciones de densidad .

La solución a los problemas que plantea el AD varian de acuerdo a si se conocen o no: las probabilidades a priori, el costo asociado a una asignación errónea y la familia de las densidades [Seber:17].

La solución consiste en elegir una regla de decisión que

permita hacer una asignación óptima de acuerdo a un criterio predeterminado. El criterio más utilizado es aquel que hace mínimo el costo esperado de mala clasificación (CEMCL). Para definir dicho costo considere:

$f_1(\underline{x}\ ;\ \mu_1, \Sigma_1)$ y $f_2(\underline{x}\ ;\mu_2, \Sigma_2)$ funciones de densidad multivariada asociadas con el vector de variables aleatorias $\underline{x}$ para los EMC$_1$(EMC$_2$),con medias $\mu_1(\mu_2)$ y matriz de varianza covarianza $\Sigma_1(\Sigma_2)$.

$C(g'/g)$= costo de clasificar un fruto del EMC$_g$, equivocadamente en el EMC$_{g'}$.

P_g= probabilidad a priori de pertenencia de un fruto al EMC$_g$.

$\mathcal{R}=(\mathcal{R}_1\ ,\ \mathcal{R}_2)$ una partición del espacio muestral en $\mathcal{R}^q$, tal que si un fruto cae en $\mathcal{R}_g$ ($\underline{x}\in\mathcal{R}_g$), es asignado al EMC$_g$. Con $\mathcal{R}_1$ y $\mathcal{R}_2$ ajenos y exhaustivos.

$P(g'/g,\ \mathcal{R}) = \int_{\mathcal{R}_{g'}} f_g(\underline{x})\,d\underline{x} = P(\underline{x}\in R_{g'}/EMC_g)$= probabilidad de asignar un fruto en el EMC$_{g'}$, dado que pertenece al EMC$_g$ y se considera la partición $\mathcal{R}$.

$P_g\ P(g'/g,\mathcal{R})$ = Probabilidad de que un fruto sea del EMC$_g$ y se clasifique erróneamente en el EMC$_{g'}$.

Así entonces el CEMCL se expresa como:

$$CEMCL = C(2/1)\ P_1\ P(2/1) + C(1/2)\ P_2\ P(1/2) \qquad \ldots(2)$$

En este caso se demuestra [Anderson T. W : 16] que las regiones $\mathcal{R}_1$ y $\mathcal{R}_2$ que definen la partición de $\mathcal{R}$ óptima son :

$$\mathcal{R}_1 = \left(\underline{x} \mid \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq K\right) \ ;\ \mathcal{R}_2 = \left(\underline{x} \mid \frac{f_1(\underline{x})}{f_2(\underline{x})} < K\right) \qquad \ldots(3)$$

$$\text{con } K = -\frac{C(1/2)}{C(2/1)} \frac{P_2}{P_1}$$

Si los costos de mala clasificación son iguales, la regla de decisión es: asigne $\underline{x}$ al EMC$_i$ si

$$P_i f_i(\underline{x}) = \max_{g=1,2} (P_g f_g(\underline{x})) \qquad ...(4)$$

o bien al EMCi si es donde $\underline{x}$ tiene la probabilidad a posteriori de pertenencia mayor[19], dado que por el teorema de bayes dicha probabilidad para el EMC$_i$ es:

$$P(EMC_i / \underline{x}) = P_i f_i(\underline{x}) / \Sigma_g P_g f_g(\underline{x}) \quad (g=1,2) \qquad ...(5)$$

Si además las densidades que caracterizan a los EMC son normales, puede mostrarse que $\mathcal{R}_1$ se define como a continuación se indica, al calcular los logaritmos naturales en (3):

$$\mathcal{R}_1=\{\underline{x} \mid DM_1^2(\underline{x})+\ln(|\Sigma_1|)-2\ln(P_1) \leq DM_2^2(\underline{x})+\ln(|\Sigma_2|)-2\ln(P_2) \} \qquad ...(6)$$

donde $DM_g^2(\underline{x})=(\underline{x}-\mu_g)'\Sigma^{-1}(\underline{x}-\mu_g)=(\mu_g-\underline{x})'\Sigma^{-1}(\mu_g-\underline{x})$

se refiere a la distancia que existe entre el fruto por clasificar con información $\underline{x}$ y la media poblacional μ_g del EMC$_g$, y se denomina distancia de Mahalanobis al cuadrado[25] de un fruto con respecto al EMCg.

La region $\mathcal{R}_1$ puede expresarse también como:

$$\mathcal{R}_1=\{\underline{x} \mid Q_{1-2}(\underline{x}) \geq \ln(P_2/P_1) \} \qquad ...(7)$$

donde $Q_{1-2}(\underline{x}) = -1/2\ \underline{x}'(\Sigma_1^{-1}-\Sigma_2^{-1})\underline{x} + \underline{x}'(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2)$
$- 1/2\ (\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2) + 1/2 \ln(|\Sigma_2|/|\Sigma_1|) = FC_{1-2}(\underline{x})$

es la función de clasificación cuadrática de dos EMC.

Ahora bien, si las matrices de varianzas-covarianzas son iguales para cada EMC ($\Sigma_1=\Sigma_2=\Sigma$) puede mostrarse que R_1 se define como:

$$R_1=\{ \underline{x} : \underline{x}'(\Sigma^{-1}\mu_1-\Sigma^{-1}\mu_2) - 1/2(\mu_1'\Sigma^{-1}\mu_1-\mu_2' \Sigma^{-1}\mu_2) \geq \ln(P_2/P_1) \} \quad \ldots(8)$$

al hacer un desarrollo algebraico sobre la parte izquierda de la desigualdad se obtiene :

$$(\mu_1' \Sigma^{-1}\underline{x}- \frac{1}{2} \mu_1' \Sigma^{-1}\mu_1) - (\mu_2' \Sigma^{-1}\underline{x}- \frac{1}{2} \mu_2' \Sigma^{-1}\mu_2) \quad \ldots(9)$$

y al reagrupar términos se tiene las expresiones siguientes:

$$= (\mu_1 -\mu_2)' \Sigma^{-1}\underline{x} - 1/2 (\mu_1' \Sigma^{-1} \mu_1 - \mu_2' \Sigma^{-1} \mu_2.)$$

$$= (\mu_1 -\mu_2)'\Sigma^{-1} \underline{x} -1/2 (\mu_1 -\mu_2)' \Sigma^{-1} (\mu_1 +\mu_2)$$

$$= (\mu_1 -\mu_2)'\Sigma^{-1} [\underline{x} - 1/2 (\mu_1 +\mu_2)]$$

si $\quad \delta_{12} = \mu_1-\mu_2 \quad$ y $\quad \mu=(\mu_1 + \mu_2)/2 \quad$ entonces

$$= \delta_{12}' \Sigma_p^{-1} [\underline{x} - \mu]$$

Por otro lado de (6), un fruto se clasifica en R_1 si:

$$[DM^2_1 (\underline{x})- 2\ln (P_1)] -[DM^2_2 (\underline{x})-2\ln(P_2)] < 0$$

al hacer un desarrollo algebraico y considerar, $\Sigma_1=\Sigma_2=\Sigma$ se obtiene :

$$[\mu_1'\Sigma^{-1}\underline{x}- \frac{1}{2}\mu_1'\Sigma^{-1}\mu_1+ 2\ln(P_1)]-[\mu_2' \Sigma^{-1}\underline{x}- \frac{1}{2}\mu_2'\Sigma^{-1}\mu_2+2\ln(P_2)] > 0 \quad \ldots(10)$$

donde $\mu_g\Sigma^{-1}\underline{x} - 1/2 \mu_g\Sigma^{-1}\mu_g + \ln(P_g)= FCL_g(\underline{x})$ $\quad \ldots(11)$

se refiere a la función de clasificación lineal para el EMC_g.

Al considerar (11), (10) puede reescribirse como:

$$FCL_1(\underline{x}) - FCL_2(\underline{x}) = (\mu_1-\mu_2)'\Sigma^{-1}\underline{x} - \frac{1}{2}(\mu_1\Sigma^{-1}\mu_1 - \mu_2\Sigma^{-1}\mu_2) + \ln(P_1/P_2) > 0 \quad ...(12)$$

$$= FCL_{1-2}(\underline{x}) > 0$$

y se refiere a la función de clasificación lineal de la diferencia de dos EMC.

En este caso que se conoce la forma de las densidades, (normales en nuestro caso) pero se desconocen los parámetros μ_g y Σ_g, el procedimiento usual es sustituir éstos por sus estimadores. Los estimadores respectivos son:

$$\bar{\underline{X}}_{g.} = \frac{1}{n_g} \sum_i \underline{X}_{gi} = (\bar{X}_{g.1} \quad \bar{X}_{g.2} \ldots \bar{X}_{g.q})' = \underline{X}_g \, 1_* / n_g \qquad ...(13)$$

donde 1_* = vector de unos con dimensiones $n_g x 1$.

En cuanto al estimador de Σ_g, se refiere a S_g, una matriz de dimensiones $q \times q$:

$$S_g = \begin{bmatrix} S_g(1\,1) & S_g(1\,2) & \ldots & S_g(1\,q) \\ \vdots & \vdots & \ldots & \vdots \\ S_g(q\,1) & S_g(q\,2) & \ldots & S_g(q\,q) \end{bmatrix} = [\, S_g(j\,j') \,] \qquad ...(14)$$

Y donde cada término de la matriz se calcula como:

$$S_g(j\,j') = \frac{1}{n_g - 1} \sum_i (X_{gij} - \bar{X}_{g.j})(X_{gij'} - \bar{X}_{g.j}) = \frac{D_g(j\,j')}{n_g - 1} \qquad ...(15)$$

con $D_g(j\,j') = \sum_i X_{gij} X_{gij'} - \frac{X_{g.j} X_{g.j'}}{n_g}$ si $j \neq j'$ y

$D_g(j\,j) = \sum_i X^2_{gij} - X^2_{g.j} / n_g$ si $j = j'$

Además S_g puede escribirse como:

$$S_g = [\, S_g(j\ j')\,] = \frac{1}{n_g - 1}\ [D_g(j\ j')] = \frac{Q_g}{n_g - 1} \qquad \ldots(16)$$

donde $Q_g = [\, D_g(j\ j')\,]$ = matriz de suma de cuadrados y productos cruzados dentro del EMC_g.

De manera matricial Q_g puede expresarse como:

$$Q_g = \sum_i (\underline{X}_{gi} - \overline{\underline{X}}_{g.})(\underline{X}_{gi} - \overline{\underline{X}}_{g.})' \qquad \ldots(17)$$

o en forma matricial compacta a partir de las desviaciones de los datos multivariados con respecto al vector de promedios del grupo como :

$$Q_g = \tilde{\underline{X}}_g\ \tilde{\underline{X}}_g' \qquad \ldots(18)$$

donde $\tilde{\underline{X}}_g = (\underline{X}_{g1} - \overline{\underline{X}}_{g.}\ \ \underline{X}_{g2} - \overline{\underline{X}}_{g.}\ \ \ldots\ \ \underline{X}_{gng} - \overline{\underline{X}}_{g.})$ se refiere a las desviaciones de los datos multivariados con respecto al vector de promedios del grupo (datos centrados) y que pueden expresarse como:

$$\tilde{\underline{X}}_g = (\underline{X}_{g1}\ \underline{X}_{g2}\ \ldots\ \underline{X}_{gng}) - (\overline{\underline{X}}_{g.}\ \ \overline{\underline{X}}_{g.}\ \ \ldots\ \ \overline{\underline{X}}_{g.}) \qquad \ldots(19)$$

$$= (\underline{X}_g - \frac{1}{n_g}\ \underline{X}_g\ 1_* 1_*')(\underline{X}_g - \frac{1}{n_g}\ \underline{X}_g\ 1_{ngxng}) = \underline{X}_g\ (I_{ngxng} - 1_{ngxng}/n_g)$$

donde 1_{ngxng} = es una matriz de unos de dimensiones ngxng.

Al considerar (19) en (18) se obtiene finalmente:

$$Q_g = \underline{X}_g\ (I_{ngxng} - 1_{ngxng}/n_g)\ \underline{X}_g' \qquad \ldots(20)$$

Por otro lado la matriz de correlación $\rho_g = [\, \rho_g(j\ j')\,]$ con : $\rho_g(j\ j') = \sigma_g(j\ j') / [\, \sigma_g(j\ j)\ \sigma_g(j'j')\,]^{1/2}$ es estimada por

$$R_g(j\ j') = S_g(j\ j') / [\, S_g(j\ j)\ \ S_g(j'j')\,]^{1/2} \qquad \ldots(21)$$

Una relación de interés entre la correlación y Sg en el EMCg es:

$$R_g = DIAG(S_g)^{-1/2} \; S_g \; DIAG(S_g)^{-1/2} \qquad \ldots(22)$$

y $\quad S_g = DIAG(S_g)^{1/2} \; R_g \; DIAG(S_g)^{1/2}$ donde

DIAG(Sg)=matriz diagonal con elementos [Sg(1 1) Sg(2 2)..Sg(q q)] de dimensiones qxq.

Por lo anterior es que a Rg se le denomina matriz de Sp estandarizada, pues se calcula como Sg solo que con datos estandarizados.

Por otro lado cuando se tienen m EMC y al suponer que cada uno de ellos proviene de una función de densidad $N_q(\mu_g, \Sigma_g)$, donde $\Sigma_1=\Sigma_2=\ldots=\Sigma_m$ (= Σ). El estimador para Σ es , Sp:

$$\hat{\Sigma} = S_p = \frac{(n_1-1)S_1+(n_2-1)S_2+\ldots+(n_m-1)S_m}{n_1+n_2+\ldots+n_m-m} \qquad \ldots(23)$$

$$= \frac{Q_1+Q_2+\ldots+Q_m}{n-m} \quad , \quad \text{con}$$

Sp=Estimador ponderado de la matriz de varianzas y covarianzas.

Asi entonces un estimador de la distancia al cuadrado de Mahalanobis para el EMCg es:

$$\hat{DM}^2_g(\underline{x}) = (\bar{\underline{x}}_g - \underline{x})' S_p^{-1} (\bar{\underline{x}}_g - \underline{x}) \qquad \ldots(24)$$

Y en base a la $\hat{DM}^2_g(\underline{x})$ también puede calcularse , para el caso tratado, la probabilidad de pertenencia a posteriori estimada[19] para el EMCg , como :

$$\hat{P}(EMC_g/\underline{x}) = P_g \exp(-\frac{1}{2}\hat{DM}^2_g(\underline{x})) / \sum_g P_g \exp(-\frac{1}{2}\hat{DM}^2_g(\underline{x})) \qquad \ldots(25)$$

Y entonces un fruto es clasificado en el EMCg, si en éste se produjo la probabilidad de pertenencia a posteriori mayor.

Asimismo estimadores para la función de clasificación lineal y la de diferencias de EMC son respectivamente:

$$\hat{FCL}_g(\underline{x}) = \bar{\underline{X}}_g' \, S_p^{-1} \, \underline{x} - \frac{1}{2} \bar{\underline{X}}_g' \, S_p^{-1} \, \bar{\underline{X}}_g + \ln(P_g) \qquad \ldots(26)$$

donde $\bar{\underline{X}}_g' \, S_p^{-1}$ = a los coeficientes estimados de la función de clasificación, los otros términos estiman la constante. Y

$$\hat{FC}_{1-2}(\underline{x}) = (\bar{\underline{X}}_1 - \bar{\underline{X}}_2)' S_p^{-1} \underline{x} - \frac{1}{2} (\bar{\underline{X}}_1' S_p^{-1} \bar{\underline{X}}_1 - \bar{\underline{X}}_2' S_p^{-1} \bar{\underline{X}}_2) + \ln(P_1/P_2) \qquad \ldots(27)$$

$$= (\bar{\underline{X}}_1 - \bar{\underline{X}}_2)' S_p^{-1} \underline{x} - 1/2 (\bar{\underline{X}}_1 - \bar{\underline{X}}_2)' S_p^{-1} (\bar{\underline{X}}_1 + \bar{\underline{X}}_2) + \ln(P_1/P_2)$$

$$= (\bar{\underline{X}}_1 - \bar{\underline{X}}_2)' S_p^{-1} \left[\underline{x} - \frac{1}{2} (\bar{\underline{X}}_1 + \bar{\underline{X}}_2) \right] + \ln(P_1/P_2)$$

sea $\hat{\delta}_{12} = \bar{\underline{X}}_1 - \bar{\underline{X}}_2$ y $\bar{\underline{X}}_{..} = (\bar{\underline{X}}_1 + \bar{\underline{X}}_2)/2$ entonces

$$\hat{FC}_{1-2}(\underline{x}) = \hat{\delta}_{12}' \, S_p^{-1} \left[\underline{x} - \bar{\underline{X}}_{..} \right] + \ln(P_1/P_2)$$

En particular si en los m EMC las densidades son normales y son iguales las matrices de varianzas-covarianzas e iguales los costos de clasificación errónea; si se pretende clasificar un fruto nuevo con mediciones $\underline{x}$ a cualquiera de los m EMC, se puede proceder de cualquiera de las siguientes formas :

1.- Calcular la distancia de Mahalanobis al cuadrado del fruto con cada uno de los EMC_g y restarle $2\ln(P_g)$ y entonces asignar el fruto al EMC_g que tenga el valor mínimo. Como se comento ello equivale a asignar el fruto al grupo en el cual tenga la máxima probabilidad aposteriori de pertenencia.

2.- Calcular la función de clasificación lineal para los g EMC y asignar el fruto al EMC_g con el máximo valor. Esto equivale a calcular las funciones de clasificación lineal para las

posibles diferencias de los m grupos y asignar el fruto al EMC_g si para toda $g \neq g'$ ocurre que $\hat{FCL}_{g-g'}(\underline{x}) > 0$.

Por otro lado si con los datos del estudio ocurre que las matrices de varianza y covarianza no son iguales , la regla de clasificación que minimiza el CEMCL es :

Asigne $\underline{x}$ al EMC_g si para toda $g \neq g'$ ocurre que:

$$Q_{g-g'}(\underline{x}) = FCC_{g-g'}(\underline{x}) \geq 0$$

donde $Q_{g-g'}(\underline{x}) = FCC_{g-g'}(\underline{x})$ = función de clasificación cuadrática para la diferencia de poblaciones.

De forma similar a las $\hat{FCL}$, para obtener la $\hat{FCC}$ se utilizan los estimadores de los parámetros

Es conveniente comentar que con cualquiera de las reglas de clasificación ya mencionadas se construye la matriz de clasificación[21] que nos indica : cuantos de los frutos son clasificados correctamente y de los restantes en que grupo se asignan erróneamente. Puede construirse también la denominada matriz de clasificación Jackknife,[22,27,28] que consiste en omitir fruto por fruto para construir las reglas de clasificación con las cuales se clasifican los frutos omitidos.

2.2.2 Funciones Discriminantes.

El enfoque de funciones discriminantes permite obtener un perfil de los EMC, para ello se investiga, si la representacion de las diferencias que se tienen entre los frutos de los m EMC con las q variables discriminantes (PDIFEEMC), pueden tenerse en una buena proporción en un nuevo espacio con pocas dimensiones . Dicho de

otra manera se pretende saber que proporción de la variabilidad total que tienen los datos originales, entre los EMC , puede recuperarse en una o dos dimensiones.

Para ello , el vector q-variado $\underline{x}$ asociado a cada fruto se transforma en un nuevo vector $\underline{Z}=(Z_1\ Z_2 \ldots\ Z_s)'$ con $s \leq q$, donde Z_c son combinaciones lineales de las q variables discriminantes que se denominan funciones discriminantes o variables canónicas (en lo que sigue llamaremos a las variables generadas canónicas y a su espacio asociado discriminante). En el nuevo espacio, Z_1 es una variable que hace maximó el %RDIFEEMC , Z_2 hace maximó el %RDIFEEMC no expresado en Z_1, ... , Z_r hace maximó el %RDIFEEMC no expresado en $Z_1, Z_2, \ldots, Z_{r-1}$ y finalmente Z_s es el %RDIFEEMC no incluido en las variables anteriores.

Como puede notarse al utilizar todas las variables del nuevo espacio generado, puede tenerse la totalidad de la RDIFEEMC , sin embargo como ya se comento, el interés radica en conocer si con Z_1 o bien con Z_1 y Z_2 (en el peor de los casos) se tiene una buena representación de las diferencias existentes entre los EMC. Pues de ser asi, ello posibilita la elaboración de sus perfiles, si en Z_1 se tiene una buena representación las diferencias pueden representarse en un eje, para el caso de Z_1 y Z_2 la representación se hace en el llamado plano discriminante principal[32,33] .

El procedimiento que se sigue para formar las variables canónicas Z_c es el siguiente:

1.- Para la variable canónica uno ($VC_1=Z_1$), sea:

$$Z_1 = a_{11} x_1 + a_{12} x_2 + \ldots + a_{1q} x_q = \underline{a}'_1 \underline{x} \qquad \ldots(29)$$

donde $(a_{11}\ a_{12} \ldots a_{1q})' = \underline{a}'_1$ son los llamados coeficientes discriminantes, por determinar, asociados a las variables discriminantes j ($j=1,2,\ldots,q$) para la VC_1.

Los coeficientes de $\underline{a}'_1$, por determinar, se obtienen al maximizar

$$\frac{\underline{a}'_1 E \underline{a}_1}{\underline{a}'_1 D \underline{a}_1} \qquad \ldots(30)$$

donde E [D] matriz de suma de cuadrados y productos cruzados de las variables originales entre [dentro] los EMC .

$$E = \sum_g \sum_i (\bar{\underline{X}}_{g.} - \bar{\underline{X}}_{..})(\bar{\underline{X}}_{g.} - \bar{\underline{X}}_{..})' = [\ E(j\ j')\ ] \qquad \ldots(31)$$

$$= [\ \sum_g \sum_i (\bar{X}_{g.j} - \bar{X}_{..j})(\bar{X}_{g.j'} - \bar{X}_{..j'})\ ]$$

con $$E(j\ j') = \sum_g X_{g.j}\, X_{g.j'}/n_g - X_{..j}\, X_{..j'}/n \qquad \text{si } j \neq j'$$

$$E(j\ j) = \sum_g X^2_{g.j}/n_g - X^2_{..j}/n \ , \ \text{si } j=j \qquad y$$

$$D = \sum_g \sum_i (\underline{X}_{gi} - \bar{\underline{X}}_{g.})(\underline{X}_{gi} - \bar{\underline{X}}_{g.})' = [\ D(j\ j')\ ] \qquad \ldots(32)$$

$$= [\ \sum_g \sum_i (X_{gij} - \bar{X}_{g.j})(X_{gij'} - \bar{X}_{g.j'})]$$

$$D(j\ j') = \sum_g \sum_i X_{gij}\, X_{gij'} - \sum_g X_{g.j}\, X_{g.j'}/n_g \qquad \text{si } j \neq j'$$

$$D(j\ j) = \sum_g \sum_i X^2_{gij} - \sum_g X^2_{g.j}/n_g \qquad \text{si } j=j'$$

Por otro lado la suma de cuadrados y productos cruzados de

las q variables originales del total ajustado por la media (T) es:

$$T = \sum_{g} \sum_{i} (X_{gi} - \bar{X}..)(X_{gi} - \bar{X}..)' = [\ T(j\ j')\ ] \qquad ...(33)$$

$$= [\ \sum_{g} \sum_{i} (X_{gij} - \bar{X}..j)(X_{gij'} - \bar{X}..j')\ ]$$

$$T(j\ j') = \sum_{g} \sum_{i} X_{gij}\, X_{gij'} - X..j\, X..j'/n \qquad \text{si } j \neq j'$$

$$T(j\ j) = \sum_{g} \sum_{i} X^{2}_{gij} - X^{2}.j/n \qquad \text{si } j = j'$$

Se puede demostrar que T = D + E ; para ello sólo basta considerar la equivalencia de los términos específicos j j' ya definidos.

Como se ha podido observar T , D y E representan la generalización de suma de cuadrados de un Análisis de Varianza para un modelo con un criterio de clasificación.

Otra forma de cálculo de E , D y T es considerando expresiones matriciales en forma compacta . Así por ejemplo para el cálculo de D , al considerar $S_g(j\ j')$ de (15) en términos de $D_g(j\ j')$ y luego al substituir en (32) se obtiene :

$$D = [\ D(j\ j')\ ] = [\ \sum_{g} D_g(j\ j')\ ] \qquad ...(34)$$

Ahora considerando (16) en (34), D puede expresarse como:

$$D = \sum_{g} Q_g \qquad ...(35)$$

donde Q_g puede calcularse con las expresiones (17), (18) o (20). O bien, por analogía al cálculo con (20) al formar la matriz de desviaciones de todos los datos multivariados de los m grupos con respecto a su vector de promedios respectivo (datos centrados por

EMC) , que denotamos como:

$$\underset{\sim}{X}^{d} = (\underset{\sim}{X} - \underset{\sim}{X}\,\underset{\sim}{P}) = \underset{\sim}{X}(I_{n\times n} - \underset{\sim}{P}_{n\times n}) \qquad \ldots(36)$$

donde, $\underset{\sim}{P} = \begin{bmatrix} \frac{1}{n_1} 1_{n_1\times n_1} & \underset{\sim}{O}_{n_2\times n_2} & \cdots & \underset{\sim}{O}_{n_m\times n_m} \\ \underset{\sim}{O}_{n_2\times n_2} & \frac{1}{n_2} 1_{n_2\times n_2} & \cdots & \underset{\sim}{O}_{n_m\times n_m} \\ \vdots & \vdots & \vdots & \\ \underset{\sim}{O}_{n_m\times n_m} & \underset{\sim}{O}_{n_m\times n_m} & \cdots & \frac{1}{n_m} 1_{n_m\times n_m} \end{bmatrix}$ $\qquad \ldots(37)$

Entonces $\qquad D = \underset{\sim}{X}^{d}\,\underset{\sim}{X}^{d\,\prime} \qquad \ldots(38)$

Ahora al considerar (36) en (38) se obtiene :

$$D = \underset{\sim}{X}_{q\times n}\,(I_{n\times n} - \underset{\sim}{P}_{n\times n})\,\underset{\sim}{X}'_{n\times q} \qquad \ldots(39)$$

Procediendo de forma similar a D se obtiene E y T considerando:

$$E = \overset{e}{\underset{\sim}{X}}\,\overset{e}{\underset{\sim}{X}}' = \underset{\sim}{X}_{q\times n}\,(\underset{\sim}{P}_{n\times n} - \frac{1}{n} 1_{n\times n})\,\underset{\sim}{X}'_{n\times q} \qquad \ldots(40)$$

con $\overset{e}{\underset{\sim}{X}} = (\underset{\sim}{X}\,\underset{\sim}{P} - \underset{\sim}{X}\,\frac{1}{n} 1_{n\times n}) = \underset{\sim}{X}(\underset{\sim}{P} - \frac{1}{n} 1_{n\times n})$ matriz de desviaciones de los vectores promedio por grupo con respecto al vector de los promedios generales. y

$$T = \overset{t}{\underset{\sim}{X}}\,\overset{t}{\underset{\sim}{X}}' = \underset{\sim}{X}_{q\times n}\,(I_{n\times n} - \frac{1}{n} 1_{n\times n})\,\underset{\sim}{X}'_{n\times q} \qquad \ldots(41)$$

con $\overset{t}{\underset{\sim}{X}} = (\underset{\sim}{X} - \underset{\sim}{X}\,\frac{1}{n} 1_{n\times n}) = \underset{\sim}{X}(I_{n\times n} - \frac{1}{n} 1_{n\times n})$ matriz de desviaciones de los datos multivariados con respecto al vector de promedios generales.

Tomando en consideración los resultados anteriores en (30), puede

demostrarse que:

$$\text{Máx}\frac{\underline{a}'_1 E \underline{a}_1}{\underline{a}'_1 D \underline{a}_1} = \text{Mín}\frac{\underline{a}'_1 T \underline{a}_1}{\underline{a}'_1 E \underline{a}_1} = \text{Máx}\frac{\underline{a}'_1 E \underline{a}_1}{\underline{a}'_1 T \underline{a}_1} \qquad \ldots(42)$$

Para obtener una solución única a este problema de maximización consideremos por ejemplo la última igualdad y en ella la restricción $\underline{a}'_1 T \underline{a}_1 = 1$. Con la técnica de los multiplicadores de Lagrange (que maximiza una función sujeta a restricciones) se obtiene que la función a maximizar es :

$$\Psi_1(\underline{a}_1, \lambda) = \underline{a}'_1 E \underline{a}_1 - \lambda (\underline{a}'_1 T \underline{a}_1 - 1) \qquad \ldots(43)$$

y al aplicar dicha técnica se tiene :

$$\partial\Psi_1(\underline{a}_1, \lambda) / \partial \underline{a}_1 = 2 E \underline{a}_1 - 2 \lambda T \underline{a}_1 = 0 \qquad \ldots(44)$$

$$\partial\Psi_1(\underline{a}_1, \lambda) / \partial \lambda = - \underline{a}'_1 T \underline{a}_1 + 1 = 0 \qquad \ldots(45)$$

al premultiplicar (44) por $\underline{a}'_1$ y considerar (45) se deduce que $\lambda = \underline{a}'_1 E \underline{a}_1$ es el valor máximo que toma (43) relacionado con el %RDIFEEMC y generado por los valores de $\underline{a}_1$.

Para obtener λ y $\underline{a}_1$, al premultiplicar (44) por T^{-1} se obtiene $(T^{-1}B - \lambda I)\underline{a}_1 = 0$ y tomando en cuenta (45) se concluye que λ es la máxima raiz característica de $T^{-1}B$ (sea λ_1) y $\underline{a}_1$ su vector característico asociado, que satisface $\underline{a}'_1 T \underline{a}_1 = 1$.

2.- La variable canónica dos (VC2=Z2) es generada como:

$$Z_2 = a_{z1} X_1 + a_{z2} X_2 + \ldots + a_{zq} X_q = \underline{a}'_2 \underline{X} \qquad \ldots(46)$$

y es una combinación lineal que maximiza el %RDIFEEMC , no correlacionada con la VC1 (Z_1) , esto es, se quiere encontrar un

vector $\underline{a}z$ que maximice:

$$\frac{\underline{a}'z\ E\ \underline{a}\ z}{\underline{a}'z\ T\ \underline{a}\ z} \quad \text{condicionada a que} \quad \underline{a}'z\ T\ \underline{a}\ z=1 \text{ y } \underline{a}'z\ T\ \underline{a}\ 1=\underline{0} \qquad \ldots(47)$$

En este caso la función por maximizar es :

$$\Psi z(\underline{a}\ z,\lambda,\mu)=\underline{a}'z\ E\ \underline{a}\ z-\lambda(\underline{a}'\ z\ T\ \underline{a}\ z-1)-\mu\ \underline{a}'z\ T\ \underline{a}\ 1 \qquad \ldots(48)$$

nuevamente al aplicar la técnica de los multiplicadores se tiene :

$$\partial\Psi z(\underline{a}\ z,\lambda,\mu)\ /\ \partial\underline{a}\ z=2\ E\ \underline{a}\ z-2\ \lambda\ T\ \underline{a}\ z-2\ \mu\ T\ \underline{a}\ 1=0 \qquad \ldots(49)$$

$$\partial\Psi z(\underline{a}\ z,\lambda,\mu)\ /\ \partial\ \lambda=-\underline{a}'\ z\ T\ \underline{a}\ z+1=0 \qquad \ldots(50)$$

$$\partial\Psi z(\underline{a}\ z,\lambda,\mu)\ /\ \partial\ \mu=\underline{a}'z\ T\ \underline{a}\ 1=0 \qquad \ldots(51)$$

al premultiplicar (49) por $\underline{a}'z$ se tiene:

$\underline{a}'z\ E\ \underline{a}\ z-\lambda\ \underline{a}'z\ T\ \underline{a}\ z-\mu\ \underline{a}'z\ T\ \underline{a}\ 1=0$ y al considerar (50) y (51):

$\lambda=\underline{a}'z\ E\ \underline{a}\ z$ es el valor máximo que toma (48) relacionado con el %RDIFEEMC y generado por los valores de $\underline{a}\ z$.

Ahora al premultiplicar (49) por $\underline{a}'1$ se tiene:

$$\mu\ \underline{a}'1\ T\ \underline{a}\ 1=\underline{a}'1\ E\ \underline{a}\ z-\lambda\ \underline{a}'1\ T\ \underline{a}\ z \qquad \ldots(52)$$

por otro lado como $(T^{-1}E-\lambda I)\underline{a}\ 1=0$ y como $\lambda 1\neq 0$ entonces:

$$\underline{a}'1=\underline{a}'1\ ET^{-1}/\ \lambda 1 \qquad \ldots(53)$$

por (51) sabemos que $\underline{a}'1\ T\ \underline{a}\ z=0$ y considerando (53) en esta expresión se tiene que $\underline{a}'1\ T\ \underline{a}\ z=\underline{a}'1\ E\ T^{-1}T\ \underline{a}\ z=\underline{a}'1\ E\ \underline{a}\ z=0$.

Por lo tanto dado que $\underline{a}'1\ T\ \underline{a}\ z=0$, $\underline{a}'1\ E\ \underline{a}\ z=0$ y por (45) se deduce en (52) que $\mu=0$. Esto indica que la ecuación (48) es equivalente a la ecuación (43). Entonces resulta que la segunda raiz característica de TE^{-1} (sea $\lambda 2$) está generado por el vector característico normalizado $\underline{a}\ z$ ($\underline{a}'z\ T\ \underline{a}\ z=1$) y no

correlacionado con $\underline{a}_1$ ($\underline{a}'_1 T \underline{a}_2 = 0$).

Es importante anotar que el número de raíces características diferentes de cero, se refiere al rango de $T^{-1}E$ que corresponde con $s = \min (m-1, q)$.

En general, si se han extraído $r-1 < s$ variables canónicas, entonces para obtener la VCr se tiene que maximizar:

$$\frac{\underline{a}'_r E \underline{a}_r}{\underline{a}'_r T \underline{a}_r} \qquad \ldots(54)$$

sujeta a: $\underline{a}'_r T \underline{a}_{r=1}$ y $\underline{a}'_r T \underline{a}_{r-1} = \underline{a}'_r T \underline{a}_{r-2} = \ldots = \underline{a}'_r T \underline{a}_1 = 0$.

En este caso la función por maximizar es :

$$\Psi_r(\underline{a}_r, \lambda, \mu_1, \mu_2, \ldots, \mu_r) = \underline{a}'_2 E \underline{a}_2 - \lambda(\underline{a}'_r T \underline{a}_r - 1) - \sum_{k=1}^{r-1} \mu_k \underline{a}'_r T \underline{a}_k \qquad \ldots(55)$$

nuevamente al aplicar la técnica de los multiplicadores se tiene :

$$\partial\Psi_r / \partial \underline{a}_r = 2 E \underline{a}_r - 2 \lambda T \underline{a}_r - 2 \sum_{k=1}^{r-1} \mu_k T \underline{a}_k = 0 \qquad \ldots(56)$$

$$\partial\Psi_r / \partial \lambda = - \underline{a}'_r T \underline{a}_r + 1 = 0 \qquad \ldots(57)$$

$$\partial\Psi_r / \partial \mu_k = \underline{a}'_r T \underline{a}_k = 0 \quad r \neq k \ (k = 1, 2, \ldots, r-1) \qquad \ldots(58)$$

al premultiplicar $\underline{a}'_l$ (para toda $l < r$) por (56) y considerando (58), se tiene que :

$$\mu_l \underline{a}'_l T \underline{a}_l = \underline{a}'_l E \underline{a}_r - \lambda \underline{a}'_l T \underline{a}_r \qquad \ldots(59)$$

por otro lado, como $(T^{-1}B - \lambda I)\underline{a}_l = 0$ entonces

$$\underline{a}'_l = \underline{a}'_l B T^{-1} / \lambda_l \text{ , ya que } \lambda_l \neq 0 \qquad \ldots(60)$$

Por (58) sabemos que $\underline{a}'_l T \underline{a}_r = 0$ para toda $l < r$, asimismo

considerando (60) en esta expresión se tiene que:

$$\underline{a}'_l\, T\, \underline{a}_r = \underline{a}'_l B T^{-1} T\, \underline{a}_r = \underline{a}'_l B\, \underline{a}_r = 0, \quad \text{para toda } l<r \qquad \ldots(61)$$

Al considerar $\underline{a}'_l\, E\, \underline{a}_r = 0$, $\underline{a}'_l\, T\, \underline{a}_r = 0$ y $\underline{a}'_l\, T\, \underline{a}_l = 1$ en (59) se deduce que $\mu_l = 0$ para $l<r$. Esto indica que la ecuación (55) para toda $l<r$ es equivalente a la ecuación (43). De aquí se obtiene entonces que λ_r es la r-ésima raíz característica de $T^{-1}E$ y $\underline{a}_r$ es su vector característico asociado, tal que $\underline{a}'_r\, T\, \underline{a}_r = 1$ y no correlacionado con $\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_{r-1}$.

Resumiendo, para recuperar la representación que se tiene en q dimensiones con las variables discriminantes, en el espacio de las variables canónicas, se calculan las q raíces características (λ_c) y sus vectores característicos asociados de $T^{-1}E$, tales que $\underline{a}'_c\, T\, \underline{a}_c = 1$ y donde el número de raíces diferentes de cero son $s = \min(m-1, q)$.

En el enfoque de funciones discriminantes a partir de los raíces características se calculan medidas que indican la calidad de la regla de separación. Así por un lado se obtiene el porciento de variación total explicada por cada una de las variables canónicas[31] (%RDIFEEM) y por otro se calculan las correlaciones canónicas entre las VC (funciones discriminantes) y las variables indicadoras de pertenencia a cada EMC[30]. Para ello:

$$\Re_c = \frac{\lambda_c^{1/2}}{(1+\lambda_c)^{1/2}} \quad c = 1,2,\ldots,s\ [\min(m-1, q)] \qquad \ldots(62)$$

permite obtener las correlaciónes canónicas. Y para el % de variación total se emplea la expresión:

$$\%DVC_c = \frac{\lambda_c}{\sum_{c=1}^{s} \lambda_c} \qquad \ldots(63)$$

Es importante anotar que los vectores característicos de $T^{-1}E$ y los obtenidos a partir de $D^{-1}E$ son iguales, cuando dichos vectores se obtienen normalizados de igual manera. Además sus raíces tienen la siguiente relación :

$$\lambda_r = \eta_r / (1 + \eta_r) \qquad \ldots(64)$$

con η_r raíz característica r-ésima de $D^{-1}E$.

En $D^{-1}E$ puede pensarse que lo que se pretende es maximizar las diferencias entre los EMC y minimizarla dentro de ellos. Por esto es que se comenta en la literatura estadística que , en el enfoque de las funciones discriminantes para la separación de los m EMC, existe una estrecha relación entre los coeficientes discriminantes para formar las VC y el análisis de varianza univariado con un criterio de clasificación utilizando las VC , pues a partir de las VCs es posible calcular los datos canónicos[24] como :

$$Z_{cgi} = a_{c1} X_{gi1} + a_{c2} X_{gi2} + \ldots + a_{cq} X_{giq} = \underline{a}'_c \underline{X}_{gi} \qquad \ldots(65)$$

$$\bar{Z}_{cg.} = a_{c1} \bar{X}_{g.1} + a_{c2} \bar{X}_{g.2} + \ldots + a_{cq} \bar{X}_{g.q} = \underline{a}'_c \bar{\underline{X}}_{g.}$$

$$\bar{Z}_{c..} = a_{c1} \bar{X}_{..1} + a_{c2} \bar{X}_{..2} + \ldots + a_{cq} \bar{X}_{..q} = \underline{a}'_c \bar{\underline{X}}_{..}$$

con Z_{cgi} = valor de la VC$_c$ para el fruto i del EMC$_g$.

$\bar{Z}_{ig.}$ y $\bar{Z}_{i..}$ = valores promedio de la VC$_c$ para el EMC$_g$ y para el total de la muestra ($n = \sum_{g=1}^{m} n_g$).

y con ellos se puede mostrar que los elementos de $\underline{a}$ c

(coeficientes discriminantes) máximizan la distancia entre los $\bar{Z}_{cg}$ (g=1,2,...,m) relativa a la distancia dentro de los ellos, expresada en la estadística F utilizada como criterio de prueba para contrastar las diferencias existentes entre los EMC, es decir, para toda variable canónica se cumple que:.

$$\frac{\sum_g \sum_i (\bar{Z}_{1g.} - \bar{Z}_{1..})^2}{\sum_g \sum_i (Z_{1gi} - \bar{Z}_{1g.})^2} = \frac{\underline{a}'_1 \left(\sum_g \sum_i (\bar{\underline{X}}_{g.} - \bar{\underline{X}}_{..})(\bar{\underline{X}}_{g.} - \bar{\underline{X}}_{..})' \right) \underline{a}_1}{\underline{a}'_1 \left(\sum_g \sum_i (\bar{\underline{X}}_{gi} - \bar{\underline{X}}_{g.})(\bar{\underline{X}}_{gi} - \bar{\underline{X}}_{g.})' \right) \underline{a}_1} = \frac{\underline{a}'_1 \, E \, \underline{a}_1}{\underline{a}'_1 \, D \, \underline{a}_1}$$

Interpretación con las Funciones Discriminantes.

Para propósitos prácticos .en la aplicación de las FD se pretende que las dos primeras variables canonicas expliquen casi la totalidad de la variabilidad del espacio original, pues de ser así en el plano discriminante principal, es posible gráficar las diferencias obtenidas entre los EMC con las variables discriminantes. En este contexto cuando se pretende elaborar los perfiles de los EMC surge la pregunta: Cuáles variables discriminantes contribuyen mas en la diferenciación de ellos?. Es importante anotar que cuando se calcula dicha contribución esta debe ser comparativa, es decir, independiente de las escalas de medida empleadas. Por ello es que los perfiles se forman considerando los pesos o coeficientes estandarizados, asociados a las variables discriminantes estandarizadas (centradas y reducidas).

Los coeficientes estandarizados se obtienen al considerar los

valores de las variables discriminantes de manera estandarizada, de modo que ellas tengan media cero y varianza uno. Asi entonces cuando se considera la expresión (42) pero con matrices estandarizadas, los coeficientes estandarizados de las funciones discriminantes (variables canónicas) se obtienen al considerar:

$$\text{Máx}\frac{\underline{w}'_1 E^e \underline{w}_1}{\underline{w}'_1 D^e \underline{w}_1} = \text{Min}\ \frac{\underline{w}'_1 T^e \underline{w}_1}{\underline{w}'_1 E^e \underline{w}_1} = \text{Máx}\frac{\underline{w}'_1 E^e \underline{w}_1}{\underline{w}'_1 T^e \underline{w}_1} \qquad \ldots(66)$$

Recordemos que cuando los datos no están estandarizados la matriz de dispersión ponderada Sp se expresa como:

$$Sp = \frac{\sum_g Q_g}{n-m} = \frac{D}{n-m} = \frac{\sum_g (n_g - 1)\, S_g}{\sum (n_g - 1)} \qquad \ldots(67)$$

Cuando las variables son estandarizadas Sp corresponde a la matriz de correlación R, que puede calcularse de acuerdo a (22) como :

$$R = DIAG(Sp)^{-1/2}\ Sp\ \ DIAG(Sp)^{-1/2} \qquad \ldots(68)$$

de dimensiones q x q. De manera similar la suma de cuadrados y y productos cruzados dentro(D) y entre (E) los EMC_g estandarizados son:

$$D^e = DIAG(Sp)^{-1/2}\ D\ \ DIAG(Sp)^{-1/2} = R \qquad \ldots(69)$$

$$E^e = DIAG(Sp)^{-1/2}\ E\ \ DIAG(Sp)^{-1/2}$$

Asi entonces considerando la primera igualdad de la expresión (42) con variables discriminantes estandarizadas, el problema se

traduce en encontrar los raices y vectores caracteristicos de $R^{-1}E^{\bullet}$ de modo que los vectores caracteristicos de las variables canónicas $\underline{w}c\ (c=1,2,\ldots,s)$ satisfagan que $\underline{w}'c\ R\ \underline{w}\ c = 1$.

Otra forma de obtener los coeficientes estandarizados, es considerando que el valor de la VCc para cada fruto, con q variables discriminantes no estandarizadas es :

$$Zcgi = \sum_j acj\ xgij = \underline{a}'c\ \underline{X}gi \qquad \ldots(70)$$

de donde $\bar{Z}cg. = a\ c1\ \bar{X}g.1 + a\ c2\ \bar{X}g.2 + \ldots + a\ cq\ \bar{X}g.q = \underline{a}'c\ \bar{\underline{X}}g.$

y entonces el valor promedio de la VCc al considerar los m EMC es:

$$\bar{\bar{Z}}c.. = (\bar{Z}c1. + \bar{Z}c2. + \ldots + \bar{Z}cm.)/m = \underline{a}'c(\bar{\underline{X}}1. + \bar{\underline{X}}2. + \ldots + \bar{\underline{X}}m.)/m$$

por lo cual los datos canónicos centrados se expresan como:

$$\mathcal{Z}cgi = Zcgi - \bar{\bar{Z}}c.. = \underline{a}'c\ [\ \underline{X}gi - (\bar{\underline{X}}1. + \bar{\underline{X}}2. + \ldots + \bar{\underline{X}}m.)/m\ ]$$

$$= \underline{a}'c\ [\ \underline{X}gi - \bar{\underline{X}}..\ ]$$

$$= \underline{a}'c\ (\ \underline{X}gi - (\ \bar{X}..1\quad \bar{X}..2\quad \ldots\quad \bar{X}..q)'\ )$$

$$= a\ c1\ (Xgi1 - \bar{X}..1) + a\ c2\ (Xgi2 - \bar{X}..2) + a\ cq\ (Xgiq - \bar{X}..q)$$

$$= a\ c1\ \acute{o}gi1 + a\ c2\ \acute{o}gi2 + \ldots + a\ cq\ \acute{o}giq \qquad \text{entonces:}$$

$$\mathcal{Z}cgi = \sum_j a\ cj\ \acute{o}gij = \sum_j a\ cj\ (\ Xgij - \bar{X}..j) \qquad \ldots(71)$$

donde c= es un subíndice que denota el número de variable canónica .

$\mathcal{Z}cgi$= desviación de un dato canónico con respecto al promedio de ellos, o dato canónico centrado.

$\acute{o}gij = Xgij - X..j$ = desviacion del dato $Xgij$ con respecto al valor promedio de la variable en cuestión j. Dato original centrado.

a cj = coeficiente discriminante de la VCc , asociado a la variable j .

Así $\mathcal{Z}cgi$ puede expresarse como :

$$\mathcal{Z}cgi = \sum_j a\ cj\ Xgij - \sum_j a\ cj\ \bar{X}..j \qquad ...(72)$$

Tanto (70),(71)y(72) se refieren a la VCc no estandarizada, sin embargo en la paqueteria estadística se reporta (72), en donde el segundo término se denomina constante de la función discriminante (o variable canónica) no estandarizada.

Para obtener los coeficientes discriminantes estandarizados se multiplica y divide (71) término a término por la desviación estandar promedio de la variable discriminante correspondiente (ésta se refiere a la raiz cuadrada de los términos de la diagonal de la matriz Sp). Y se obtiene :

$$\mathcal{Z}cgi = \sum_j a\ cj\ Sp(j\ j)(Xgij-\bar{X}..j\ ) \ /\ Sp(j\ j) = \sum_j W\ cj\ \zeta gij \qquad ...(73)$$

o, $$\mathcal{Z}cgi = \sum_j W\ cj\ Xgij\ /\ Sp(j\ j) - \sum_j a\ cj\ \bar{X}..j \qquad ...(74)$$

donde W cj =coeficiente discriminante estandarizado, para la variable j, de la VCc.

$\zeta gij = (Xgij - \bar{X}..j) / Sp(j\ j) = ógij\ /\ Sp(j\ j)$, valor estandarizado, de la variable Xgij original.

Las expresiones (73) y (74) se refieren a la VCc estandarizada, pero la expresión (73) es la que generalmente se reporta en los paquetes de cómputo. Los valores canónicos para cualquier fruto pueden obtenerse con las expresiones mencionadas , pero se prefieren (72) y (73) para las VCc no estandarizada y estandarizada respectivamente , pues con ellas se tiene consistencia en el lenguaje. Asi cuando se habla de coeficientes discriminantes estandarizados, éstos están asociados a variables

orijinales estandarizadas en el sentido estadístico, en el otro caso, los no estandarizados están asociados a las variables originales solamente centradas y es por ello que aparece el término constante.

2.3 Equivalencia entre las Funciones de Clasificación y las de Discriminación.

El enfoque de funciones discriminantes (variables canónicas) consiste en determinar para cada dato muestra multivariado, un conjunto de a lo mucho q variables canónicas. Como ya habiamos comentado el número de raices características distintas de cero para $D^{-1}E$ son $s = min(m-1, q)$ y las restantes raices $q-s$ se consideran $\eta_{s+1}=\eta_{s+2}=\ldots=\eta_q = 0$ con vectores $\underline{a}_{s+1}, \underline{a}_{s+2}, \ldots, \underline{a}_q$.

Para mostrar la equivalencia se consideran las q VC ,aunque , algunas de ellas realmente no aporten separación sustancial entre los grupos. Las q VC se definen a partir de un dato q multivariado original $\underline{x}$ como sigue:

$$\underline{Z}_{qx1} = \underline{L}'_{qxq}\ \underline{x}_{qx1} = \begin{bmatrix} \underline{a}'_1 \\ \underline{a}'_2 \\ \vdots \\ \underline{a}'_q \end{bmatrix} \underline{x} = \begin{bmatrix} \underline{a}'_1\ \underline{x} \\ \underline{a}'_2\ \underline{x} \\ \vdots \\ \underline{a}'_q\ \underline{x} \end{bmatrix} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_q \end{bmatrix} \qquad \ldots(75)$$

donde $\underline{L} = (\underline{a}_1\ \underline{a}_2\ \ldots\ \underline{a}_q)$ esta formada por los vectores característicos para las q posibles variable canónicas.

$\underline{a}_c = (a_{c1}\ a_{c2}\ \ldots a_{cq})'$ vector característico para la variable canónica c ($c=1,2,\ldots,q$). Donde

$Z_c = \underline{a}'_c\ \underline{x} = a_{c1}\ x_1 + a_{c2}\ x_2 + \ldots + a_{cq}\ x_q$. Y

$\underline{x}=(x_1\ x_2 \ldots x_q)'$= vector de q variables discriminantes.

Ahora bien, como $\underline{Z}$ es una transformación no singular de $\underline{x}$. Los vectores $\underline{a}_c$ por construcción tienen las siguientes propiedades:

$$\underline{a}'_c \, D \, \underline{a}_c = 0 \qquad \ldots(76)$$

$$\underline{a}'_c \, E \, \underline{a}_{c'} = 0 \quad (c \neq c' \; ; \; c, c' = 1, 2, \ldots, q)$$

además como son normalizados ($\underline{a}'_c \, D \, \underline{a}_c = 1$) entonces $\underline{L}$ vuelve identidad a la matriz D y diagonaliza la matriz E, es decir:

$$\underline{L}' D \, \underline{L} = I_{qxq} \qquad \ldots(77)$$

$$\underline{L}' E \, \underline{L} = DIAG \, [\, \eta_1 \; \eta_2 \; \ldots \; \eta_q \,]$$

Ahora bien de (77) se obtiene que $(\underline{L}' D \, \underline{L})^{-1} = \underline{L}^{-1} \, D^{-1} \, \underline{L}'^{-1} = I_{qxq}$ y al pre y post multiplicar por $\underline{L}$ y $\underline{L}'$ a $\underline{L}^{-1} \, D^{-1} \, \underline{L}'^{-1} = I_{qxq}$ se obtiene la igualdad:

$$D^{-1} = \underline{L} \, \underline{L}' \qquad \ldots(78)$$

Por otro lado la matriz de varianza covarianza de $\underline{Z} = \underline{L}' \underline{x}$ es $\underline{L}' \Sigma \, \underline{L}$ y un estimador de ella es $(n - m)^{-1} \underline{L}' D \, \underline{L}$, quien por (77) es igual con :

$$(n - m)^{-1} \, I_{q \times q} \qquad \ldots(79)$$

Con lo mencionado puede entonces demostrarse la equivalencia de los dos enfoques presentados.

Para ello consideremos que para cada EMC$_g$ se tiene : $\bar{\underline{Z}}_g = \underline{L}' \bar{\underline{X}}_g$ donde a partir de las q VC para el EMC$_g$ ($g = 1, 2, \ldots, m$) se obtiene $\bar{\underline{Z}}_g = (\bar{Z}_{g.1} \; \bar{Z}_{g.2} \ldots \bar{Z}_{g.q})'$ que se denomina centroide[26] del EMC$_g$ considerando las q Vc y se calcula a partir del centroide

$\bar{X}_g = (\bar{X}_{g.1}\ \bar{X}_{g.2} \ldots \bar{X}_{g.q})'$ de los datos de las variables originales.

Cuando se quiere clasificar un nuevo fruto con mediciones $x_o = (x_{1o}\ x_{2o} \ldots x_{qo})'$ en los EMC , se calculan los q datos canónicos $Z_o = (Z_{1o}\ Z_{2o} \ldots Z_{qo})'$, con la expresión $Z_o = L' x_o$.

Y entonces el fruto con medidas Z_o es asignada al EMC_g con el que sea más próximo al calcular la distancia de Mahalanobis:

$$\hat{DM}^2_g(Z_o) = (\bar{Z}_g - Z_o)' \left[(n-m)^{-1} I_q \right]^{-1} (\bar{Z}_g - Z_o)$$

$$= (n-m) \left[\bar{Z}_g' \bar{Z}_g - 2 \bar{Z}_g' Z_o + Z_o' Z_o \right] \qquad \ldots(80)$$

Por otro lado, la función de clasificación para variables originales cuando las probabilidades a priori son iguales es, de acuerdo a (11):

$$\hat{FCL}_g(x_o) = (n-m)\ \bar{X}_g' D^{-1} x_o - \frac{1}{2} (n-m)\ \bar{X}_g' D^{-1} \bar{X}_g \qquad \ldots(81)$$

$$= \bar{X}' S_p^{-1} x_o - 1/2\ \bar{X}_g' S_p^{-1} \bar{X}_g$$

al considerar (78), (81) puede reescribirse como:

$$\hat{FCL}_g(x_o) = (n-m)\ \bar{X}_g' L L' x_o - \frac{1}{2} (n-m)\ \bar{X}_g' L L' \bar{X}_g$$

$$= (n-m) \left[L' \bar{X}_g \right]' [L' x_o] - \frac{1}{2} (n-m) \left[L' \bar{X}_g \right]' [L' \bar{X}_g]$$

$$= (n-m)\ \bar{Z}_g' Z_o - \frac{1}{2} (n-m)\ \bar{Z}_g' \bar{Z}_g \qquad \ldots(82)$$

entonces $\hat{FC}_g(x_o)$ puede expresarse en funcion de $\hat{DM}^2_g(Z_o)$ como :

$$\hat{FC}_g(x_o) = -\frac{1}{2} \hat{DM}^2_g(Z_o) + \frac{1}{2} (n-m) Z_o' Z_o \qquad \ldots(83)$$

ya que al sustituir (80) en (83) se obtiene (82). En (83) puede observarse que el Máx de $\hat{FC}_g(x_o)$ y el Mín de $\hat{DM}^2_g(Z_o)$ ocurren para el mismo valor del EMC_g, lo cual prueba la equivalencia de los dos

enfoques [Kshirsagar A. M. and Arseven E: 14]

Otra forma de demostrar la equivalencia es a partir de la primer equivalencia de (80), es decir :

$$\hat{DM}_g^2(\underline{Z}_o) = (\bar{\underline{Z}}_g-\underline{Z}_o)' [(n-m)^{-1} I_q]^{-1} (\bar{\underline{Z}}_g-\underline{Z}_o)$$

$$=(n-m)[\underline{L}'\bar{\underline{X}}_g - \underline{L}'\underline{x}_o]'[\underline{L}'\bar{\underline{X}}_g-\underline{L}'\underline{x}_o]$$

$$=(n-m)[\underline{L}'(\bar{\underline{X}}_g-\underline{x}_o)]'[\underline{L}'(\bar{\underline{X}}_g-\underline{x}_o)]$$

$$=(n-m)(\bar{\underline{X}}_g-\underline{x}_o)'\underline{L}\,\underline{L}'(\bar{\underline{X}}_g-\underline{x}_o), \text{ por (78) como } \underline{L}\,\underline{L}' = D^{-1}$$

entonces $=(n-m)(\bar{\underline{X}}_g-\underline{x}_o)' D^{-1} (\bar{\underline{X}}_g-\underline{x}_o)$

$$=(\bar{\underline{X}}_g-\underline{x}_o)' S_p^{-1} (\bar{\underline{X}}_g-\underline{x}_o) = \hat{DM}_g^2(\underline{x}_o)$$

2.4 Uso de las funciones Discriminantes para clasificar.

Aunque las variables canónicas son derivadas con el propósito de tener una representación de los datos con pocas variables para separar y proporcionar los perfiles de los EMC tan bien como sea posible, también puede construirse con ellas una regla de clasificacion. Para ello se considera la distancia al cuadrado de Mahalanobis con los datos canónicos.

De esta manera un fruto se asigna al EMCg. si para toda g diferente de g' ocurre que:

$$(\bar{\underline{Z}}_g-\underline{Z}_o)'(\bar{\underline{Z}}_g-\underline{Z}_o) < (\bar{\underline{Z}}_{g'}-\underline{Z}_o)'(\bar{\underline{Z}}_{g'}-\underline{Z}_o) \quad \ldots(84)$$

$\bar{\underline{Z}}_g$= centroide para el grupo $g=1,2,\ldots,m$

$$\sum_{c=1}^{q} (\bar{Z}_{cg}-Z_{co})^2 < \sum_{c=1}^{q} (\bar{Z}_{cg'}-Z_{co})^2$$

con $\sum_{c=1}^{q} (\bar{Z}_{cg}-Z_{co})^2 = \sum_{c=1}^{s} (\bar{Z}_{cg}-Z_{co})^2 + \sum_{c=s+1}^{q} (\bar{Z}_{cg}-Z_{co})^2$ donde los $s = \min(m-1, q)$ primeros términos son realmente la parte útil para la clasificación pues corresponden a los valores

característicos distintos de cero.

Es conveniente comentar finalmente que los dos procedimentos presentados para la discriminación son equivalentes, cuando:

1.- En las funciones de clasificación las densidades son normales y las matrices de varianzas covarianzas de los grupos, las probabilidades a priori y los costos de mala clasificación son iguales. Y

2.- Se consideran todas las posibles variables canónoicas.

A n e x o

Cuadros de datos y ejemplos de su procesamiento con rutinas del paquete de computo estadistico BMDP.

CUADRO A: Datos observados en las 3 épocas de corte para obtener el índice de cosecha.
(NOTA E8 = E9)

PAGE 113 BMDP1D Indice de cosecha de chicozapote: datos de campo y laboratorio al momento de corte.

CASE NO.	LABEL	1 ev 11 ps 21 cf	2 ea 12 lae 22 fo	3 na 13 ns	4 nf 14 ps	5 dp 15 gv	6 de 16 fv	7 vo 17 gtv	8 pfcc 18 sv	9 pfsc 19 pol	10 gb 20 at
1		e8 4157.90 .0672	earlla .700 .881	a1 3	f1 2.651	82.500 2.516	72.700 2.065	210 4.581	214.400 4.707	200 .950	27.500 9.288
2		e8 2452.90 .0787	earlla .900 .790	a1 1	f2 .960	82.200 3.416	64.900 5.430	172 8.847	172.800 .749	159.200 .962	29 9.596
3		e8 2452.90 .0747	earlla .600 .774	a2 1	f1 .749	81.700 2.663	63.200 2.670	152 5.333	155.300 3.472	143.700 .921	27.500 8.805
4		e8 3342.90 .0684	earlla .600 .766	a3 1	f1 1.166	90.500 4.167	69.300 4.050	200 8.217	198.900 4.909	185.300 1.225	26 13.126
5		e8 2452.90 .0835	earlla .800 .809	a4 1	f1 .895	82.800 3.493	67 2.934	174 6.427	180.800 3.003	165.700 .579	29.500 9.430
6		e8 2452.90 .0781	earlla .900 .857	a4 2	f2 2.093	84.600 3.032	72.500 2.519	218 5.551	215.200 3.328	198.400 .712	28.250 8.879
7		e8 4157.90 .0771	earlla .700 .687	a5 1	f1 .863	94.200 4.001	64.700 3.111	176 7.113	191.900 6.238	177.100 .601	26 13.351
8		e8 3342.90 .0768	earlla .500 .740	a6 1	f1 1.130	95.800 3.747	70.900 3.986	248 7.732	243.400 5.494	224.700 .901	27.750 13.226
9		e8 2452.90 .0653	ear17a .900 .816	a1 1	f1 1.108	80.300 3.571	65.500 3.261	170 6.832	171.400 7.325	160.200 1.070	26 14.158
10		e8 3342.90 .0782	ear17a .700 .890	a1 2	f2 1.852	80.700 3.365	71.800 4.418	200 7.783	209.700 4.168	193.300 .710	24.750 11.951
11		e8 2452.90 .0582	ear17a .700 .900	a2 2	f1 2.034	83.700 3.042	75.300 2.771	235 5.813	226.700 3.780	213.500 .846	29.250 9.593
12		e8 2452.90 .0644	ear17a .900 1.154	a3 10	f1 8.534	74.100 2.582	85.500 2.831	250 5.414	279.700 2.156	261.700 .172	23 7.570
13		e8 3342.90 .0750	ear17a .800 .727	a4 1	f1 1.007	95.600 4.271	69.500 4.474	273 8.744	228 4.579	210.900 .904	29.750 13.323
14		e8 2452.90 .0601	ear17a .800 .027	a5 3	f1 2.642	88.500 3.025	73.200 2.685	280 5.711	229.700 4.396	215.900 .537	25 10.107
15		e8 2452.90 .0659	ear17a .600 .825	a6 2	f1 2.032	89.200 2.685	73.600 3.013	220 5.697	221.400 4.194	206.800 .857	28 9.892
16		e8 2452.90 .0638	ear17a .800 .807	a7 2	f1 1.936	83.600 3.810	67.500 2.837	184 6.647	182.900 2.880	171.400 .789	19 9.527

CUADRO A (continuación)

PAGE 114 BMDPID Indice de cosecha de chicozapote: datos de campo y laboratorio al momento de corte

C A S E NO.	LABEL	1 ev 11 fs 21 cf	2 ea 12 lae 22 fo	3 ne 13 ns	4 nf 14 ps	5 dp 15 gv	6 de 16 fv	7 vo 17 gtv	8 pfcc 18 sv	9 pfsc 19 pol	10 gb 20 at
17		e8	earl7a	a8	f1	84.100	78	249	257.500	239.700	28.750
		2452.90	.500	4	3.678	3.221	3.134	6.354	2.298	.691	8.653
		.0691	.727								
18		e8	earl7a	a9	f1	63.600	67.800	185	195.500	179.400	26.500
		1887.90	.700	2	1.164	3.037	2.672	5.710	2.759	.743	8.468
		.0824	.811								
19		e8	earl7a	a9	f2	90	71.300	215	213.900	195.900	26.750
		2452.90	.900	2	2.118	2.601	1.888	4.490	4.280	.601	8.770
		.0842	.772								
20		pc	earlla	a1	f1	78.400	63.200	155	160.300	147.800	29.250
		2452.90	1	1	1.085	3.044	2.915	5.960	8.081	892	14.041
		.0780	.806								
21		pc	earlla	a1	f2	93.700	65.500	200	207.200	192.900	31
		1072.90	.800	1	.952	2.924	2.886	5.810	8.617	1.025	14.427
		.0690	.699								
22		pc	earlla	a2	f1	66	61.100	120	124.300	110.500	17
		2452.90	1	1	.887	2.933	3.893	6.826	6.758	1.246	13 584
		.111	.926								
23		pc	earlla	a3	f1	75	64.100	148	150.300	138.600	29.500
		1072.90	.400	1	1.047	3.471	3.111	6.583	8.392	1.297	14.975
		.0778	.855								
24		pc	earlla	a4	f1	103.700	80.400	219	275.800	256.200	24.250
		1887.90	.400	3	3.823	2.851	3.457	6.308	9.936	.865	16.244
		.0711	.775								
25		pc	earlla	a4	f2	93.200	73.600	266	275.800	256.200	24.750
		1057.90	.500	2	2.224	2.976	3.277	6.253	7.507	1.167	13.760
		.0711	.790								
26		pc	earlla	a5	f1	96.400	64.100	188	191.300	176.900	23.750
		1887.90	.800	1	1.186	3.617	2.685	6.303	8.182	1.153	14.484
		.0753	.665								
27		pc	earlla	a6	f1	104.100	84.800	366	376.400	353.900	25.250
		1072.90	.500	4	3.682	3.259	3.245	6.504	8.854	1.290	15.358
		.0598	.815								
28		pc	earl7a	a1	f1	89.600	73.900	249	253	233.800	31
		1887.90	.800	3	3.366	2.733	3.666	6.399	6.305	.979	12.704
		.0759	.825								
29		pc	earl7a	a1	f2	88	72.200	270	239.700	209.600	29.500
		1057.90	.600	3	2.918	2.621	3.895	6.518	5.012	.927	11.530
		.126	.820								
30		pc	earl7a	a2	f1	76.700	71.600	200	199.100	186.700	21.250
		1887.90	.600	1	1.114	3.159	4.703	7.862	4.072	1.119	11.934
		.0623	.934								
31		pc	earl7a	a3	f1	72	69	180	181.700	168.400	25.250
		2452.90	1.400	4	3.985	3.195	3.894	7.089	6.978	1.286	14.067
		.0732	.958								
32		pc	earl7a	a4	f1	91.800	69.100	210	212.500	196.400	29.750
		1887.90	.300	1	1.034	2.912	4.530	7.443	14.078	1.146	21.520
		.0758	.753								

CUADRO A (continuación)

 BMDP1D indice de cosecha de chicozapote: datos de campo y laboratorio al momento de corte

CASE NO.	LABEL	1 ev 11 fs 21 cf	2 ea 12 lae 22 fb	3 na 13 ns	4 nf 14 ps	5 dp 15 gv	6 de 16 fv	7 vo 17 gtv	8 pfcc 18 sv	9 pfsc 19 pol	10 gb 20 at
33		pc 1072.90 .0843	ear17a .500 .932	a5 7	f1 6.202	78.900 3.408	73.500 3.303	218 6.712	212.300 3.482	194.400 1.048	26.500 10.194
34		pc 1087.90 .0690	ear17a .700 .835	a6 3	f1 3.016	94.700 3.445	79.100 4.749	289 8.194	301.500 8.107	280.700 1.098	28.750 16.301
35		pc 1887.90 .0786	ear17a .600 .739	a7 1	f1 .998	81.900 2.446	60.500 3.211	149 5.658	140 6.363	129 .834	26.250 12.020
36		pc 2452.90 .102	ear17a 1.500 .962	a8 5	f1 3.917	62.800 4.118	60.400 3.375	120 7.493	123.800 4.484	111.200 1.284	30 11.977
37		pc 1072.90 .0757	ear17a .300 .758	a9 3	f1 2.908	96.700 3.073	73.300 2.660	260 5.733	264.300 8.272	244.300 .803	29.250 14.004
38		pc 1072.90 .0776	ear17a .100 .700	a9 2	f2 2.084	102 2.967	71.400 2.240	257 5.207	260.200 6.846	240 .619	20.500 12.053
39		sc 500 .0232	earl1a .400 .822	a1 4	f1 4.034	78.300 2.955	64.400 4.325	126 7.280	129.100 10.518	126.100 .604	28.750 17.799
40		sc 200 .0303	earl1a .200 .722	a1 2	f2 1.862	79.800 1.584	73.600 4.049	130 5.634	131.900 10.042	127.900 .523	29.500 15.676
41		sc 150 .0151	earl1a .100 .764	a2 2	f1 1.920	92.700 3.133	70.800 3.531	250 6.664	251.900 7.894	248.100 .626	23 14.558
42		sc 450 .00593	earl1a .400 .744	a3 6	f1 5.247	102.300 2.374	76.100 4.446	317 6.819	320.300 11.026	318.400 .719	22.750 17.845
43		sc 400 .0119	earl1a .300 .885	a4 5	f1 4.018	96.600 3.536	85.500 3.938	223 7.075	226.500 8.079	223.800 .593	21.750 15.154
44		sc 350 .0150	earl1a .400 .761	a4 1	f2 1.102	81.900 2.680	62.300 4.572	143 7.252	146.800 10.450	144.600 .493	24.750 17.702
45		sc 200 .0223	earl1a .200 .830	a5 1	f1 1.253	82 3.031	68.100 4.336	120 7.367	123 10.616	120.200 .385	31 17.784
46		sc 300 .00751	earl1a .400 .771	a6 1	f1 1.432	92.500 2.271	71.300 3.635	237 5.906	239.700 8.489	237.900 .663	30.250 14.395
47		sc 300 .0121	ear17a .400 .744	a1 4	f1 4.026	83.700 2.328	62.300 3.971	120 6.299	124.300 5.876	122.800 .546	24.750 12.175
48		sc 350 .0150	ear17a .300 .739	a1 2	f2 2.199	86.200 3.039	63.700 4.045	166 7.084	173.200 4.861	170.600 .605	25.250 11.945

CUADRO N (continuación)

 BMDP1D indice de cosecha de chicozapote: datos de campo y laboratorio al momento de corte.

CASE NO.	LABEL	1 ev 11 fs 21 cf	2 ea 12 lae 22 fo	3 na 13 ns	4 nf 14 ps	5 dp 15 gv	6 de 16 fv	7 vo 17 gtv	8 pfcc 18 sv	9 pfsc 19 pol	10 gb 20 at
49		sc	ear17a	a2	f1	97.100	74.500	260	261.800	258.900	26
		400	.300	1	1.002	3.097	4.303	7.400	4.547	.525	11.947
		.0111	.767								
50		sc	ear17a	a3	f1	105.200	86.300	319	321.100	318.300	31
		450	.300	1	1.521	3.901	4.066	7.968	7.787	.672	15.755
		.00872	.820								
51		sc	ear17a	a4	f1	99.300	68.100	189	199.400	198.200	28.750
		250	.200	7	6.353	3.734	4.633	8.367	12.876	.566	21.243
		.00602	.686								
52		sc	ear17a	a5	f1	74.300	69	177	180.600	178.500	30
		300	.300	6	4.516	3.253	3.943	7.196	10.807	.454	18.003
		.0116	.929								
53		sc	ear17a	a6	f1	75	70.400	190	191.200	178.300	28.750
		300	.200	4	4.051	3.346	3.935	7.281	11.172	.543	18.453
		.0675	.939								
54		sc	ear17a	a7	f1	107.400	80.300	320	326.100	306	25.750
		300	.300	2	2.083	3.324	3.733	7.057	8.663	.539	15.720
		.0616	.748								
55		sc	ear17a	a8	f1	86.700	73.100	237	241.800	237.700	27
		300	.400	2	1.964	3.317	3.786	7.103	15.433	.626	22.536
		.0170	.843								
56		sc	ear17a	a9	f1	104.700	85.200	319	332.200	330.100	17
		250	.200	1	1.344	2.272	3.707	5.979	7.581	.568	13.560
		.00632	.814								
57		sc	ear17a	a9	f2	83.200	61.200	158	160.900	158.300	18.750
		200	.200	1	1.238	3.194	2.797	5.992	7.480	.474	13.471
		.0162	.736								

NUMBER OF CASES READ. 57

VARIABLE NO.	NAME	STATED VALUES FOR MINIMUM	MAXIMUM	MISSING	CODE	GROUP INDEX	CATEGORY NAME	INTERVALS .GT.	.LE.
1	ev				8.000	1	e8		
					11.00	2	pc		
					12.00	3	sc		
2	ea				1.000	1	ear11a		
					2.000	2	ear17a		
3	na				1.000	1	a1		
					2.000	2	a2		
					3.000	3	a3		
					4.000	4	a4		
					5.000	5	a5		
					6.(	6			
					7.000	7	a7		

BMDP3D - ONE-SAMPLE AND TWO-SAMPLE T-TESTS

BMDP Statistical Software, Inc.
1440 Sepulveda Boulevard
Los Angeles, California 90025
Phone (213) 479-7799
Telex 4972934

Program Version: 1987
(VAX/VMS)
Manual Edition: 1983, 1985 reprint. State NEWS in the PRINT paragraph for a summary of new features.

19-OCT-87 AT 11:53:52

PROGRAM INSTRUCTIONS

```
'bmdp3d
 /input file is dla.code is chicofru.
/variable grouping is ev.
'test hotelling.title is'comparacion multivariada entre epocas
de corte'.correlation.nonparametric.robust.variables=5 to 10.
/end
```

'ROBLEM TITLE IS
19-OCT-87 11:53:52

```
 UMBER OF VARIABLES TO READ IN. . . . . .        10
NUMBER OF VARIABLES ADDED BY TRANSFORMATIONS. .   0
TOTAL NUMBER OF VARIABLES . . . . . .           10
 UMBER OF CASES TO READ IN. . . . . .  TO END
 ASE LABELING VARIABLES . . . . . .
MISSING VALUES CHECKED BEFORE OR AFTER TRANS. . NEITHER
"LANKS ARE. . . . . .  MISSING
 NPUT FILE. . '.dla
REWIND INPUT UNIT PRIOR TO READING. . DATA. . .   YES
NUMBER OF WORDS OF DYNAMIC STORAGE. . . . . .  19998
```

.NPUT BMDP FILE
CODE. . . IS chicofru
ONTENT . IS DATA
ABEL . . IS
15-OCT-87 16:47:39

''ARIABLES

1 ev	2 ea	3 na	4 nf	5 gv
6 fv	7 gtv	8 sv	9 pol	10 at

ARIABLES TO BE USED

1 ev	2 ea	3 na	4 nf	5 gv
6 fv	7 gtv	8 sv	9 pol	10 at

MULTIVARIATE STATISTICS FOR GROUP e8 VS. GROUP sc
THERE ARE 19 CASES, 19 OF THEM COMPLETE IN GROUP e8
THERE ARE 19 CASES, 19 OF THEM COMPLETE IN GROUP sc
NULL HYPOTHESIS IS THAT BOTH GROUPS HAVE EQUAL MEANS FOR ALL VARIABLES

DEGREES OF FREEDOM, BELOW, REDUCED BY 2
BECAUSE OF LINEAR DEPENDENCIES AMONG THE VARIABLES TESTED.

MAHALANOBIS D SQUARE	12.2930		
HOTELLING T SQUARE	116.7836		
F VALUE	26.7629	P-VALUE	0.0000
DEGREES OF FREEDOM	4, 33		

MULTIVARIATE STATISTICS FOR GROUP pc VS. GROUP sc
THERE ARE 19 CASES, 19 OF THEM COMPLETE IN GROUP pc
THERE ARE 19 CASES, 19 OF THEM COMPLETE IN GROUP sc
NULL HYPOTHESIS IS THAT BOTH GROUPS HAVE EQUAL MEANS FOR ALL VARIABLES

DEGREES OF FREEDOM, BELOW, REDUCED BY 2
BECAUSE OF LINEAR DEPENDENCIES AMONG THE VARIABLES TESTED.

MAHALANOBIS D SQUARE	16.1812		
HOTELLING T SQUARE	153.7218		
F VALUE	35.2279	P-VALUE	0.0000
DEGREES OF FREEDOM	4, 33		

CORRELATION MATRIX FOR GROUP 1 e8

		gv	fv	gtv	sv	pol	at
		5	6	7	8	9	10
gv	5	1.0000					
fv	6	0.6210	1.0000				
gtv	7	0.8451	0.9439	1.0000			
sv	8	0.3123	-0.1156	0.0527	1.0000		
pol	9	0.3809	0.3738	0.4155	0.3276	1.0000	
at	10	0.7678	0.5142	0.6743	0.7730	0.5063	1.0000

```
at         VARIABLE NUMBER  10        GROUP    1 eB        VS. GROUP  2 pc                    STATISTICS  P-VALUE  DF
*******************************   ***********************************************    ***********************************
                                                                                      SEPARATE  T   -4.83 0.0000    34.9
GROUP        eB           pc          1 eB       (N=  19)       2 pc       (N=  19)  POOLED    T   -4.83 0.0000    36
MEAN        10.4059      13.9566                                                      TRIM SEP.  T  -5.22 0.0000    31.2
STD DEV      2.0559       2.4567    HH                             X  X               TRIM POOL T   -5.22 0.0000    32
S.E.M.       0.4717       0.5636    HH     H                       X  X               LEVENE F FOR
SAMPLE SIZE      19           19    HH     H                       X  XX              VARIANCES      0.04 0.8486 1,     36
MAXIMUM     14.1575      21.5201    HHH    H                       X  XX X
  Z-SCORE    1.8248       3.0787    HHHH H HH                    X XX XXXX        X
  AT CASE         9           32  M--------------------M  M--------------------M
MINIMUM      7.5696      10.1936  I   AN H=     1 CASES  A  I   AN X=     1 CASES  A
  Z-SCORE   -1.3795      -1.5318  N                      X  N                      X
  AT CASE        12           33
2ND MAX     13.3510      16.3009
2ND MIN      8.4685      11.5299
TRIM MEAN   10.3520      13.7331

NONPARAMETRIC ANALYSIS
**********************
RANK SUMS FOR GROUPS eB        pc       ARE RESPECTIVELY       233.0      508.0
MANN-WHITNEY TEST STATISTIC =       43.0   TWO-TAIL P-VALUE = 0.0001

at         VARIABLE NUMBER  10        GROUP    1 eB        VS. GROUP  3 sc                    STATISTICS  P-VALUE  DF
*******************************   ***********************************************    ***********************************
                                                                                      SEPARATE  T   -6.87 0.0000    32.0
GROUP        eB           sc          1 eB       (N=  19)       3 sc       (N=  19)  POOLED    T   -6.87 0.0000    36
MEAN        10.4059      16.1011                                                      TRIM SEP.  T  -6.39 0.0000    28.2
STD DEV      2.0559       2.9719    HH                                      X         TRIM POOL T   -6.39 0.0000    32
S.E.M.       0.4717       0.6818    HH     H                             X  X         LEVENE F FOR
SAMPLE SIZE      19           19    HH     H                             X  X         VARIANCES      2.06 0.1594 1,     36
MAXIMUM     14.1575      22.5359    HHH    H                       X   XX  X
  Z-SCORE    1.8248       2.1652   HHHH H HH                       XXXXXX  XX    XX
  AT CASE         9           55  M--------------------M  M--------------------M
MINIMUM      7.5696      11.9450  I   AN H=     1 CASES  A  I   AN X=     1 CASES  A
  Z-SCORE   -1.3795      -1.3985  N                      X  N                      X
  AT CASE        12           48
2ND MAX     13.3510      21.2428
2ND MIN      8.4685      11.9475
TRIM MEAN   10.3520      15.9670

NONPARAMETRIC ANALYSIS
**********************
RANK SUMS FOR GROUPS eB        sc       ARE RESPECTIVELY       209.0      532.0
MANN-WHITNEY TEST STATISTIC =       19.0   TWO-TAIL P-VALUE = 0.0000
```

****

STEP NUMBER 3
VARIABLE ENTERED 17 lf

VARIABLE	F TO REMOVE	FORCE LEVEL	TOLERNCE	*	VARIABLE	F TO ENTER	FORCE LEVEL	TOLERNCE
DF =	2 52			*	DF =	2 51		
11 fs	51.78	1	0.20526	*	15 cf	25.67	2	0.94173
12 lae	17.15	1	0.10097	*	18 vf	0.59	2	0.46154
17 lf	15.59	1	0.05682	*	19 ff	0.71	2	0.16630

U-STATISTIC(WILKS' LAMBDA) 0.0994877 DEGREES OF FREEDOM 3 2 54
APPROXIMATE F-STATISTIC 37.620 DEGREES OF FREEDOM 6.00 104.00

F - MATRIX DEGREES OF FREEDOM = 3 52

	e8	pc
pc	15.89	
sc	116.61	56.35

CLASSIFICATION FUNCTIONS

GROUP =	e8	pc	sc
VARIABLE			
11 fs	0.04273	0.03630	0.01832
12 lae	105.59888	103.80839	62.09622
17 lf	-0.04990	-0.04771	-0.02839
CONSTANT	-48.86314	-36.66289	-11.58414

CLASSIFICATION MATRIX

GROUP	PERCENT CORRECT	NUMBER OF CASES CLASSIFIED INTO GROUP - e8	pc	sc
e8	94.7	18	1	0
pc	94.7	0	18	1
sc	100.0	0	0	19
TOTAL	96.5	18	19	20

JACKKNIFED CLASSIFICATION

GROUP	PERCENT CORRECT	NUMBER OF CASES CLASSIFIED INTO GROUP - e8	pc	sc
e8	94.7	18	1	0
pc	84.2	2	16	1
sc	100.0	0	0	19
TOTAL	93.0	20	17	20

BMDP7M - STEPWISE DISCRIMINANT ANALYSIS.

BMDP Statistical Software, Inc.
1964 Westwood Blvd. Suite 202 Phone (213) 475-5700
Los Angeles, California 90025 Telex 4992203

Program Version: April 1985
(VAX/VMS)
Manual Edition: 1983, 1985 reprint. State NEWS in the PRINT
paragraph for a summary of new features.

1-OCT-87 AT 13:32:32

PROGRAM CONTROL INFORMATION

```
/bmdp7m
/problem    title is 'indice de cosecha chicozapote:dpto fis.post conafrut'.
/input      variables are 19. groups are 3.
            format is free.
/variable   names are ev,ea,na,nf,dp,de,vo,pfcc,pfsc,gb,fs,lae,ns,ps,gv,fv,gtv,
            sv,pol,at,cf,fo,lf,vf,ff.
            add=6.
            grouping is ev.
            use=15,16,17,18,19,20.
            retain.
/transform  at=gtv+sv. cf=(pfcc-pfsc)/pfcc. fo=de/dp. lf=fs*lae. vf=fs*vo. ff=fs*fo.
/group      codes(1) are 08,11,12.
            names(1) are e8,pc,sc.
/discriminant method=1.
              tolerance=0.10.
              enter=2.0,2.0.
              remove=1.7,1.7.
              contrast=0,1,-1. contrast=1,-1,0.
/print      corr. with. no step. no posterior. no point.
/plot       group=1,2. group=2,3. group=1,2,3.
/end
```

***** TRAN PARAGRAPH IS USED *****

PROBLEM TITLE IS
indice de cosecha chicozapote:dpto fis.post conafrut

```
NUMBER OF VARIABLES TO READ IN. . . . . .          19
NUMBER OF VARIABLES ADDED BY TRANSFORMATIONS. .     6
TOTAL NUMBER OF VARIABLES . . . . . .              25
NUMBER OF CASES TO READ IN. . . . . .  TO END
CASE LABELING VARIABLES . . . . . .
MISSING VALUES CHECKED BEFORE OR AFTER TRANS. .  NEITHER
BLANKS ARE. . . . . .  MISSING
NUMBER OF WORDS OF DYNAMIC STORAGE. . . . . .    19998
```

VARIABLES TO BE USED

```
   15 gv        16 fv        17 gtv       18 sv        19 pol
   20 at
```

INPUT FORMAT IS

(

STEP NUMBER 4
VARIABLE ENTERED 17 gtv

VARIABLE	F TO REMOVE	FORCE LEVEL	TOLERNCE	*	VARIABLE	F TO ENTER	FORCE LEVEL	TOLERNCE
DF =	2 51			*	DF =	2 50		
15 gv	3.87	1	0.48329	*	16 fv	0.24	1	0.00000
17 gtv	2.98	1	0.41007	*	18 sv	1.15	1	0.00000
19 pol	41.15	1	0.81638	*				
20 at	19.60	1	0.78667	*				

U-STATISTIC(WILKS' LAMBDA) 0.1582916 DEGREES OF FREEDOM 4 2 54
APPROXIMATE F-STATISTIC 19.297 DEGREES OF FREEDOM 8.00 102.00

F - MATRIX DEGREES OF FREEDOM = 4 51

	e8	pc
pc	11.64	
sc	23.37	26.71

CLASSIFICATION FUNCTIONS

GROUP =	e8	pc	sc
VARIABLE			
15 gv	6.12815	4.25224	2.15579
17 gtv	3.32359	2.63279	4.81754
19 pol	9.14097	18.72783	-1.12761
20 at	0.44477	1.08237	1.55400
CONSTANT	-27.76148	-33.82701	-33.18811

CLASSIFICATION MATRIX

GROUP	PERCENT CORRECT	NUMBER OF CASES CLASSIFIED INTO GROUP - e8	pc	sc
e8	84.2	16	3	0
pc	89.5	2	17	0
sc	100.0	0	0	19
TOTAL	91.2	18	20	19

JACKKNIFED CLASSIFICATION

GROUP	PERCENT CORRECT	NUMBER OF CASES CLASSIFIED INTO GROUP - e8	pc	sc
e8	73.7	14	4	1
pc	84.2	2	16	1
sc	94.7	1	0	18
TOTAL	84.2	17	20	20

BIBLIOGRAFIA

[1] Méndez R. I.,Cañedo D. L y García R. H. "Principios de Investigación Médica". DIF . México 1977.

[2] Pelayo Z. C. y Pedraza G. E . ."Cosecha y Acondicionamiento de Frutas y Hortalizas ". Manuales Técnicos para la elaboración de Cursos de Capacitación. No. 3 . SEP / CONAFRUT / SCFI. 1982.

[3] Bósquez E .,"Qué son la Fisiología y Tecnología Postcosecha ". Boletín Informativo No. 1. CONAFRUT . México. 1985.

[4] Pelayo Z. C." El Papel de la Fisiología y Tecnología Postcosecha en la Producción Frutícola y Hortícola de México ". CONAFRUT. México 1981.

[5] Johnson A. R.Wichern W. D. "Applied Multivariate Statistical Analysis ". Prentice-Hall, Inc..Englewood Cliffs,New Jersey. 1982.

[6] Manual, " Statistical Software : BMDP ". University of California Press. Berkeley, L.A London. 1985.

[7] López C. C ."Metodología estadística en la determinación del índice de cosecha", Tesis Lic Actuaria. Facultad de Ciencias. UNAM-México. 1989.

[8] Hair J.F,Anderson R. E. "Multivariate Data Analysis ", Mc Millan , Neww-York. 1987.

[9] Tatsuoka M.M. "Multivariate Analysis : Tecniques for Educational and Psychological Research". John Wiley and Sons, 1981.

[10] Ochoa R.M. " Una aplicacion del Análisis Discriminante a problemas de Biología",Tesis de licenciatura , Facultad de

Ciencias-UNAM, México 1980.

[11] Searle S.R.," Linear Models ", John Wiley and Sons, 1981.

[12] Hernández A. y Ochoa R. M ," Aplicaciones de técnicas de Análisis Multivariado en el estudio de embalses temporales", Comunicaciones Técnicas : serie naranja No 300, IIMAS-UNAM, México, 1982.

[13] Méndez R.I y Rodríguez S., " Dos ejemplos de aplicación de Análisis Discriminante en Medicina ", Comunicaciones Técnicas: serie naranja No 179 ,IIMAS-UNAM, México. 1978.

[14] Kshirsagar A. M and Arseven E, "Note on the equivalency of two Discriminant Procedures".The American Statistician, February 1975, vol 29 , No 1 , 38-39 pp.

[15] Baz G ,"Análisis Discriminante no Paramétrico".Comunicaciones técnicas: serie azul No 70,IIMAS-UNAM, México, 1983.

[16] Anderson T. W, " An Introduction to Multivariate Analysis", John Wiley and Sons", 1974.

[17] Seber G. A. F, "Multivariate Observations", John Wiley and Sons, 1984.